
# UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

"Entrenamiento dinámico de modelos acústicos de reconocedores de voz para los corpora en español de México: DIMEx100 niños y adultos "

Tesis:

Para obtener el título de

**Ingeniero en Computación**

Presenta:

**Edith Moya García**

Directores de Tesis:

**Dr. Manuel Hernández Gutiérrez**

**Dr. Ivan V. Meza Ruiz**

Huajuapan de León, Oaxaca. Abril 2011.

# Índice general

# Índice de figuras

# Índice de cuadros

# Capítulo 1

# Introducción

En los años 50's se dio inicio a diversos estudios por parte de investigadores y desarrolladores para explorar la posibilidad de crear sistemas donde los seres humanos interactuaran de forma hablada con máquinas (interacción humano-computadora). Dentro de estas investigaciones se ha buscado que las máquinas sean capaces de comunicarse mediante lenguaje hablado para realizar diversas tareas, todo esto por medio de reconocimiento automático de voz (ASR, por sus siglas en inglés). El objetivo de ASR es que una máquina de forma automática convierta las palabras, que son emitidas por el ser humano, a texto. En la actualidad las investigaciones que se han llevado a cabo en reconocimiento de voz han tenido notables avances, pero aún con varias limitaciones en este tipo de tecnologías donde se busca construir sistemas computacionales que interactúen en lenguaje natural con el ser humano. A pesar de las limitaciones se busca extender las áreas de la vida cotidiana en las que puedan ser aplicados estos sistemas, así como crear aplicaciones para un mayor tipo de hablantes [Pérez Pavón, 2006, Juárez Vázquez, 2009].

Debido a la creciente importancia tanto de la construcción como el uso de reconocedores de voz y a la disponibilidad de los recursos tan particulares, en esa tesis se busca realizar entrenamientos dinámicos de modelos acústicos para caracterizar empíricamente los corpora *DIMEx100 niños* y *DIMEx100 adultos* que son utilizados para crear reconocedores de voz del español de México para dos tipos de hablantes: niños y adultos. De manera general, se sabe que la construcción de un reconocedor de voz implica el uso de un corpus por lo que para este trabajo se emplearán -como ya se mencionó- un corpus de niños y otro de adultos. Que si por una parte estos recursos son limitados en cuanto a tamaño; por otra, en el aspecto fonético, no lo son por la cuidadosa planeación que se les dio al crearlos, por la cantidad de hablantes con las que cuenta y por ser corpora de habla leída. Además de que ambos corpus se puede decir que son paralelos ya que comparten las mismas frases. Su única diferencia radica en el tipo de hablante.

En la sección 1.1 se exponen las razones que motivaron a realizar esta tesis, posteriormente en la sección 1.2 se enlistan los objetivos de este trabajo y finalmente en la sección 1.3 se presenta la organización general de esta tesis.

## 1.1. Motivación

La comunicación hablada ha sido importante para los seres humanos desde su existencia para poder transmitir ideas y comunicarse. Hoy en día es común ver como las personas hacen un constante uso de la tecnología en diferentes áreas de su vida cotidiana principalmente usando computadoras y teléfonos celulares, en los que puede ser aplicado el reconocimiento automático de voz como medio para manipularlos. Ejemplo de las áreas donde se tiene

presente este tipo de tecnologías del habla se pueden mencionar: la medicina, la educación, el entretenimiento, sólo por mencionar algunos. Sin embargo, para el desarrollo de aplicaciones en español de México se cuentan con recursos muy limitados y con muy pocas investigaciones en esta área, pero se debe de tener en cuenta para futuras investigaciones que un reconocedor de voz para el español de México es un recurso muy valioso y útil para el desarrollo de diversas aplicaciones en el país [Pérez Pavón, 2006].

Por lo mencionado anteriormente se tiene el gran interés de crear sistemas computacionales que integren la comunicación hablada como medio de interacción, haciendo uso de sistemas conversacionales de diálogo como interfaz. Un sistema de diálogo está formado por las siguientes partes: entendimiento del lenguaje, manejador del diálogo y generador de respuesta (Fig. 1.1). En la etapa de entendimiento del lenguaje se encuentra uno de los módulos que es sumamente importante en este tipo de sistemas: el reconocedor de voz.

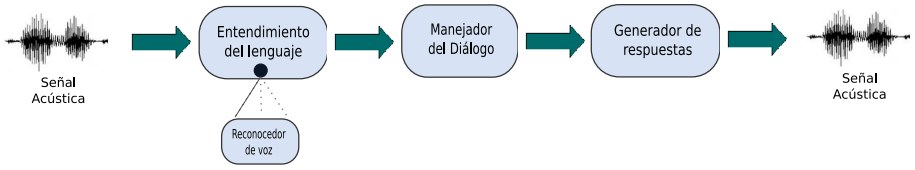


Figura 1.1: Arquitectura básica de un sistema de diálogo [Tapia et al., 2010].

Estos aspectos de los tipos de hablantes en conjunto con las características del corpus motivaron el desarrollo de esta tesis. Los corpora presentan características muy particulares, pues comparten las mismas oraciones con la única diferencia del tipo de hablante que se utilizó para grabarlas.

Un factor que se debe tener presente para el desarrollo de reconocedores de voz es el tipo de hablante al que va dirigido (niños, adultos, adultos mayores, mujeres, etc.), debido a que

cada tipo de hablante posee características que lo hace único y distinto, que a su vez deben de tomarse en cuenta durante el proceso de reconocimiento para que las palabras emitidas por el hablante sean reconocidas e interpretadas correctamente. Algunos ejemplos de estas características que poseen cada uno de estos tipos de hablantes son: la pronunciación, el acento, la tonalidad de su voz, cantidad de vocabulario con la que cuenta, edad, género, entre otros. Además se debe de tenerse en cuenta que de acuerdo al tipo de hablante al que va dirigido el reconocedor de voz, se recopilan recursos específicos que cumplan con las características de cada tipo de hablante. Como ejemplo de este tipo de recursos se puede mencionar al corpus *DIMEx100 niños* y *DIMEx100 adultos*. Algunos de los factores más comunes e importantes que se deben de tomar en cuenta al momento de desarrollar un reconocedor de voz para niños y adultos son: por la diferencia de edad la voz de un niño suele ser más aguda, la pronunciación que el hablante le da a las palabras, la cantidad de vocabulario que posee el hablante, el lugar de origen del hablante por el acento y pronunciación que le da a las palabras, entre otros [Blomberg and Elenius, 2004, Russell and D' Arcy, 2007].

La mayoría de las investigaciones que se han realizado en el área de reconocimiento de voz para diversos tipos de hablantes (hablantes nativos del país y del idioma en estudio) por universidades y laboratorios han sido en el extranjero. Estas investigaciones tienen como objetivo crear sistemas que reconozcan la voz de hablantes de diversos idiomas, como lo son el inglés británico, alemán, italiano, sueco, entre otros [Giuliani and Gerosa, 2003, Wilpon and Jacobsen, 1996, Batliner et al., 2005]. En México el desarrollo e investigación en tecnologías orientadas al reconocimiento de voz del español de México son aún muy escasas y recientes en comparación con las realizadas en el extranjero. A pesar de los pocos proyectos de investigación con que se cuentan en el país, uno de

los proyectos que se interesa actualmente por el estudio de este tipo de tecnologías es el proyecto *DIME* (Diálogos Inteligentes Multimodales en Español) desarrollado en el IIMAS de la UNAM, el cual cuenta con un reconocedor de voz del español de México para adultos [Pineda, 2008].

Las tecnologías e investigaciones desarrolladas hasta el momento en el área del reconocimiento de voz tanto en el extranjero como en el país, buscan adaptar y extender las capacidades de los sistemas de reconocimiento de voz, a fin de que sean capaces de interactuar con distintos tipos de hablantes. El desarrollo de este tipo de aplicaciones para niños tiene un gran futuro, ya que para estos hablantes se tiene un amplio campo de estudio para desarrollar múltiples aplicaciones en la educación, entretenimiento, salud, etc.; debido a que actualmente los niños se encuentran en constante acercamiento con la tecnología, en sus hogares, en la escuela y cuando juegan [Russell and D' Arcy, 2007].

Los recursos fonéticos empleados en la creación de reconocedores de voz son de suma importancia para obtener sistemas con un alto desempeño en el reconocimiento, ya que estos recursos incluyen las características del tipo de hablante al que va dirigido. Para este trabajo de investigación tanto el corpora de habla leída de niños como de adultos con respecto a la parte fonética se cuido que fueran balanceados, es decir, que contaran con todos los fonemas del español de México en todos sus contextos posibles. Además ambos corpora cuentan con una característica que los hace muy peculiares ya que son paralelos en cuanto a las frases que cada uno de ellos contiene. El primer hablante del corpus *DIMEx100 niños* ha leído la misma frase correspondiente al primer hablante del corpus *DIMEx100 adultos*, y así sucesivamente para todos los hablantes de ambos corpora.

El objetivo principal de esta tesis es realizar entrenamientos dinámicos de los modelos

acústicos para caracterizar de manera empírica reconocedores de voz en español de México, los cuales han sido creados con distintas proporciones de los corpora de habla leída *DIMEx100 niños* y *DIMEx100 adultos*, así como del corpus de habla espontánea *Golem-Universum*. Para cumplir con el objetivo de caracterizar los recursos disponibles se realizaron diferentes experimentos donde se evalúa dinámicamente modelos acústicos. Con los experimentos se espera determinar si los recursos limitados con que se cuentan son aptos y tienen las características necesarias que se necesitan para mejorar el desempeño un reconocedor de voz en español de México o si es conveniente realizar algunos cambios en los recursos.

Los recursos disponibles para llevar a cabo esta tesis -como se ha mencionado- son dos corpora fonéticos del español de México: el corpus *DIMEx100 adultos* y el corpus *DIMEx100 niños* ambos son corpora leídos del español de México, además se utiliza un corpus de habla espontánea, el cual fue recolectado durante evaluaciones realizadas al juego "Adivina la carta", así como un reconocedor de voz en español de México para adultos, que ha sido entrenado con el corpus *DIMEx100 adultos*. Este reconocedor también ha sido probado en tiempo real con habla espontánea por hablantes adultos. El reconocedor de voz presentó un desempeño aceptable durante el proceso de reconocimiento [Pérez Pavón, 2006, Pineda et al., 2009].

## 1.2. Objetivo

Debido a la importancia que tiene el desarrollo de tecnologías del habla en la actualidad y a las deficiencias que se han encontrado en el rendimiento del sistema *Golem-Universum* durante el proceso de reconocimiento, en particular para niños, el objetivo principal de esta tesis es realizar entrenamientos dinámicos de modelos acústicos para caracterizar de manera

empírica reconocedores de voz en español de México creados principalmente con los corpora *DIMEx100 niños* y *DIMEx100 adultos*.

Mediante la realización de diferentes evaluaciones empíricas utilizando los corpora de habla leída se espera poder caracterizarlos para identificar la mejor forma de combinar dos tipos de hablantes: niños y adultos. Estos recursos que se tienen disponibles son limitados pero tienen la ventaja de que son ricos fonéticamente, pues cuentan con un significativo número de hablantes distintos, y son representativos del español de México. Además de que cuentan con la particularidad de que ambos corpora tienen el mismo conjunto de oraciones.

Para cumplir con el objetivo de realizar diversos entrenamientos dinámicos de los modelos acústicos para caracterizar los corpora, se plantearon las siguientes preguntas que ayudan a observar el comportamiento de un reconocedor de voz cuando es entrenado con diferentes proporciones de datos de entrenamiento:

- ¿Cuál es el comportamiento de un reconocedor que es entrenado para un tipo de hablante, pero interactúa con otro tipo de hablantes para el que fue entrenado?
- ¿Cómo se comporta un reconocedor de voz que es creado para distintos tipos de hablantes?
- ¿Qué sucede si la información de un tipo de hablante es agregada poco a poco durante el proceso de entrenamiento del reconocedor de voz?
- De acuerdo a las características que poseen los corpora que se tienen disponibles, ¿Cuál es el desempeño que presenta un reconocedor de voz si se mezclan de manera complementaria los datos de los dos corpora para el entrenamiento?

Para dar respuesta a estas interrogantes se diseñaron un conjunto de experimentos, los

cuales se describen más a detalle en el capítulo 4. Estos experimentos fueron diseñados teniendo en cuenta los siguientes factores metodológicos:

- Los resultados obtenidos de cada evaluación deben de ser comparables.
- La mayor carga de trabajo, durante las evaluaciones realizadas a cada uno de los reconocedores de voz, debe de estar presente en los modelos acústicos, de tal manera que los modelos de lenguaje fueran pobres.
- Los experimentos son exhaustivos.

## 1.3. Organización de la tesis

Los siguientes capítulos de la tesis se encuentran organizados de la siguiente manera.

En el capítulo dos se da un panorama detallado de los elementos principales que integran un reconocedor de voz, también se describe de manera general su funcionamiento y los elementos que lo conforman; además de dar una introducción a algunos modelos y teorías utilizadas durante el procesamiento del lenguaje natural, principalmente se aborda el concepto de Modelos Ocultos de Markov (HMM por sus siglas en inglés, Hidden Markov Models).

Posteriormente en el capítulo tres se presentan los avances que se han tenido en los últimos años en el área del procesamiento de lenguaje, así como los experimentos realizados en el reconocimiento de voz para niños y adultos; además se presenta el contexto en el que se encuentra ubicado este trabajo de tesis; también se muestra la importancia que tienen los recursos lingüísticos para la creación de un buen reconocedor de voz, como caso de estudio se

presenta el corpus *DIMEx100*, del cual se da una descripción de las principales características y elementos que lo componen.

En el capítulo cuatro, se presenta la descripción de los experimentos y los resultados obtenidos de la caracterización de los corpora de habla leída.

En el capítulo cinco, se presentan los experimentos donde es evaluado un reconocedor de voz con habla espontánea. Estas evaluaciones ayudan a complementar la caracterización de los corpus de habla leída. Además en este capítulo se muestra la comparación de los resultados de las evaluaciones entre los diferentes experimentos que se realizaron.

Finalmente, en el capítulo seis se presentan las conclusiones, así como el trabajo a futuro.

# Capítulo 2

# Reconocimiento de voz

Actualmente para desarrollar un sistema reconocedor de voz se requiere que los elementos que lo integren tengan un mejor diseño con la finalidad de crear sistemas robustos capaces de reconocer dominios más amplios en cuanto a tipos de hablantes. Por ejemplo se pueden realizar mejoras al modelo acústico, al modelo de lenguaje o enriquecer el diccionario de pronunciación. Es importante tener conocimiento de la formulación matemática de cada elemento que compone a un reconocedor de voz para comprender mejor los problemas que se pueden presentar durante el desarrollo de este tipo de sistemas y así poder solucionarlos computacionalmente.

Este capítulo se encuentra organizado de la siguiente manera: en la sección 2.1 se describe brevemente el funcionamiento de un reconocedor de voz y se da una breve introducción de la formulación matemática involucrada en el reconocimiento de voz, en la sección 2.2 se explica el funcionamiento de cada uno de los elementos básicos que integran a un reconocedor de voz. En la sección 2.3 se describe el funcionamiento y la interacción de un reconocedor de

voz en los sistemas de diálogo, además se mencionan algunos ejemplos de estos sistemas con reconocimiento de voz para distintos tipos de hablantes.

## 2.1. Funcionamiento de un sistema reconocedor de voz

Un reconocedor de voz es un dispositivo que es capaz de interpretar los sonidos emitidos por el hablante convirtiéndolos automáticamente a una transcripción ortográfica. Dentro de los componentes que integran un sistema reconocedor de voz, se se pueden mencionar los siguientes: la etapa de procesamiento o extracción de características, en esta etapa se toman muestras de la señal acústica emitida por el hablante, la cual es separada por rangos de tiempo (10, 15 o 20 milisegundos), posteriormente la señal dividida es convertida a una representación espectral lo que da como resultado los vectores de características espectrales; en la etapa de reconocimiento fonético se usan técnicas de probabilidad como son redes neuronales o Modelos Gausianos para calcular las probabilidades individuales de cada uno de los segmentos en que fue dividida la señal acústica; finalmente, en la etapa de decodificación se obtiene la secuencia de palabras que tenga la mayor probabilidad. Este proceso se lleva a cabo utilizando diccionarios de pronunciación y un modelo de lenguaje, además de utilizar el algoritmo Viterbi o A* [Jurafsky and James H., 2008].

Los elementos que integran un sistema reconocedor de voz son: el modelo acústico, modelo de lenguaje y el modelo de pronunciación. En el modelo acústico se encuentra la variabilidad acústica de una lengua. Este modelo recibe como entrada una señal acústica de la que se extraen sus propiedades y posteriormente se obtiene un vector de características, el cual será

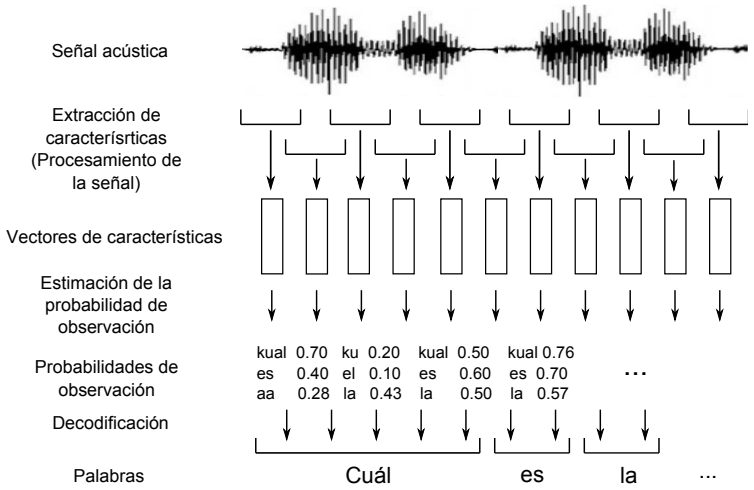


Figura 2.1: Arquitectura de un reconocedor de voz [Jurafsky and James H., 2008].

comparado para obtener el patrón que tenga la mayor probabilidad de ocurrencia. Una de las técnicas probabilísticas utilizadas para la creación de modelos acústicos son los Modelos Ocultos de Markov (HMM, por sus siglas en inglés). El modelo de lenguaje representa la probabilidad de que una secuencia de palabras forme parte de un lenguaje, debido a que en este modelo se encuentran las propiedades lingüísticas del lenguaje; además en este modelo se calcula la probabilidad a priori de la secuencia de palabras usando un modelo de predicción de la palabra llamado n-gramas [Jurafsky and James H., 2008]. El modelo de pronunciación está formado por los diccionarios de pronunciación los cuales son utilizados junto con el modelo acústico para el alineamiento automático del vector de características de la palabra. Los diccionarios de pronunciación son construidos a partir de un corpus y contienen las pronunciaciones más comunes de cada palabra que se encuentran contenidas en el corpus.

La ecuación que representa al modelo del canal ruidoso o modelo de inferencia de Bayes, es un modelo probabilístico de variación de pronunciación y ortografía. En esta ecuación se

puede identificar los modelos que integran un reconocedor de voz (ecuación 2.3). De acuerdo con este modelo los problemas que se presentan en los sistemas reconocedorea de voz son modelados como una tarea de clasificación donde una cadena de texto genera otra cadena de texto, siendo para este caso, la pronunciación correcta que ha sido emitida por el hablante. La entrada del reconocedor en este modelo es una señal acústica que ha pasado por un canal de comunicación ruidoso y que posteriormente es convertida a una cadena de texto. El canal de comunicación introduce ruido a la señal acústica debido a varios factores entre ellos: variación de pronunciación del hablante, variación de la realización de los fonemas, variaciones acústicas en el canal de comunicación, entre otros [Jurafsky and James H., 2008].

Para el caso de una palabra, en este modelo se espera encontrar la palabra $w$ la cual corresponde a una secuencia de fonemas que es proporcionada por el hablante. La búsqueda de la palabra se realiza en un vocabulario $V$ (universo de palabras) en el que se selecciona la secuencia de observaciones $O$ más probables. La ecuación que representa la palabra con mayor probabilidad de ocurrencia está dada por:

$$\hat{w} = \arg\max_{w \in V} P(w|O) \tag{2.1}$$

Para calcular $P(w|O)$, es necesario aplicar la regla de Bayes, de tal forma que la ecuación (2.1) queda transformada de la siguiente manera:

$$\hat{w} = \arg\max_{w \in V} \frac{P(O|w)\ P(w)}{P(O)} \tag{2.2}$$

El cálculo de la probabilidad de observación de la secuencia de observaciones $P(O)$ en la ecuación (2.2) es irrelevante debido a que la secuencia de observaciones $O$ para cada palabra siempre es la misma. Tomando en cuenta la consideración anterior la ecuación que representa

al modelo del canal ruidoso en reconocedores de voz queda representada por:

$$\hat{w} = \arg\max_{w \in V} \underbrace{P(O|w)}_{Prob.\ de\ observacion} \underbrace{P(w)}_{Prob.\ a\ priori} \tag{2.3}$$

Como se puede apreciar en la ecuación (2.3) se observan todos los elementos de un reconocedor de voz. La probabilidad de observación se calcula con el modelo acústico, de pronunciación y la probabilidad a priori se calcula a través del modelo de lenguaje. Además cabe mencionarse que esta ecuación se encuentra expresada en términos de una palabra $w$, pero puede ser generalizada para una secuencia de palabras $W = w_1, w_2, ..., w_n$, siguiendo el mismo procedimiento descrito durante esta sección.

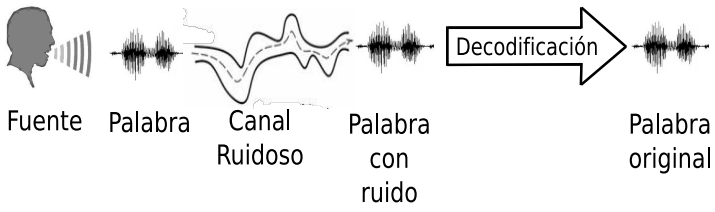


Figura 2.2: Modelo del canal ruidoso.

## 2.2. Elementos básicos de un reconocedor de voz

En la sección anterior se presentó la ecuación (2.3) en la que se pueden observar los elementos que integran a un reconocedor de voz, así como el proceso de reconocimiento. En esa ecuación $P(O|w)$ representa el modelo acústico y de pronunciación y $P(w)$ al modelo de lenguaje.

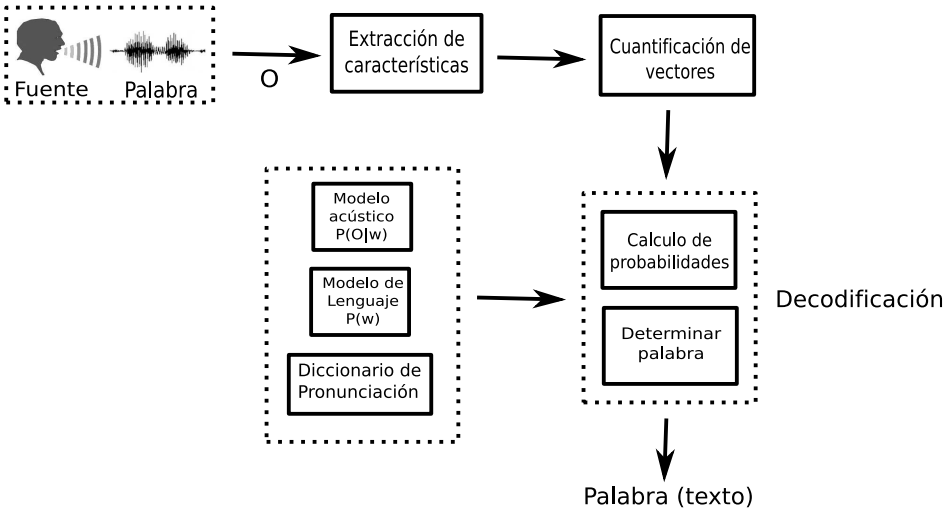


Figura 2.3: Proceso de reconocimiento de palabras.

En la figura 2.3 se muestra el diagrama de bloques que describe el proceso de reconocimiento de una palabra de acuerdo al modelo del canal ruidoso. La entrada del sistema es una señal acústica que normalmente contiene ruido. La señal es procesada en el módulo de extracción de características y el módulo de cuantificación de vectores; la función de estos módulos es extrar la señal propia del audio eliminando el ruido que pueda contener para posteriormente obtener los vectores de características de la señal. En el módulo de decodificación se realizan las tareas del cálculo de las probabilidades y de la selección de la palabra que tenga la probabilidad más alta, la cual es la palabra reconocida por el sistema reconocedor de voz, en este módulo es donde el modelo de lenguaje, modelo acústico y el diccionario de pronunciación trabajan conjuntamente para obtener la palabra reconocida.

A continuación se describen cada uno de los modelos que integran al reconocedor de voz, además de mencionar brevemente los algoritmos y modelos que utilizan cada uno de ellos para realizar su función dentro del proceso de reconocimiento de voz.

### 2.2.1. Modelo acústico

El modelo acústico es la parte del reconocedor de voz que contiene la variabilidad acústica de la señal de entrada. Para la creación del modelo acústico los elementos que deben de estar presentes son: el diccionario de pronunciación, que contiene todas las posibles pronunciaciones de las palabras propias del contexto de reconocimiento y un conjunto de datos acústicos, en este caso se refiere a un corpus de habla, la expresión que forma parte de la ecuación del modelo del canal ruidoso presentada en la sección anterior, hace referencia al modelo acústico para una palabra, está dada por $P(O|w)$. Esta expresión indica que el proceso que se lleva a cabo para obtener la hipótesis de la palabra emitida por el hablante, consiste en asignar probabilidades a cada una de las subpalabras (fonemas) que forman la palabra, creando un conjunto de modelos estadísticos, los cuales son los símbolos de observación en un Modelo Oculto de Markov (HMM, por sus siglas en inglés).

Los modelos acústicos fonéticos están basados en HMMs. El primer paso que se debe de realizar para crear un HMMs de una palabra consiste en crear un diccionario fonético de las palabras, con base a las pronunciaciones propias del contexto al que está dirigido el reconocedor de voz. Cada uno de los símbolos que forman la palabra le corresponde un HMM elementario (ver fig 2.4); al concatenar todos los HMMs elementarios se obtiene el HMM de la palabra donde, el estado final de el HMM es concatenado con una transición nula, seguido del estado inicial del próximo HMM. Para estimar los parámetros desconocidos en un HMM se hace uso del algoritmo Baum-Welch que a su vez utiliza el algoritmo forward-backward. El proceso descrito anteriormente para una palabra es generalizado para una secuencia de palabras ahora concatenando un estado de silencio entre cada HMM de la palabra, tal como se muestra en el esquema de la figura 2.5 [Jelinek, 1999].

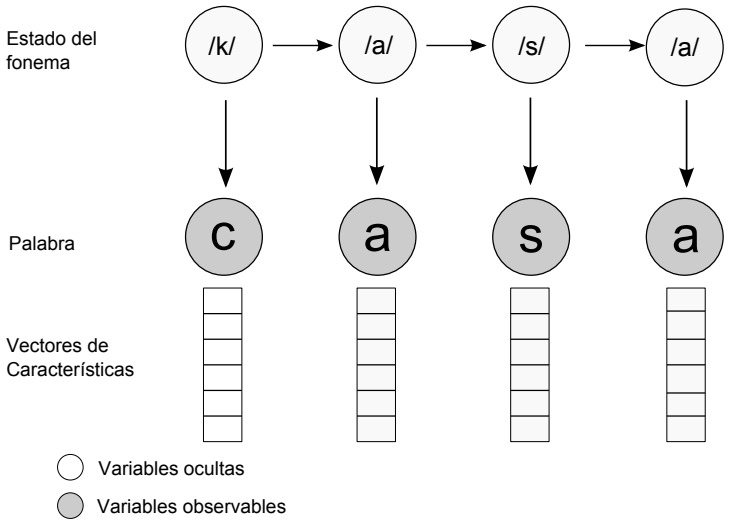


Figura 2.4: HMM de una palabra.

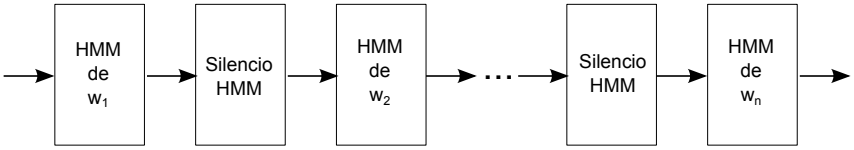


Figura 2.5: Conjunto de HMMs para el texto de entrenamiento $w_1, w_2, ..., w_n$ [Jelinek, 1999].

Durante el proceso de la creación de los modelos acústicos como ya se menciono se utilizan diversos algoritmos, para el cálculo de las probabilidades se hace uso de los HMMs porque son modelos generativos de las características vocales. Por otra parte el análisis de la señal acústica se lleva a cabo con técnicas de filtrado de LPC (por sus siglas en ingles, Linear Predictive Coding) PLP (por sus siglas en ingles, Perceptual Linear Predictive) y MFC (por sus siglas en ingles, Mel Frecuency Ceptral Coeficients).

Existen dos métodos utilizados para la creación de modelos acústicos: los métodos deterministicos y los estocásticos. Los métodos deterministicos utilizan técnicas de coincidencia de

patrones acústicos, por otro lado los en los métodos estocásticos se caracteriza la variabilidad de la voz [Juárez Vázquez, 2009].

### 2.2.2. Modelo de lenguaje

El modelo de lenguaje es un modelo estadístico de secuencia de palabras, donde son asignadas probabilidades a priori que le corresponde a una cadena de palabras, con la finalidad de conocer cuál fue la palabra original articulada por el hablante. El objetivo de los modelos de lenguaje es reducir el espacio de búsqueda y capturar el contexto de uso de las palabras a reconocer calculando la probabilidad de ocurrencia de una o más palabras.

Aplicando la regla de Bayes a $P(W)$, queda formalmente expresado como:

$$P(W) = \prod_{i=1}^{n} P(w_i|w_1, ..., w_{i-1}) \tag{2.4}$$

Para predecir la siguiente palabra de una secuencia de palabras, se calcula la probabilidad de que una palabra sea emitida dado que $w_1, ..., w_{i-1}$ han sido previamente dichas; calcular esta probabilidad resulta complicado de realizar pero es una tarea esencial en el reconocimiento de voz. Para realizar este cálculo se utiliza el modelo de N-gramas que resuelven parcialmente el problema pues reduce el número de palabras a N. El modelo de N-gramas consiste en calcular estadísticamente la siguiente palabra con mayor probabilidad de pertenecer a una secuencia de palabras, es decir, el modelo utiliza N-1 palabras previas para predecir la siguiente palabra [Jurafsky and James H., 2008].

Este proceso de predicción de una palabra puede ser difícil de realizar, debido a factores tales como ruido en la palabra o ambigüedad. Para no realizar el cálculo de la probabilidad de todas las palabras previas se calcula para el caso de un bigrama únicamente la palabra

anterior, en un trigrama las dos palabras anteriores y así sucesivamente; al realizarse estos cálculos hacen más complicado y robusto al modelo de lenguaje. Esta modelación se basa en los modelos ocultos de Markov que asumen el hecho de que únicamente es necesario considerar cierta parte de la historia de la secuencia de palabras (palabras anteriores) para predecir la siguiente palabra. Este supuesto es conocido como el supuesto de Markov.

Para realizar el cálculo de la probabilidad condicional de los N-gramas de $N$ palabras previas, la ecuación queda expresada como:

$$P(W_n|W_1^{n-1}) \approx P(W_n|W_{n-N+1s}^{n-1}) \tag{2.5}$$

Los problemas que se presentan al utilizar este modelo de N-gramas es que dependen del corpus utilizado para el entrenamiento del modelo de lenguaje, por lo que si se tiene corpus pequeño el vocabulario con el que cuenta puede ser muy restringido. Cuando se presenta que las probabilidades obtenidas son 0 o cercanas a este valor se aplican técnicas de suavizado donde se les reasigna un nuevo valor distinto de cero.

### 2.2.3. Diccionario de pronunciación

Un diccionario de pronunciación está formado por un conjunto de palabras a las que se le asigna una correspondiente secuencia de sonidos (pronunciación), los cuales están representados por una serie de símbolos previamente establecidos. Las razones por las que es importante crear diccionarios de pronunciación son principalmente porque en ocasiones algunas palabras puede tener una o varias pronunciaciones, dependiendo del tipo de hablante que emitió la palabra, además de los distintos significados que toma una palabra de acuerdo al contexto en la que es usada. El conjunto de palabras contenidas en el corpus usualmente

corresponde al contexto al que está dirigido el reconocedor de voz.

En su mayoría los diccionarios de pronunciación utilizados en esta tesis fueron extraídos de los corpora empleados durante el proceso de entrenamiento. Cada uno de los corpora se encuentran etiquetados manualmente, por lo que las pronunciaciones contenidas en los diccionarios cuentan con varias pronunciaciones para una misma palabra.

Como ya se mencionó los diccionarios utilizados en esta tesis fueron construidos a partir de los corpora *DIMEx100 niños*, *DIMEx100 adultos* y del corpus *Golem-Universum*; además de palabras propias que enmarcan al contexto del juego "Adivina la carta", tales como nombres propios de personas, nombres de colores, características de objetos astronómicos, entre otros. El juego "Adivina la carta" tiene por objetivo que el jugador adivine una carta con objetos astronomicos que previamente ha sido seleccionada por el sistema, diciendole al sistema una serie de caracteristicas de acuerdo a lo que observa en un conjunto de cartas. En capítulo 3 sección 3.2.1, se explica más detalladamente el juego.

En el cuadro 2.1 se muestran algunas palabras extraídas del corpus *DIMEx100 niños* que pertenecen a uno de los diccionarios utilizados en los experimentos, con sus respectivas pronunciaciones.

De acuerdo a la cantidad de palabras que contenga el diccionario de pronunciación y el modelo de lenguaje dependerá que tan restrictivo es el reconocimiento del hablante, y por ende se tendrá un modelo de lenguaje más robusto; por otro lado si se cuenta con una gran cantidad de palabras, el reconocimiento es menos restrictivo pero pueden llegar a ocurrir otros problemas como la ambigüedad. Por lo mencionado anteriormente el corpus debe tener una buena riqueza léxica, ya que juega un papel muy importante para la creación

| **Palabra** | **Pronunciación** |
|---|---|
| JAVA | x a b a |
| JAVA(2) | Z a b a |
| JAVIER | x a b i e r1 |
| JEFA | x e f a |
| JEFE | x e f e |
| JEFES | x e f e s |
| JETTA | Z e t a |
| JIMENES | x i m e n e s |
| JIMENEZ | x i m e n e s |

Cuadro 2.1: Fragmento de un diccionario de pronunciación.

de diccionarios de pronunciación.

## 2.3. Reconocimiento de voz en los sistemas de diálogo hablado

En el área de la interacción humano-computadora se construyen los sistemas de diálogo hablado, los cuales pueden ser definidos como programas computacionales que tienen como objetivo establecer una comunicación con el fin de facilitar la interacción entre la computadora y el ser humano. Los sistemas de diálogo hablado son utilizados en la actualidad en diversas actividades cotidianas. Por ejemplo en las aerolíneas son utilizados para proporcionar cierta información a los pasajeros de forma automática, como son horarios de salida, el precio de los boletos, entre otras. En la telefonía estos sistemas se encargan de la marcación o contestación automática por voz al ingresar una simple entrada al sistema.

Por otra parte también se han desarrollado ciertas aplicaciones dirigidas a determinados

tipos de hablantes: niños, adultos, adultos mayores, mujeres, hombres, los cuales no son tan usuales encontrarlos en la vida cotidiana. Algunos ejemplos de estos sistemas de diálogo se describen brevemente a continuación (cuadro 2.2):

- **Sistema Fonexi.** Sistema desarrollado por la Universidad Autónoma de Yucatán, que cuenta con un juego de Memorama dirigido a niños con problemas de lenguaje con el objetivo de ayudarlos durante el proceso de rehabilitación y evaluación del habla [Miranda-Palma et al., 2007]. El juego cuenta con un reconocedor de voz para niños en español de México; para su desarrollo se utilizaron librerías de *Hidden Markov Model Toolkit* (HTK)[1]; el corpus de voz utilizado para desarrollar el reconocedor de voz, fue creado con hablantes nativos de el español de México [Miranda-Palma et al., 2007].

- **Proyecto Golem.** Proyecto desarrollado en el DCC (Departamento de Ciencias de la Computación) del IIMAS de la UNAM [Pineda, 2008]. En este proyecto fue creado el robot *Golem*, el cual tiene como tarea principal el dar vistas guiadas con una interacción en lenguaje natural de las investigaciones desarrolladas en el DCC. El proyecto cuenta con un reconocedor de voz automático en español de México para dos tipos de hablantes: hombres y mujeres adultos, y ha sido entrenado con el sistema SPHINX-3[2] y del corpus *DIMEx100 adultos* [Pineda et al., 2004].

- **Proyecto TRAINS.** El proyecto fue desarrollado con el propósito de crear y diseñar un asistente de planificación para un dominio conversacional especifico proporcionando ayuda de forma natural a un humano para que pueda resolver problemas que se le presenten relacionados con un sistema de trenes de carga por medio de una conversación

[1] http://htk.eng.cam.ac.uk/

[2] Software de la universidad de Carnegie Mellon. http://cmusphinx.sourceforge.net/

en inglés [Allen et al., 1994]. El reconocedor de voz de este sistema está orientado a reconocer voz de personas adultas. El corpus utilizado en este proyecto está formado por hablantes nativos del inglés de América del Norte [Heeman and Allen, 1995].

- **Chester.** Este sistema es un asistente conversacional que tiene por objetivo proporcionar una alternativa a adultos mayores para que desde sus hogares puedan monitorear y cuidar su salud, por medio de recordatorios de las horas en que deben tomar determinados medicamentos o cuando tengan algún molestar preguntarle al sistema *Chester* que acción deben de efectuar, todo esto por medio de una conversación coherente entre el sistema y el usuario. Para el desarrollo del sistema reconocedor de voz de *Chester* se hizo del diccionario CMU que contiene palabras en idioma inglés [Allen et al., 2006].

### 2.3.1. Etapas de un sistema de diálogo

Los sistemas de diálogo hablado se dividen en etapas o módulos que realizan diferentes procesos. La figura 1.2 muestra un esquema general de un sistema de diálogo hablado con los módulos más comunes. El módulo del reconocimiento del habla se convierte la entrada recibida (señal acústica) a texto únicamente de las palabras que fueron reconocidas por el sistema. Posteriormente, en el módulo de interpretación semántica se determina el significado de la secuencia de palabras que fueron reconocidas en el módulo anterior. En el módulo de gestor de diálogo se determina la tarea que se debe de realizar para ello se debe considerar el significado que se le dio a la entrada al sistema. En el módulo de generación de lenguaje se crea el enunciado que será la respuesta del sistema, a partir de la representación interna proporcionada en el módulo de gestión del dialogo. Finalmente se tiene el módulo

| | SISTEMAS DE DIÁLOGO HABLADO | | | |
|---|---|---|---|---|
| **Características** | **Sistema Fonexi** | **Golem** | **TRAINS** | **Chester** |
| **Objetivo** | Ayudar en la pronunciación de algunas palabras y evaluación para niños con problemas de lenguaje. | Implementar sistemas conversacionales del español de México y reconocimiento de voz [Pineda et al., 2004]. | Proporcionar ayuda a usuarios de un sistema de trenes de carga por medio de una conversación. | Desarrollar un asistente conversacional que proporcione ayuda en el monitoreo de la salud en adultos mayores. |
| **Tipo de hablantes** | Reconocedor de voz para niños | Reconocedor de voz para adultos | Reconocedor de voz para adultos | Reconocedor de voz para adultos mayores |
| **Tipo de palabras a reconocer** | Palabras cortas de un vocabulario restringido | Frases largas y cortas de habla espontánea | Frases largas de habla espontánea | Frases largas y cortas de habla espontánea |
| **Idioma** | Nativos del español de México | Nativos del español de México | Nativos del inglés de Norte América | Adultos mayores nativos del inglés |
| **Herramientas para el desarrollo del reconocedor de voz** | Hidden Markov Model Toolkit (HTK) | Sphinx-3 | Sphinx-3 | Sphinx-3 |

Cuadro 2.2: Características de los sistemas de diálogo hablado para distintos tipos de hablantes.

de conversión de texto en habla donde se toma el enunciado creado y se transforma a señal acústica, que será la respuesta esperada por el usuario [Llisterri and Machuca, 2006].

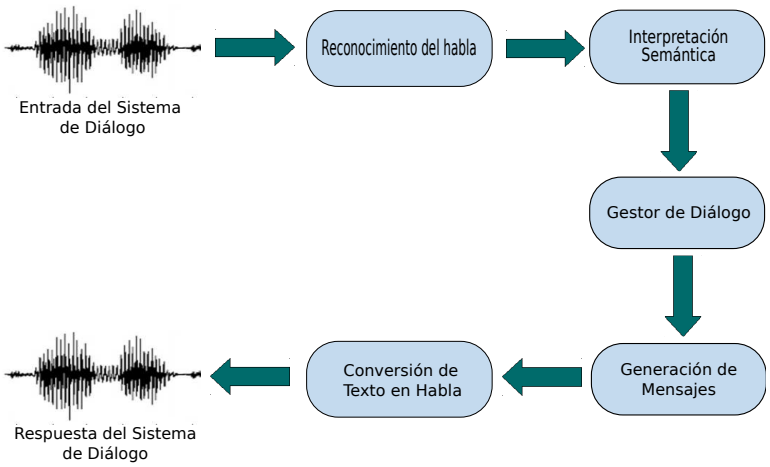


Figura 2.6: Módulos de un sistema de diálogo hablado.

Para el trabajo de investigación de esta tesis, se tiene centrada toda la atención e interés en mejorar el desempeño del módulo de reconocimiento de voz para hablantes del español de México, este módulo es sumamente importante debido a que es donde se proporcionan las características propias de la aplicación, como son el dominio de las palabras o frases y tipos de hablantes que únicamente el sistema va a ser capaz de reconocer, etc. Los resultados obtenidos en esta tesis, serán empleados para mejorar el módulo del sistema reconocedor de voz del proyecto *Golem-Universum* para que el desempeño del reconocimiento de voz en niños sea lo más eficientemente posible, con los recursos limitados que se tiene disponibles.

# Capítulo 3

# Antecedentes

Debido al gran número de aplicaciones en donde son involucrados los niños como usuarios, actualmente existe un amplio campo de estudio para crear sistemas orientadas al reconocimiento de voz para este tipo de hablantes. Tales como tutores de lectura o para el aprendizaje de determinado idioma, aplicaciones para juegos educativos, etc. El desarrollo de aplicaciones orientadas a niños no ha sido tan estudiada hasta la fecha tanto como en adultos, debido a las características tan diferentes que presenta la voz de un niño, principalmente en la tonalidad de la voz [Russell and D' Arcy, 2007].

Investigaciones realizadas en reconocedores de voz para niños, buscan mejorar el desempeño de este tipo de sistemas en los últimos años, con el fin de crear aplicaciones robustas, confiables y con un óptimo funcionamiento. Cabe mencionar que en esas investigaciones se han llevado a cabo diversos tipos de experimentos, principalmente para evaluar el reconocimiento para cada uno de los tipos de hablantes, así como la aplicación de diversas técnicas y métodos, todo esto con el único objetivo de mejorar las tecnologías orientadas al reconoci-

miento de voz en niños. En este mismo sentido se busca el desarrollo de nuevas aplicaciones que sean capaces de reconocer diversos tipos de hablantes, así como mejorar las aplicaciones ya existentes.

En este capítulo en la sección 3.1 se explican las características más importantes del corpus *DIMEx100 niños*, y en la sección 3.2 se presentan algunos proyectos y trabajos de investigación relevantes en donde se muestran algunas evaluaciones realizadas en reconocedores de voz para niños y adultos que forman parte de los antecedentes para esta tesis, estas evaluaciones tienen como objetivo determinar qué factores influyen en el desempeño de un reconocedor de voz, con la finalidad de que sea capaz de interactuar principalmente con niños y adultos.

## 3.1. El corpus *DIMEx100*

La disponibilidad de recursos fonéticos del español de México para crear sistemas orientados en el procesamiento del habla son muy limitados, a pesar de los notables avances que se han logrado en los últimos años. Principalmente, es necesario contar con un recurso lingüístico a nivel fonético lo suficientemente robusto para el desarrollo de aplicaciones de tecnologías de habla, que cuente con requerimientos específicos tales como considerar los alófonos[1] más comunes de cierto idioma [Pineda et al., 2004]. Como ejemplo de este tipo de recursos lingüísticos se tiene a los corpora fonéticos, los cuales consisten en una colección de frases de cierto idioma, seleccionadas de acuerdo a ciertos criterios lingüísticos.

Para la realización de esta tesis se tiene la disponibilidad de usar los corpora *DIMEx100*

[1] Un alófono es la pronunciación que se le da a las vocales y consonantes de un alfabeto, cuando estas son articuladas.

*adultos* y *DIMEx100 niños*, los cuales han sido desarrollados en el DCC del IIMAS, UNAM, como parte del proyecto *DIME-II* y del proyecto *Golem-Universum*, respectivamente. Ambos corpora fueron creados a partir de oraciones seleccionadas de la Web y cada una de las oraciones fueron ordenadas de menor a mayor grado de complejidad buscando que siempre se encontraran balanceados fonéticamente; a continuación se presentan ejemplos de oraciones que forma parte de los corpora[2]:

1. *Programa Nacional de Sociedad de la Información.*

2. *Debe tenerse presente la posibilidad de obtener falsos negativos con las pruebas cuánticas de tuberculina.*

3. *¿A qué se refiere protección en tiempo real?*

La distribución de la cantidad de las oraciones para cada uno de los corpus se seleccionaron de la siguiente manera: para ambos corpora, cada uno de los 100 hablantes leyó un total de 50 oraciones, obteniéndose un total de 5000 oraciones grabadas. Particularmente para el corpus *DIMEx100 adultos* cada hablante grabó adicionalmente 10 oraciones comunes, teniendo un total de 1000 oraciones adicionales; por lo tanto el corpus *DIMEx100 adultos* está compuesto de 6000 oraciones, mientras que el corpus *DIMEx100 niños* contiene 5000 oraciones [Pineda et al., 2004, Meza et al., 2010a].

Las características de los hablantes seleccionados para las grabaciones de estos corpora fueron principalmente la edad, nivel de estudios y lugar de origen, como se muestra en el cuadro 3.1. De acuerdo a estas características los hablantes que participaron en la creación

[2]Los ejemplos que presentan son oraciones individuales que pertenecen a los hablantes 17, 44 y 66 respectivamente.

del corpus *DIMEx100 adultos* fueron hablantes del centro del país, específicamente de la ciudad de México, en su mayoría estudiantes, investigadores y profesores de la Universidad Nacional Autónoma de México (UNAM); el rango de edad de los hablantes fue de 16 a 32 años de edad, por lo que el promedio de edad de los hablantes grabados fue de 24 años. Con respecto al género de los hablantes 49 % fueron hombres y 51 % mujeres [Pineda et al., 2009, Pineda et al., 2004].

Para la creación del corpus *DIMEx100 niños* se consideraron características semejantes de los hablantes del corpus *DIMEx100 adultos*, en el caso de la edad los niños seleccionados debían de tener entre 10 y 14 años de edad, ser estudiantes del último año de primaria o secundaria, además de ser hablantes nativos del español de México del centro del país [Meza et al., 2010a].

De acuerdo a sus características fonéticas, a la cuidadosa planeación llevada a cabo para la construcción de los corpora podemos concluir que se encuentran fonéticamente completos y balanceados, debido a que cuentan con un significativo número de fonemas[3] del español de México, además de que incluyen todos las unidades alofónicas en los tres niveles de etiquetación que se realizaron principalmente para el corpus *DIMEx100 adultos*. Cada nivel de etiquetación tiene las pronunciaciones más comunes de cada una de las unidades del español de México. Por todo lo mencionado anteriormente los corpora *DIMEx100 niños* y *DIMEx100 adultos* son un excelente recurso tanto para estudios fonéticos del español de México, así como principalmente para construir modelos acústicos y diccionarios de pronunciación; los cuales son empleados en la construcción de reconocedores automáticos de voz del español Méxicano [Pineda et al., 2009].

[3]Un fonema es una abstracción de los diferentes sonidos que forman parte de una lengua hablada.

Cabe resaltar, que estos recursos son peculiares dado que ambos corpora están compuestos por las mismas 500 oraciones leídas según corresponda el tipo de hablante del corpus (niño o adulto). En particular, la característica que hemos mencionado es la razón por la cual se busca caracterizar a los corpora. Para lograr este propósito se han planteado las siguientes preguntas:

- ¿Cuál es el comportamiento de un reconocedor de voz que fue creado específicamente para un tipo de hablante y que interactúa con un tipo de hablante distinto?
- De acuerdo a las características de los corpora, ¿Es posible complementar la información de ambos utilizada para el entrenamiento de un reconocedor de voz?
- ¿Cuál es el desempeño de un reconocedor de voz cuando se le agregan datos de entrenamiento poco a poco de un tipo de hablante?
- Y finalmente, si un reconocedor de voz es creado para distintos tipos de hablantes, ¿Cuál es su desempeño durante el reconocimiento?

Al dar respuesta a estas preguntas se espera conseguir el objetivo propuesto para esta tesis: caracterizar los corpus en español de México utilizados en la creación de reconocedores de voz.

## 3.2. Trabajo previo

En esta sección se describen brevemente algunos de los aspectos más relevantes en trabajos y proyectos de investigación que se han realizado en el extranjero y en el país; los cuales

| **Características** | **DIMEx100 adultos** | **DIMEx100 niños** |
|---|---|---|
| Cantidad de hablantes | 50 mujeres y 50 hombres | 50 niños y 50 niñas |
| Edad de los hablantes | 16 a 32 años | 10 a 14 años |
| Lugar de origen de los hablantes | Principalmente de la ciudad de México | Principalmente de la ciudad de México |
| Nivel de escolaridad | Medio superior y superior | Básica (primaria) |
| Número de frases por hablante | 50 frases distintas | 50 frases distintas |
| Número de frases comúnes | 10 frases | No disponible |
| Tamaño de las frases | de 7-15 palabras | de 7-15 palabras |
| Tipo de habla del corpus | Grabaciones leídas | Grabaciones leídas |
| Transcripciones | Ortográfica, T22, T44, T54 | Ortográfica |

Cuadro 3.1: Tabla comparativa de las características de los corpora *DIMEx100 niños* y *DIMEx100 adultos*.

tienen como objetivo mejorar y adaptar la tecnología desarrollada en reconocimiento de voz para el tipo de hablante de niños, haciendo uso de diferentes técnicas y procedimientos para lograr un buen desempeño en el reconocimiento. Primero se abordarán investigaciones en reconocimiento de voz para niños y finalmente para niños y adultos.

### 3.2.1. Reconocimiento de voz para niños

**Proyecto *Preparing Future Multisensorial Interaction Research* (EU FP5 PF-Star)**

En el proyecto *Preparing Future Multisensorial Interaction Research* (EU FP5 PF-Star) [Batliner et al., 2005], se han llevado a cabo diversos experimentos a fin de evaluar el desempeño del reconocimiento de palabras cortas, definidas en un vocabulario pequeño. Estos experimentos han sido realizados en niños de diferentes rangos de edades y en diferentes idiomas. Los hablantes que colaboraron en la elaboración de los diferentes corpora que posteriormente fueron utilizados para el entrenamiento del reconocedor de voz, han sido hablantes nativos y hablantes no-nativos del idioma. También se realizaron grabaciones de sonidos espontáneos e imitados de las reacciones de los niños ante ciertos eventos que les fueron presentados.

Las edades consideradas para las grabaciones leídas y espontáneas por niños fueron de 4 a 8 años para hablantes nativos del idioma inglés e italiano; para hablantes no-nativos del idioma alemán, italiano y sueco, el rango de edad de 4 a 12 años. En tanto las grabaciones leídas con niños hablantes no-nativos del idioma inglés el rango de edad fue de 10 a 11 años, así como para hablantes nativos del idioma alemán, italiano y sueco. Las grabaciones en las que participaron habla espontáneas y emociones de niños hablantes del inglés británico y alemán la edad fue de 4 a 14 años. Como puede apreciarse este proyecto es muy particular

por la gran variedad de recursos que se utilizaron para realizar las pruebas que fueron muy exhaustivas y variadas; con respecto a las características de los hablantes se tomaron en cuenta distintos rangos de edad e idioma de los niños (nativos, no nativos) y en el caso de los corpora voz espontánea y leída [Batliner et al., 2005].

El objetivo del proyecto *EU FP5 PF-Star* es extender las capacidades de los sistemas reconocedores de voz y que sean capaces de interactuar con dos tipos de hablantes: niños y adultos, además de realizar el análisis y la síntesis en expresiones de emociones que se obtienen en el habla espontánea para la voz y el rostro del hablante. En cada uno de los experimentos que se llevaron a cabo, se evaluó el desempeño en el reconocimiento para dos tipos de hablantes: niños y adultos. Para los experimentos el reconocedor de voz fue entrenado con voces de niños o adultos, utilizando diferentes proporciones de cada uno de los corpora disponibles: para el caso de los niños el corpus *PF-STAR* y para los adultos el corpus *SpeeCo*; el corpus de adultos es de habla leída, y el de niños tanto de habla leída como espontánea. Las pruebas que se realizaron fueron con un reconocedor de voz entrenado con adultos y evaluado con niños, de manera similar un reconocedor entrenado con niños fue probado con adultos. También se utilizaron diferentes técnicas y algoritmos para mejorar el reconocimiento y para adaptar principalmente un reconocedor de voz para adultos a niños, entre los que se encuentran: la técnica de Normalización de la Longitud del Tracto Vocal (VTLN, por sus siglas en inglés) y los algoritmos de adaptación *Maximun a Posteriori* (MPA) y *Maximum Likelihood Linear Regresion* (MLLR) [Blomberg and Elenius, 2004, Elenius and Blomberg, 2005].

Además se realizaron observaciones en los formatos de las frecuencias de las voces para cada uno de los tipos de hablantes (niños y adultos). Los resultados obtenidos de estas observaciones fueron comparados y evaluados para conocer que tanto afectaban en el recono-

cimiento de voz. Los resultados conseguidos en la investigación realizada en el proyecto EU FP5 PF-Star reflejan que al utilizar apropiadamente las técnicas de normalización y adaptación se obtienen mejoras en el desempeño en los modelos acústicos, con lo cual se logra una menor tasa de error de palabra (*word error rate*). Además se llegó a la conclusión de que los modelos acústicos de los reconocedores de voz presentan un mejor desempeño cuando son evaluados con palabras cortas, por el contrario el desempeño disminuye cuando se evalúa con frases largas ya que presentan una mayor complejidad [Elenius and Blomberg, 2005].

**Proyecto *Golem-Universum***

El proyecto *Golem-Universum* ha sido desarrollado en el DCC del IIMAS (UNAM), con el objetivo de acercar la tecnología de los sistemas de diálogo hablado y el área de la Inteligencia Artificial a los niños, mediante el juego llamado "Adivina la carta"; este proyecto es el antecedente directo de esta tesis. El juego es una aplicación específica que incluye un sistema de diálogo multimodal con habla y visión y está enfocado principalmente en el diseño de una arquitectura que sea capaz de soportar sistemas de diálogo multimodales que relacionen eventos lingüísticos y visuales [Meza et al., 2010a]. El sistema cuenta con un reconocedor de voz para niños hablantes del español de México, el cual forma parte del sistema de diálogo con habla y visión. Además el sistema es capaz de sostener una conversación fluida con el usuario y de interpretar lo más correctamente posible la información hablada que le es proporcionada.

El juego "Adivina la carta" está dirigido a niños de 10 a 14 años de edad. Al inicio del juego se le pregunta al jugador su nombre y edad, si el niño es muy pequeño (menor a 10 años de edad) no se le permite jugar, en caso contrario, se le proporcionan las reglas del juego,

| **Características del proyecto** | | |
|---|---|---|
| **Objetivo del proyecto** | Evaluar el desempeño en reconocimiento de voz para niños | |
| **Técnicas y algoritmos utilizados** | Técnica de normalización del Tracto Vocal (VTLN)<br>Algoritmos de adaptación MPA (Maximun a posteriori)<br>Maximum Likelihood Linear Regresion (MLLR)<br>Observaciones en los formatos de frecuencia de los corpora | |
| **Características del reconocedor de voz** | | |
| **Tipo de reconocimiento** | Habla continua y aislada | |
| **Tipo de frases a reconocer** | Fonemas y frases largas | |
| **Idiomas que reconoce** | Inglés británico, italiano, alemán y sueco | |
| **Características de los corpora** | | |
| **Corpora** | Corpus adultos: Base de datos *SpeeCo* | |
| | Corpus niños: Base de datos *PF-STAR* | |
| **Tipo de hablantes** | Hablantes nativos y no-nativos del idioma | |
| | Niños y niñas | |
| **Tipo de habla del corpus** | Grabaciones leidas y espontáneas | |
| **Edades** | **Leído** | 4-8 años de edad (hablantes nativos)<br>4-12 años de edad (no-nativos)<br>10-11 años de edad (no-nativos)<br>10-11 años de edad (nativos) |
| | **Espontáneo** | 4-14 años de edad |

Cuadro 3.2: Características generales del proyecto EU FP5 PF-Star.

por ejemplo, el niño debe de hacer preguntas sobre las características que observe de las 10 cartas ilustradas con objetos astronómicos (planetas, galaxia, instrumentos de medición, etc.) he ir diciéndolas una por una cuando sea su turno. De acuerdo a la característica dicha por el niño, el sistema le dará respuesta diciéndole si la característica que mencionó corresponde a la carta que se ha seleccionado aleatoriamente al inicio del juego para ser adivinada. El juego finaliza después de cuatro preguntas, o cuando es adivinada la carta antes de finalizar la ronda de preguntas. Finalmente, se le pide al niño que muestre al sistema la carta que cree que es la correcta, posteriormente se le indica si ha ganado la partida del juego por medio de voz y desplegando en pantalla la imagen de la carta correcta. El módulo del juego "Adivina la carta" se encuentra disponible en el museo de las ciencias Universum [Meza et al., 2010a, Meza et al., 2010b].

En el cuadro 3.3 se muestra un fragmento del diálogo que se establece entre el sistema *Golem-Universo* y el niño.

La implementación del reconocedor de voz que forma parte de uno de los módulos del sistema de diálogo, se llevó a cabo haciendo uso del corpus de voz *DIMEx100 niños* (en la sección anterior de este capítulo se describen las características más relevantes de este corpus) y de Sphinx-3, con el objetivo de crear una aplicación robusta y con un óptimo desempeño en el reconocimiento de voz para niños, en evaluaciones que se han realizado se ha detectado que el sistema es deficiente en el reconocimiento; por esta razón se tiene la necesidad de desarrollar a base de experimentos una metodología que indique los mejores parámetros de entrenamiento que mejoren el desempeño del reconocedor de voz. Las características del corpus utilizado en este proyecto se describieron en la sección anterior.

| **Hablante** | **Diálogo** |
|---|---|
| Sistema | Hola a todos me llamo *Golem* y éste es el juego "Adivina la carta". Tenemos jugador, ¿Cómo te llamas? |
| Jugador | Mauricio |
| Sistema | Hola Mauricio, gusto en conocerte ¿Cuantos años tienes? |
| Jugador | Tengo 10 años. |
| ... | |
| Jugador | ¿Es azul? |
| Sistema | Sí, vas bien, lo que hay en la carta tiene algo azul ... |
| Jugador | ¿Es un planeta? |
| Sistema | Lo siento en mi carta no aparece un planeta. |
| ... | |
| Sistema | Lástima, la carta que había escogido es de la luna. ¡Oh!, pero antes de irte por favor mira a la cámara. Gracias por jugar conmigo regresa pronto. |

Cuadro 3.3: Fragmento de diálogo entre el sistema *Golem-Universum* y el usuario.

| Factor | Evaluación 1 | Evaluación 2 | Evaluación 3 |
|---|---|---|---|
| Desempeño del ASR | 50 % | 50 % | 40 % |

Cuadro 3.4: Desempeño del ASR para el juego "Adivina la carta"

A lo largo del desarrollo de este proyecto se han realizado tres evaluaciones dos de ellas subjetivas y una completa (esta última incluye evaluaciones subjetivas y objetivas) con el objetivo de medir el desempeño general del juego "Adivina la carta". En las evaluaciones subjetivas la evaluación que se realizó consistió en preguntarle al usuario su opinión acerca del sistema, detectando que los niños perciben que en varias ocasiones el sistema no les entiende. Por otro lado en la evaluación objetiva se mide el desempeño del reconocedor de voz utilizando la medida *word error rate* y seleccionado los datos utilizados en las evaluaciones. El resultado que se obtuvo al evaluar el desempeño del reconocimiento en un reconocedor de voz fue de 0.59 de *word error rate*. Se evaluaron un total de 10 niños a los que se les formularon una serie de preguntas principalmente para poder observar su comportamiento y percepción que el niño tiene durante el uso del sistema.

De acuerdo a los resultados presentados en el cuadro 3.4 el desempeño del reconocimiento de voz ha sido constante en cada uno de las evaluaciones, teniéndose un 50 % en las primeros dos evaluaciones y para la tercera un 40 %; estos resultados indican que el reconocedor de voz con que actualmente cuenta el sistema su calidad en el desempeño del reconocimiento es aceptable y funcional, pero puede mejorar considerablemente lográndose tener un reconocimiento robusto para dos tipos de hablantes: niños y adultos. El desempeño del reconocedor

de voz afectan directamente al funcionamiento general del sistema así como de las demás funciones que es capaz de realizar, debido a que el reconocimiento del hablante es sumamente importante para que se logre una comunicación satisfactoria entre el niño y el sistema. De manera general los demás resultados obtenidos durante las dos últimas evaluaciones para los otros aspectos que fueron evaluados (Desempeño TTS, Facilidad de realizar la tarea, Ritmo de iteración, etc.), son atribuidas a las mejoras que se han hecho al sistema, principalmente a las estrategias de recuperación de errores implementadas durante el proceso de reconocimiento de habla del niño en el proceso del juego [Meza et al., 2010b].

### 3.2.2. Reconocimiento de voz en niños y adultos

La investigación desarrollada por los laboratorios AT&T Bell, está orientada para reconocer dígitos (palabras cortas). La base de datos utilizada en el proyecto fue Rafael.0, la cual fue elegida por que se asumió que la muestra de datos es suficientemente representativa para el entrenamiento del sistema reconocedor. La base de datos (corpus) fue creada con voces de hablantes nativos del idioma Danés con edad de 8 hasta más de 80 años, ya que fuera de este rango de edad, de acuerdo a las pruebas realizadas, el error en el reconocimiento incrementaba radicalmente. Los 487 hablantes utilizados para el entrenamiento del sistema fueron balanceados con respecto al género y a los 11 dialectos del idioma Danés.

Durante el desarrollo de la investigación se utilizaron frases del idioma Danés que fueron reconocidas por el sistema, así como Modelos ocultos de Markov (HMMs) para la representación de las pronunciaciones de cada uno de los hablantes (modelos acústicos). Los modelos utilizados para el entrenamiento del sistema reconocedor fueron entrenados con la técnica de

| **Características del proyecto** | |
|---|---|
| **Objetivos del proyecto** | Acercar la tecnología de los sistemas de diálogo hablado y el área de la Inteligencia Artificial a los niños |
| **Técnicas y algoritmos utilizados** | Se utilizó para la evaluación del sistema un cuestionario para medir la satisfacción del usuario: la metodología *PARADISE* |
| **Características del reconocedor de voz** | |
| **Tipo de reconocimiento** | Habla continua y espontánea |
| **Tipo de frases a reconocer** | Frases largas y cortas |
| **Idiomas que reconoce** | Español de México |
| **Características del corpus** | |
| **Nombre del corpus** | DIMEx100 niños |
| **Edad de los hablantes** | 10-14 años de edad |
| **Cantidad de hablantes** | 100 hablantes |
| **Tipo de habla del corpus** | Grabaciones leidas |
| **Tipo de hablantes** | Hablantes nativos del Español de México |
| | Niños y niñas |

Cuadro 3.5: Características generales del proyecto *Golem-Universum*.

*Maximum Likelihood Estimation* (MLE). Para la medición del desempeño del reconocedor, durante el proceso de reconocimiento de los niños, fue utilizada la técnica de análisis *LPC*, con el fin de estimar las diferencias en las frecuencias de la longitud del tracto vocal de este tipo de hablantes, debido a que en los niños se presentan diferentes formatos dependiendo de la edad, estatura y género; además de este análisis se realizaron comparaciones entre los resultados obtenidos basados en *LPC-CEP* y *MEL-CEP* de las características de los hablantes.

Para la realización de cada uno de los experimentos de evaluación se crearon cinco categorías de hablantes, de acuerdo a la edad de cada uno de ellos; por ejemplo la categoría uno corresponde a los más jóvenes (niños) y la categoría cinco a los más grandes de edad (adultos mayores). El criterio que se utilizó durante el proceso de entrenamiento y prueba del sistema reconocedor de voz fue con respecto al género y edad para cada uno de los tipos de hablantes. La normalización de los datos se llevó a cabo de manera equitativa en cada una de las cinco categorías y el entrenamiento del sistema fue usando siempre la misma cantidad de datos, únicamente la composición de estos era la que cambiaba.

Los resultados obtenidos de los experimentos realizados reflejan que en el rango de edad de 15 a 70 años aproximadamente, la tasa de error en el reconocimiento es considerablemente buena, tomando en cuenta que el reconocedor de voz ha sido entrenado con datos suficientemente representativos del idioma a reconocer; fuera de este rango el error incrementaba dramáticamente. El valor de *word error rate* en niños y adultos mayores fue 170 % y 94 %, respectivamente, siendo más alto que para la mitad de las cinco categorías de los hablantes. Con los resultados de esos experimentos se llegó a la conclusión de que la composición de los datos de entrenamiento influye en el rango de error obtenido.

Finalmente, para explicar los resultados obtenidos se realizaron algunos análisis de género y edad [Wilpon and Jacobsen, 1996].

En este capítulo se dio un panorama general de las investigaciones más relevantes que se han realizado en los últimos años, en el área del ASR específicamente para hablantes de tipo niños; estos estudios se han desarrollado debido al interés de desarrollar aplicaciones con reconocimiento de voz para diversos tipos de hablantes, como se mencionó en el capítulo 1. Además se presentaron los resultados obtenidos al evaluar el desempeño del reconocedor de voz de las evaluaciones realizadas, el cual se encuentra implementado en el sistema *Golem-Universum*, estos resultados originan el interés para realizar esta tesis y por tanto la serie de experimentos empíricos que se realizaron, los cuales son descritos en los siguientes capítulos de esta tesis. Para finalizar, cabe resaltar que de acuerdo a los resultados y conclusiones que se presentaron en cada una de las investigaciones, podemos decir que este trabajo está centrado en estudiar a dos tipos de hablantes (niños y adultos) que presenta varias dificultades cuando se diseña un reconocedor de voz de acuerdo a sus características, principalmente ocurre un alto porcentaje de error en el reconocimiento de voz en niños. Este problema se presenta por la particularidad que presenta este tipo de hablantes, como hemos mencionado en el capítulo anterior de esta tesis.

| **Características del proyecto** | |
|---|---|
| **Objetivo del proyecto** | Mejorar desempeño para frases largas en reconocedores de voz para niños y adultos |
| **Técnicas y algoritmos utilizados** | *Maximum Likelihood Estimation (MLE)*<br>Técnica de análisis *LPC*<br>*LPC-CEP*<br>*MEL-CEP* |
| **Características del reconocedor de voz** | |
| **Tipo de reconocimiento** | Habla continua |
| **Tipo de frases a reconocer** | Frases cortas (dígitos) |
| **Idioma que reconoce** | Danés |
| **Características del corpus** | |
| **Nombre del corpus** | Rafael.0 |
| **Edad de los hablantes** | 8 hasta más de 80 años de edad |
| **Cantidad de hablantes** | 487 hablantes |
| **Tipo de habla del corpus** | Grabaciones leídas |
| **Tipo de hablantes** | Hablantes nativos<br>Niños y Adultos |

Cuadro 3.6: Características generales del proyecto AT&T Bell.

# Capítulo 4

# Modelos acústicos con los corpora *DIMEx100 niños* y *DIMEx100 adultos*

En este capítulo se presenta una descripción de cada uno de los experimentos desarrollados con el propósito de caracterizar los corpora de habla leída: *DIMEx100 niños* y *DIMEx100 adultos*. Principalmente se observan las características fonéticas que presentan y en base a esto se determina que tan buenos son estos recursos para la creación de reconocedores de voz en español de México. Estos experimentos fueron diseñados para conocer la calidad de los modelos acústicos, el cual se aprecia en el desempeño del reconocedor de voz para dos tipos de hablantes: niños y adultos. Para lograr el objetivo de este capítulo se han planteado las siguientes preguntas a las que se les quiere dar respuesta:

1. ¿Cómo se comporta un reconocedor de voz que ha sido creado para un tipo de hablante y que interactúa un tipo de hablante distinto?

2. ¿Cuál es el comportamiento de un reconocedor de voz cuando es entrenado para distintos tipos de hablantes?

3. Cuando es agregada información poco a poco de un tipo de hablante distinto de forma controlada para el entrenamiento, ¿Cuál es el comportamiento del reconocedor de voz?

4. De acuerdo a las características que presentan ambos corpora ¿pueden llegarse a complementar los datos utilizados para el entrenamiento de un reconocedor de voz?

El contenido de este capítulo se encuentra organizado de la siguiente manera: en la sección 4.1 se describen los experimentos que se han propuesto para dar respuesta a las preguntas presentadas. En la sección 4.2 se muestran los resultados de las evaluaciones realizadas en cada uno de los experimentos. En la sección 4.3 se presenta las especificaciones del ambiente de desarrollo en que se llevo a cabo los experimentos descritos en la sección 4.1. Finalmente, en la sección 4.4 se discute la importancia de los resultados de las evaluaciones.

## 4.1. Experimentos

Para realizar los experimentos de este capítulo se utilizaron en su totalidad los 100 usuarios de cada uno de los corpora de habla leída: corpus *DIMEx100 niños* y *DIMEx100 adultos*; estos experimentos son posibles de realizar de acuerdo a como se diseñaron por la característica que tienen los corpus de ser paralelos como lo hemos mencionado en capítulos anteriores. Las combinaciones de datos en los experimentos son posibles de realizarse debido

a que los corpora tienen la característica de que ambos están formados para cada usuario por las mismas oraciones, podemos decir que de alguna manera son simétricos y paralelos en cuanto a la información fonético-acústica que contienen. El conjunto de experimentos diseñados para este capítulo tienen la característica de que el mayor trabajo que se realiza durante el proceso de reconocimiento sea en los modelos acústicos, esto se logra utilizando "malos" modelos de lenguaje. Los experimentos que se presentan en este capítulo han sido diseñados para dar respuesta a cada una de las preguntas que se mencionaron anteriormente, para responder a la primera pregunta se diseñó el experimento Simple, la siguiente pregunta es contestada con el experimento Mixto balanceado, el experimento Mixto no balanceado contesta a nuestra tercer pregunta y finalmente el experimento Mixto no controlado a la cuarta pregunta. A continuación se describen cada uno de estos experimentos.

### 4.1.1. Experimento Simple

Como hemos mencionado este experimento nos ayuda a dar respuesta a la interrogante de ¿Cuál es el desempeño de un reconocedor de voz cuando interactúa con un tipo de hablante distinto para el que fue diseñado? De acuerdo a esta pregunta el objetivo de este experimento es conocer el comportamiento de un reconocedor de voz cuando es entrenado y evaluado con un mismo tipo de hablante o el caso cuando es evaluado con un hablante diferente al que fue entrenado, además de conocer la calidad de los modelos acústicos.

Cada uno de los reconocedores de voz creados en este experimento se entrenaron con 99 usuarios de uno de los corpora *DIMEx100 niños* o del *DIMEx100 adultos*, creándose un total de 200 reconocedores de voz que fueron evaluados con un usuario del mismo tipo de hablante con el que fue entrenado y con uno distinto, del proceso descrito anteriormente se obtuvieron

un total de 400 evaluaciones. Para la evaluación siempre se selecciona el mismo usuario de ambos corpora debido a que son los mismos pues ambos contienen las mismas frases con la diferencia de que son emitidas por dos tipos de hablantes distintos. En la figura 4.1 se ilustra cómo se llevó a cabo el proceso de prueba y entrenamiento. Esta figura muestra los datos que son tomados para prueba y entrenamiento en la primera iteración del experimento simple, para el proceso de validación cruzada de cada uno de los 100 usuarios de los corpora *DIMEx100 niños* y *DIMEx100 adultos*.

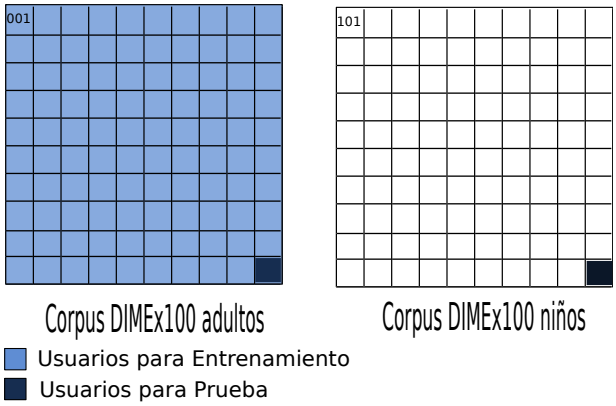


Figura 4.1: Esquema de validación cruzada para el experimento Simple.

Cabe mencionar que el proceso que se siguió para el entrenamiento y evaluación en este experimento ha sido el mismo que se realizó en [Pineda et al., 2009] pero con la diferencia de que ahí solo se utilizaron datos del corpus *DIMEx100 adultos* para el entrenamiento y la evaluación.

### 4.1.2. Experimento Mixto balanceado

Una vez realizado el experimento Simple con ambos corpora y observar el comportamiento que presenta un reconocedor de voz para un tipo de usuarios, probado con dos tipos de usuarios, niños y adultos, el siguiente experimento se realizó para dar respuesta a la pregunta ¿Cómo se comporta un reconocedor de voz que está hecho para distintos tipos de hablantes?, por lo tanto para este experimento los datos de entrenamiento consistieron de ambos corpora (198 usuarios) descartando siempre dos usuarios para la evaluación. La evaluación se realizó con cada uno de los 200 usuarios de los ambos corpora.

Esta serie de evaluaciones se realizaron como ya se ha mencionado con el objetivo de conocer que tan buenos son los corpora que se tienen disponibles para estos experimentos como recursos para la creación de reconocedores de voz, debido a que estamos interesados en conocer el desempeño que se tiene cuando los datos de entrenamiento contienen toda la información que tenemos disponibles de los dos tipos de hablantes. En la figura 4.2 se muestra un esquema de la primera iteración que es realizada en este experimento donde se observa que son tomados 198 usuarios como datos de entrenamiento (99 usuarios de cada uno de los corpora) y el primer usuario de ambos corpora para evaluar al reconocedor de voz, para este caso especifico el primer usuario. Cabe mencionar que son tomados dos usuarios, uno de cada corpora ya que cada uno de los usuarios representa el caso para la evaluación de un niño y de un adulto respectivamente.

### 4.1.3. Experimento Mixto no balanceado

El experimento Mixto no balanceado como su nombre lo indica se refiere a que la cantidad de datos utilizados para el entrenamiento no son balanceados, se van agregando poco a poco

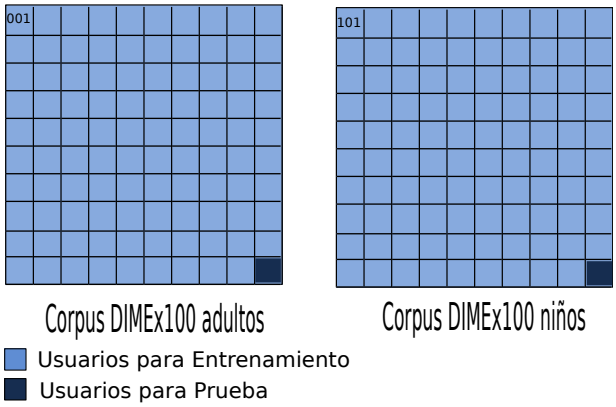


Figura 4.2: Esquema de validación cruzada para el experimento Mixto balanceado

durante cada ciclo de entrenamiento. Los experimentos anteriores han sido el punto de partida para la realización de este experimento, debido a que ahora la inquietud que se tiene es conocer el comportamiento de un reconocedor de voz cuando se le proporciona poco a poco información de un tipo de hablante y la forma en que está diseñado este experimento ayuda a dar respuesta a esta interrogante.

El proceso de selección de los datos para el entrenamiento consiste en tener como base 99 usuarios de un tipo de hablante e ir agregando durante cada iteración un usuario más diferente al que se tiene como base, para tener al final de las iteraciones 198 hablantes. Las evaluaciones realizadas consisten al igual que en los experimentos anteriores, en evaluar con un usuario niño y con un usuario adulto. En la figura 4.3 se muestra un ejemplo de los datos que son utilizados para el entrenamiento de cada uno de los 100 reconocedores de voz que han sido creados para este experimento.

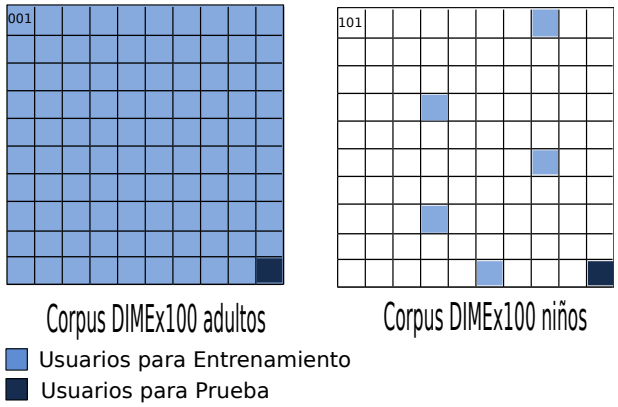


Figura 4.3: Esquema de validación cruzada para el experimento Mixto no balanceado

### 4.1.4. Experimento Mixto controlado

Este experimento fue diseñado para ayudar a contestar la pregunta ¿Es posible complementar los datos de los corpora para el entrenamiento de un reconocedor de voz?. Los datos considerados en este experimento tienen la particularidad de ser complementarios por las propiedades de los corpora de donde son tomados. Los usuarios de entrenamiento se seleccionan de manera que siempre se tengan todas las oraciones que componen al corpus con la diferencia del tipo de hablante que las está emitiendo; por ejemplo se toma un corpus en este caso del corpus *DIMEx100 adultos* los primeros 20 usuarios y del corpus *DIMEx100 niños* los siguientes 79 usuarios (ver la figura 4.4).

## 4.2. Resultados

La medida estándar utilizada para medir los porcentajes de error durante el proceso de reconocimiento de voz es usando la tasa de error de palabra (*word error rate*); para los experimentos de este capítulo se utiliza esta medida y únicamente se reporta los resultados

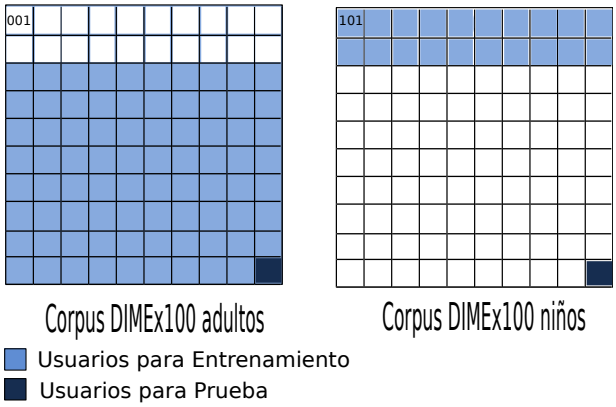


Figura 4.4: Esquema de validación cruzada para el experimento Mixto controlado.

obtenidos a nivel de palabra. Esta medida de evaluación es estándar en sistemas de reconocimiento de voz. Para realizar el cálculo de *word error rate* se utiliza la siguiente fórmula:

$$word\ error\ rate = 100\ \frac{Inserciones\ +\ Sustituciones\ +\ Eliminaciones}{Total\ de\ palabras\ de\ la\ transcripcion\ correcta} \tag{4.1}$$

Debe de tomarse en cuenta que a medida que el valor de *word error rate* en cada uno de los experimentos sea menor, significa que desempeño del reconocimiento es mejor, por el contrario si el valor es muy grande el reconocimiento es malo.

### 4.2.1. Evaluación del experimento Simple

En la figura 4.5 se muestra el diagrama donde se observa el comportamiento de todas las evaluaciones realizadas en este experimento. Como se puede observar en la gráfica el mejor reconocimiento se obtiene cuando se entrena con adultos y se evalúa con adultos; esto nos dice que las voces de los adultos presentan menos variaciones para ser reconocidas.

Como caso especifico de las evaluaciones que se realizaron se muestra en la gráfica de la figura 4.6 los resultados de las evaluaciones para dos tipos de hablantes cuando los modelos

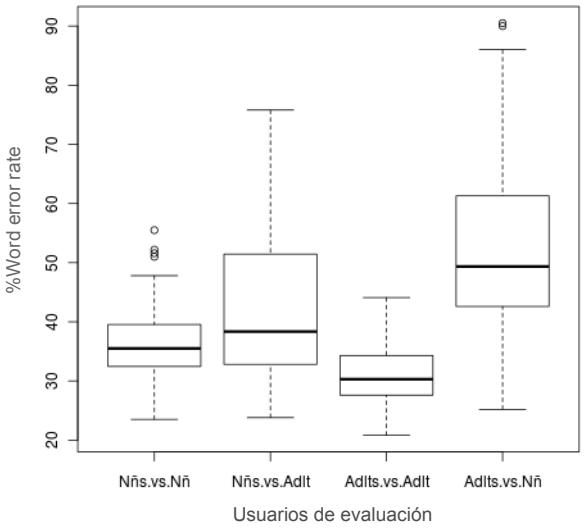


Figura 4.5: Diagrama comparativo de las evaluaciones en el experimento Simple.

acústicos han sido creados con el corpus *DIMEx100 niños*; hasta el momento los resultados que se observan en la gráfica indican que el reconocimiento es mejor cuando se evalúa con el mismo tipo de hablante que se crearon los modelos acústicos.

En el cuadro 4.1 se muestra todos los resultados de las evaluaciones realizadas en este experimento. Los datos que se muestran en la tabla nos indican que el tipo de hablante adulto es el más estable para ser reconocido. Este resultado obtenido en este experimento corresponde con el obtenido en [Pineda et al., 2009].

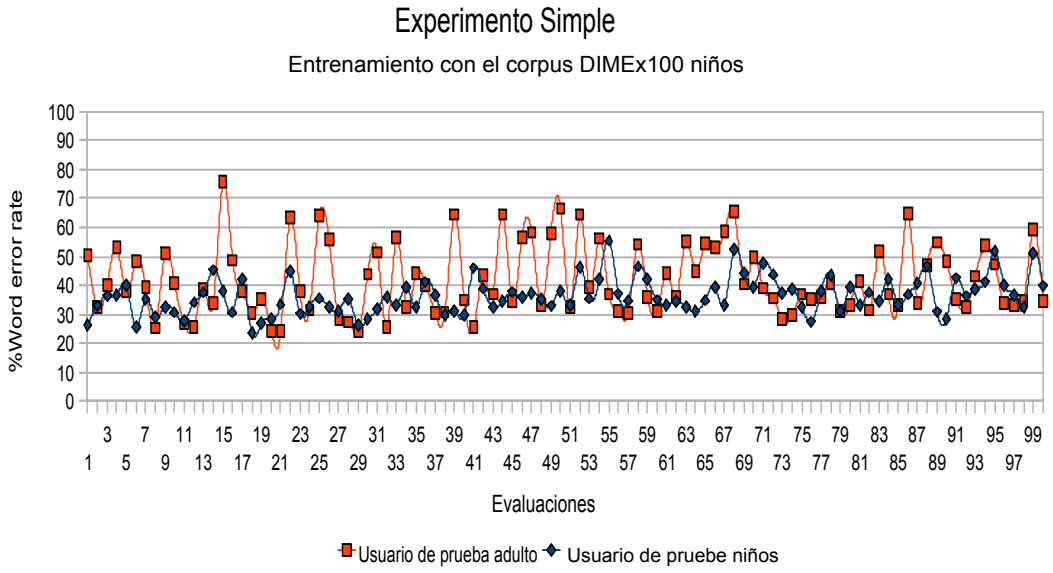


Figura 4.6: Gráfica de las evaluaciones del experimento Simple entrenado con el corpus *DIMEx100 niños*.

| **Corpus de entrenamiento** | **Corpus de evaluación** | **Promedio *word error rate*** |
|---|---|---|
| *DIMEx100 adultos* | *DIMEx100 adultos* | 30.87 % |
| *DIMEx100 niños* | *DIMEx100 niños* | 36.32 % |
| *DIMEx100 adultos* | *DIMEx100 niños* | 52.38 % |
| *DIMEx100 niños* | *DIMEx100 adultos* | 41.92 % |

Cuadro 4.1: Resultados de las evaluaciones realizadas en el experimento Simple.

### 4.2.2. Evaluación del experimento Mixto balanceado

Los resultados obtenidos de las evaluaciones realizadas en este experimento se pueden observar en la gráfica de la figura 4.7. De acuerdo a los resultados obtenidos en cada una de las evaluaciones, en promedio el mejor desempeño del reconocedor de voz se obtienen cuando es evaluado con adultos con un 31.19 % en comparación con la evaluación realizada con el corpus *DIMEx100 niños* en la que obtuvo 36.11 %; además de que el mejor desempeño se obtiene cuando es evaluado con adultos con un 21.2 % de *wor error rate*. De acuerdo con los resultados obtenidos de este experimento y del experimento Simple se puede decir que el reconocimiento de voz en hablantes adultos es hasta el momento el que mejor desempeño ha mostrado. Sin embargo, al mezclar ambos corpora el desempeño muestra mejoramiento.

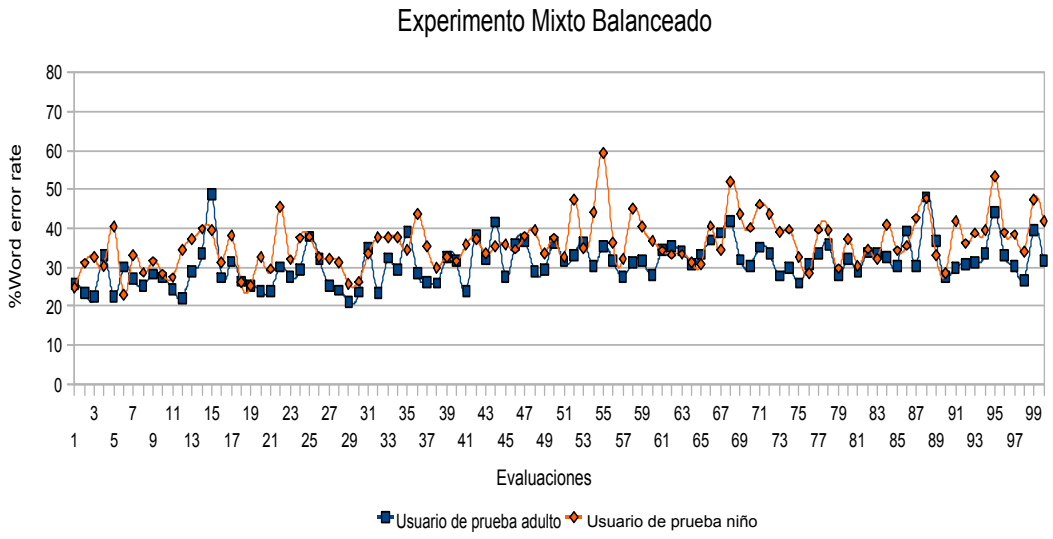


Figura 4.7: Gráfica de las evaluaciones realizadas en el experimento Mixto balanceado.

En la diagrama de la figura 4.8 se puede observar cómo se comportaron los modelos

acústicos para las evaluaciones realizadas en este experimento.

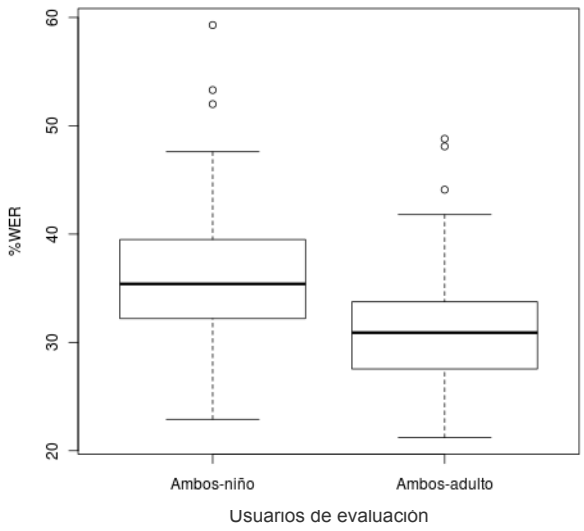


Figura 4.8: Diagrama comparativo de las evaluaciones en el experimento Mixto balanceado.

En el siguiente cuadro se resumen todos los promedios para todas las evaluaciones que se realizaron en este experimento.

| **Corpora de entrenamiento** | **Corpus de evaluación** | **Promedio *word error rate*** |
|---|---|---|
| *DIMEx100 adultos + DIMEx100 niños* | *DIMEx100 adultos* | 31.19 % |
| *DIMEx100 adultos + DIMEx100 niños* | *DIMEx100 niños* | 36.11 % |

Cuadro 4.2: Resultados de las evaluaciones realizadas del experimento Mixto balanceado.

### 4.2.3. Evaluación del experimento Mixto no balanceado

Una vez que hemos realizado los experimentos previamente explicados, de acuerdo al comportamiento de cada uno de los resultados, para el caso del experimento Simple que nos da una base del desempeño para dos tipos de hablantes, el experimento Mixto balanceado de acuerdo a los resultados que se obtuvieron nos dice que si es bueno combinar los corpora para el entrenamiento, por lo tanto con este experimento se encontraron las cantidades necesarias de entrenamiento que nos dan el mejor desempeño en el reconocimiento, con la estrategia de ir agregando un usuario más en cada evaluación.

El comportamiento que se observó en las evaluaciones realizadas para el caso donde se tienen 99 usuarios del corpus *DIMEx100 adultos* para el entrenamiento en la primera iteración e ir agregando un usuario del corpus *DIMEx100 niños* en las siguientes iteraciones se pueden apreciar en la gráfica 4.9. La curva de aprendizaje que se muestran en la gráfica corresponde al caso de la evaluación con un usuario niño y un usuario adulto respectivamente. Como se puede observar para el caso de la evaluación con niños a medida que se van agregando datos del mismo usuario con el que es evaluado el reconocimiento mejora hasta que se mantiene casi constante.

Para el caso contrario de esta evaluación donde es agregado un usuario del corpus *DIMEx100 adultos* y se tienen 99 usuarios del corpus *DIMEx100 niños* como base, no presentan variaciones considerables, podríamos decir que casi se mantienen constante el valor de *word error rate* durante las evaluaciones (ver figura 4.10) de los dos usuarios (niño y adulto) pues no presenta cambios tan notables en ambas evaluaciones. El conjunto de evaluaciones que se realizaron para este experimento únicamente fueron para el usuario 15 de los corpora, por tal motivo este experimento únicamente fue exploratorio, para poder realizar un análisis

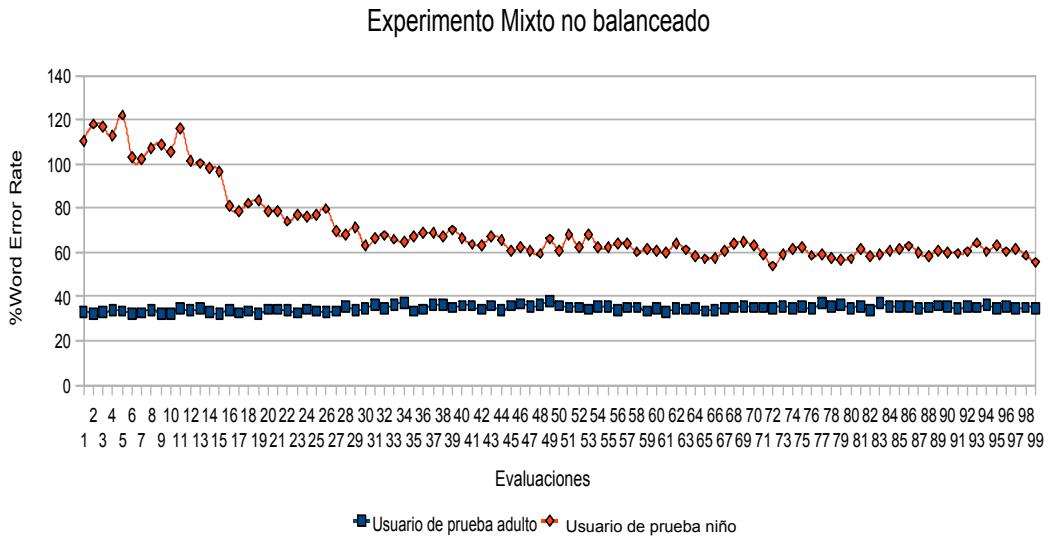


Figura 4.9: Curva de aprendizaje de las evaluaciones del experimento Mixto no balanceado agregando usuarios del corpus *DIMEx100 niños*.

mucho más completo es necesario realizar la evaluación para cada uno de los usuarios de los corpora.

### 4.2.4. Evaluación del experimento Mixto controlado

De acuerdo a las evaluaciones realizadas en el experimento Mixto no balanceado y de la pregunta que se quiere responder al realizar este experimento, de determinar si los datos de ambos corpora pueden llegan a ser complementarios se puede observar en la gráfica 4.11 que tasa de error de la palabra no muestran diferencias tan considerables para las dos evaluaciones que se realizaron con el usuario 14 de los corpora.

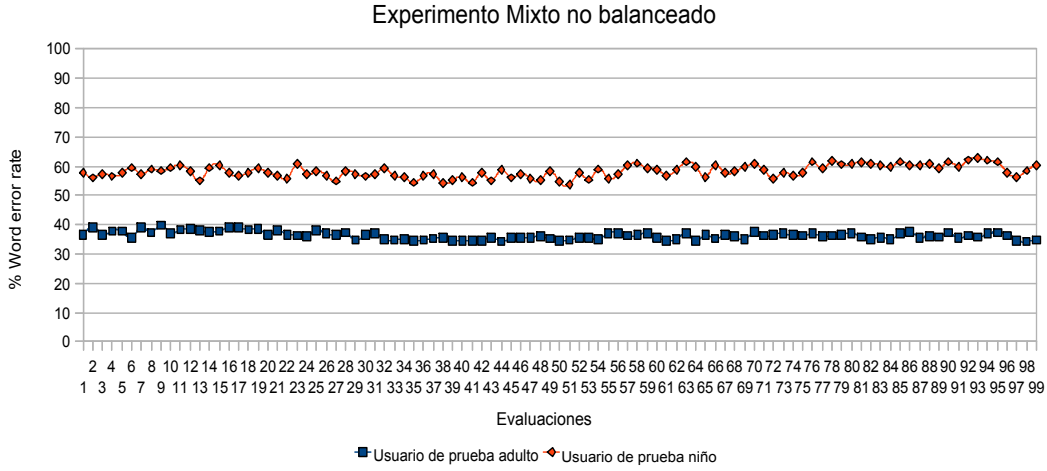


Figura 4.10: Curva de aprendizaje de las evaluaciones del experimento Mixto no balanceado agregando usuarios del corpus *DIMEx100 adultos*.

Como podemos ver en la curva de aprendizaje del experimento Mixto controlado, el reconocimiento es peor cuando la cantidad de datos de entrenamiento corresponden a un solo tipo de hablante, pero a medida que se van combinando los datos de entrenamiento manteniendo siempre las propiedades del corpus, es decir que siempre se tengan las misma frases que originalmente contiene el corpus, la curva comienza a mejorar manteniéndose sin cambios tan drásticos en la tasa de error de la palabra. Como ya se mencionó anteriormente las evaluaciones únicamente se realizaron con el usuario 14 de ambos corpus ya que este experimento ha sido realizado únicamente con fines exploratorios. Sin embargo, con este resultado se sigue encontrando que combinado ambos corpora se obtienen mejores resultados, además de que la combinación de datos debe de ser preferente lo más equitativa posible.

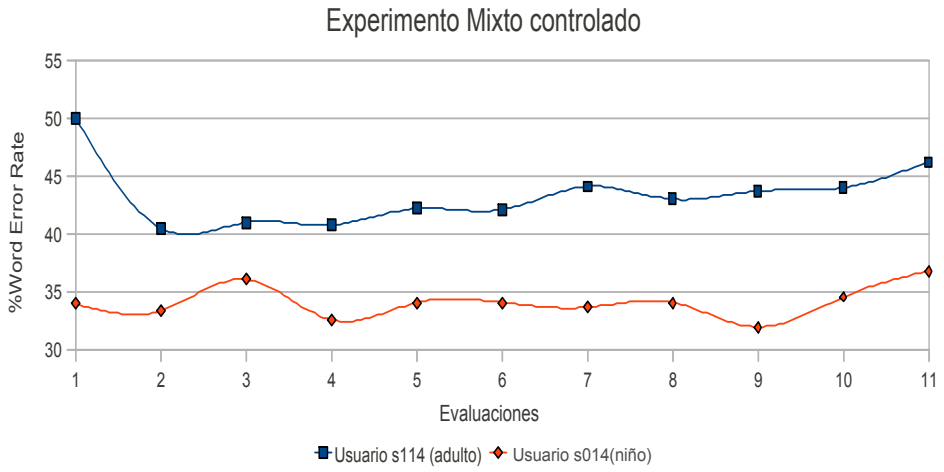


Figura 4.11: Curva de aprendizaje de las evaluaciones del experimento Mixto controlado.

## 4.3. Ambiente de desarrollo

Para la creación de los reconocedores de voz se utilizó el software *Sphinx3*, y como se ha mencionado a lo largo de este capítulo los corpora de habla leída *DIMEx100 niños* y *DIMEx100 adultos* en el nivel de etiquetación ortográfica (en el capítulo 3 se describió más a detalle las características principales de cada uno de ellos). La decisión de utilizar *Sphinx3* para el entrenamiento de los reconocedores de voz, es debido a que ya se contaba con la disponibilidad de este recurso y anteriormente había sido implementado[1]. El proceso de entrenamiento de un reconocedor de voz consiste en que el reconocedor aprenda todos los sonidos que le son indicados, en este caso por medio de los cada uno de los usuarios que forman parte de los corpora, para que pueda identificar la secuencia de unidades más probable cuando es probado; este proceso también es conocido como decodificación[Pérez Pavón, 2006].

[1]Pérez Pavón Elia Patricia (2006). Construcción de un reconocedor de voz utilizando Sphinx y el corpus DIMEx-100

Los modelos acústicos de los experimentos desarrollados en este capítulo se crearon con la herramienta *SphinxTrain*, los cuales presentan las siguientes características:

- Modelo oculto de Markov de 3 estados, continúo.
- 8 gaussianas por estado de cada modelo.
- Se crearon trifonemas.

Todos los ciclos de entrenamiento y evaluaciones de los reconocedores de voz, fueron desarrollados en el sistema operativo Linux, en la distribución de Ubuntu 10.04 y 9.10. Las características de las máquinas donde se realizaron los entrenamientos y evaluaciones de los reconocedores de voz, se muestran a continuación:

- Procesador Intel core2 Duo, CPU a 2.20 GHz, memoria RAM a 2GB
- Procesador Intel Core i3, CPU a 3.07 GHz, memoria RAM a 3.7GB
- 4 Procesadores Intel(R) Xeon(R), CPU X3450 a 2.67 GHz, memoria RAM a 16GB

En cada una de estas máquinas el tiempo de procesamiento de los datos fue diferente debido a las características tan diversas en las que se realizó cada experimento, por lo tanto en promedio el tiempo requerido para realizar toda la serie de experimentos que se presentan en este capítulo fue alrededor de 672 hrs.

Los Modelos de lenguaje para los experimentos de este capítulo son de 3-gramas siendo únicos para cada reconocedor de voz, debido a que fueron creados de acuerdo a las características de cada uno de los experimentos; de igual forma los modelos acústicos y diccionario de pronunciación son diferentes para cada reconocedor de voz.

La evaluación de los reconocedores de voz se llevó a cabo en modo *bach*, esto significa que para probarlos se utilizaron las grabaciones que contienen los corpora; podemos decir que las evaluaciones de este capítulo fueron en un mundo perfecto por que los audios se encuentran cuidados y no contienen demasiado ruido. El programa utilizado para el proceso de evaluación fue *sclite* que pertenece a las herramienta de evaluación NIST.

## 4.4. Discusión

Como mencionamos al principio de este capítulo cada uno de los experimentos que hemos presentado, han ayudado a responder las preguntas listadas al inicio de este capítulo, las cuales han sido propuestas para lograr el objetivo principal de esta tesis el cual es caracterizar los corpora de habla leída *DIMEx100 niños* y *DIMEx100 adultos*. El experimento Simple ha mostrado el comportamiento de cada tipo de hablante en un reconocedor de voz, mientras que el experimento Mixto balanceado nos ha dicho que es bueno combinar los datos de los corpora pero no las mejores proporciones que se necesitan, por otro lado los experimento Mixto no balanceado y Mixto controlado nos indican el comportamiento de un reconocedor cuando se le agrega información poco a poco de un corpus y si se complementan.

De acuerdo a los resultados que se obtuvieron en los experimentos de este capítulo, podemos concluir que lo mejor es combinar los datos de ambos corpora, para obtener un buen reconocedor de voz, pero siempre buscando que las proporciones sean equitativas de acuerdo al número de usuarios. Además en base a los resultados que arrojaron las evaluaciones, el corpus que presenta mayor estabilidad tanto para el entrenamiento como evaluación es el corpus *DIMEx100 adultos*, siendo el menos estable el corpus *DIMEx100 niños*, pero este

resultado era de esperarse debido a las características tan particulares que presenta este tipo de hablante con el que fue creado el corpus.

Los experimentos que se han realizado en este capítulo han sido la base para analizar el comportamiento de los modelos acústicos cuando son creados con corpora de voz leída. De acuerdo a los resultados que se obtuvieron al evaluar los modelos acústicos, la interrogante que ahora ha surgido es sobre el comportamiento de estos modelos pero ahora siendo evaluados con voz espontánea. En el siguiente capítulo se presentan los experimentos evaluando con voz espontánea para dar respuesta a esta nueva interrogante que se tiene.

# Capítulo 5

# Modelos acústicos con el corpus *Golem-Universum*

Un reconocedor de voz debe de presentar un buen desempeño cuando es probado en condiciones reales; es por esta razón que ha surgido el interés de conocer que tan buenos son los modelos acústicos creados para un ambiente que presenta condiciones ideales y que son evaluados con voz espontánea, por tal motivo en este capítulo se presentan los experimentos donde es utilizado el corpus *Golem-Universum* tanto como dato de entrenamiento y de evaluación. Cabe mencionar que los experimentos de este capítulo surgen de haber analizado los resultados de las evaluaciones correspondientes a los 510 reconocedores de voz que fueron diseñados para los experimentos del capítulo anterior. El objetivo que se tiene con estas nuevas evaluaciones es probar los modelos acústicos evaluado un reconocedor de voz con habla espontánea, simulando que el hablante interactua directamente con el sistema por

medio de un micrófono (pruebas en vivo). Adicionalmente de caracterizar de los tres recursos lingüísticos que se tienen disponibles, se quiere saber que tan robustos son los modelos acústicos creados y si el desempeño en el reconocimiento muestra mejoras.

En la sección 5.1 se presentan los resultados de las evaluaciones realizadas a los modelos acústicos que se crearon en los experimentos explicados en el capítulo anterior, pero ahora evaluándolos con el corpus de voz espontánea *Golem-Universum*. En la sección 5.2 se explican los experimentos donde es agregado como datos de entrenamiento el corpus de voz espontánea, además del corpus *DIMEx100 adultos* y *DIMEx100 niños*; las evaluaciones a estos nuevo modelos acústicos también son realizados con voz espontánea. En la sección 5.3 se presenta una breve discusión de los resultados obtenidos.

## 5.1. Experimentos

Los experimentos de esta sección han utilizado parte de la evaluación del corpus de habla espontánea *Golem-Universum* para probar los modelos acústicos que se crearon con datos de los corpora *DIMEx100 niños* y *DIMEx100 adultos*, además de que se realizan experimentos entrenando con voz espontánea. Los experimentos de esta sección buscan caracterizar al corpus de habla espontánea realizando nuevos ciclos de entrenamiento dinámicos de acuerdo a los resultados obtenidos de las evaluaciones hechas a los reconocedores de voz con habla leída que fueron descritos en el capítulo anterior; estos resultados son importantes para realizar estos nuevos entrenamientos y evaluaciones por que dan una referencia para saber si es bueno complementar la información de entrenamiento con el corpus *Golem-Universum*.

Para lograr el objetivo de estos experimentos, se formularon preguntas que parten de las

planteadas en la sección anterior, las cuales han quedado como se muestran a continuación:

- ¿Cuál es el comportamiento de un reconocedor de voz creado para un tipo de hablante, cuando interactua con hablantes niños en condiciones reales?
- Un reconocedor de voz que es entrenado para distintos tipos de hablantes (niños y adultos), ¿Cuál es su comportamiento si interactua con hablantes niños en condiciones reales?
- Si es agregada información poco a poco de un tipo de hablante de forma controlada para el entrenamiento, ¿Cómo se comporta el reconocedor de voz cuando es probado con hablantes niños en condiciones reales?
- Tomando en cuenta las características de los corpora de habla leída, cuando son complementados los datos de entrenamiento, ¿Cuál es el desempeño del reconocedor de voz cuando se evalúa con hablantes niños en condiciones reales?
- Si es agregada información del corpus de habla espontánea para el entrenamiento, ¿el desempeño del reconocedor de voz mejora para el reconocimiento en condiciones reales?

**Corpus *Golem-Universum***

Este corpus de habla espontánea se encuentra formado por un conjunto de 40 evaluaciones realizadas en el módulo *Golem-Universum*, donde los usuarios (niños) han interactuado con el sistema en tiempo real, bajo condiciones reales (sin cuidar las grabaciones y con el ruido propio del ambiente). Bajo este contexto entendemos que las evaluaciones que integran al corpus *Golem-Universum* se refiere a pruebas que se realizaron al módulo *Golem-Universum*,

las palabras que integran a este corpus en su mayoría son cortas referentes al contexto del juego "Adivina la carta".

### 5.1.1. Experimentos evaluando con voz espontánea

Al realizar este experimento se da respuesta a las cuatro primeras preguntas que fueron planteadas anteriormente, que básicamente todas ellas consisten en evaluar modelos acústicos que han sido creados con diferentes proporciones de entrenamiento de los corpora *DIMEx100 niños* y *DIMEx100 adultos* y que es evaluado con voz espontánea de hablantes niños. Por lo mencionado anteriormente este experimento consiste en evaluar todos los modelos acústicos pero ahora con voz espontánea, que se crearon en los experimentos Simple, Mixto balanceado, Mixto no balanceado y Mixto complementado descritos en el capítulo anterior, con el objetivo de analizar el comportamiento que presentan los modelos acústicos cuando la evaluación se realiza simulando un ambiente real. Específicamente para realizar esta evaluación se utilizó parte de las evaluaciones del corpus *Golem-Universum* para ser los usuarios de prueba.

Este experimento es muy importante ya que los resultados de las evaluaciones nos dicen que tan buenos son los modelos acústicos cuando son probados en un ambiente real, como es el caso del módulo *Golem-Universum*, con palabras espontáneas dentro del contexto del juego "Adivina la carta". Para este experimento como ya hemos mencionado se ha utilizado para la evaluación los modelo acústico de cada una de los experimentos, pero con la diferencia de que el modelo de lenguaje y el diccionario de pronunciación pertenecen a reconocedor de voz del juego "Adivina la carta".

### 5.1.2. Experimentos entrenando con voz espontánea

Una vez realizado el experimento de las evaluaciones con voz espontánea, se seleccionaron los mejores reconocedores de voz en base a los resultados obtenidos de los experimentos de la sección 5.1.1. Posteriormente de la selección se realizó nuevamente el proceso de entrenamiento pero ahora agregando datos del corpus *Golem-Universum*, con el objetivo de ver si los datos de los tres corpora se complementan y como consecuencia si mejora el reconocimiento. A continuación se muestran las proporciones de entrenamiento para este experimento:

- 99 usuarios del corpus *DIMEx100 niños* + usuarios corpus *Golem-Universum*
- 99 usuarios del corpus *DIMEx100 adultos* + usuarios corpus *Golem-Universum*
- 99 usuarios del corpus *DIMEx100 niños* + 99 usuarios del corpus *DIMEx100 adultos* + usuarios corpus *Golem-Universum*
- 99 usuarios de corpus *DIMEx100 adultos* + 85 usuarios de corpus *DIMEx100 niños* (seleccionados aleatoriamente) + usuarios corpus *Golem-Universum*
- 40 usuarios del corpus *DIMEx100 niños* + 59 usuarios de corpus *DIMEx100 adultos* + usurios corpus *Golem-Universum*

La importancia de este experimento consiste en observar que tanto puede beneficiar o afectar el agregar datos de voz espontánea para la creación de modelos acústicos, cuando estos son evaluados en un ambiente real con voz espontánea específicamente de un hablante niño. Además con las evaluaciones realizadas en este experimento se da respuesta a la pregunta: ¿Cual es el desempeño del reconocedor de voz cuando es agregada información del corpus de habla espontánea para el entrenamiento?

## 5.2. Resultados

De igual manera que en los resultados presentados en el capítulo 4, los resultados de este capítulo se utiliza la medida de tasa de error de palabra (*word error rate*) y se presentan únicamente los porcentajes obtenidos a nivel de palabra; la ecuación (4.1) es la que se utiliza para calcular este valor. A continuación se presentan los resultados de los experimentos utilizando el corpus de habla espontánea *Golem-Universum*.

### 5.2.1. Evaluaciones con voz espontánea

En la gráfica de la figura 5.1 se muestran los resultados obtenidos de las evaluaciones del experimento Simple para los dos casos que se consideraron en este experimento: entrenamiento con 99 usuarios del corpus *DIMEx100 niños* y entrenamiento con 99 usuarios del corpus *DIMEx100 adultos*; de acuerdo a los resultados que se obtuvieron en estos experimentos con respecto a los resultados de la evaluación con voz leída del capítulo 4 sección 4.2.1, el reconocimiento ha mejorado considerablemente para ambos casos.

De acuerdo a los resultados obtenidos al evaluar los modelos acústicos del experimento Simple, se siguen obteniendo mejores resultados cuando el entrenamiento se realiza con el corpus *DIMEx100 adultos*, alcanzando un promedio de *word error rate* de todas la evaluaciones de 28.1 %, por el contrario para el otro caso donde se entrena con el corpus *DIMEx100 niños* se obtuvo un porcentaje de 28.98 %.

Para el caso del experimento Balanceado los resultados de las evaluaciones han arrojado buenos resultados con un promedio de *word error rate* de 27.84 % lo que indica que los modelos acústicos que se crearon en este experimento son robustos y cuando se evalúa

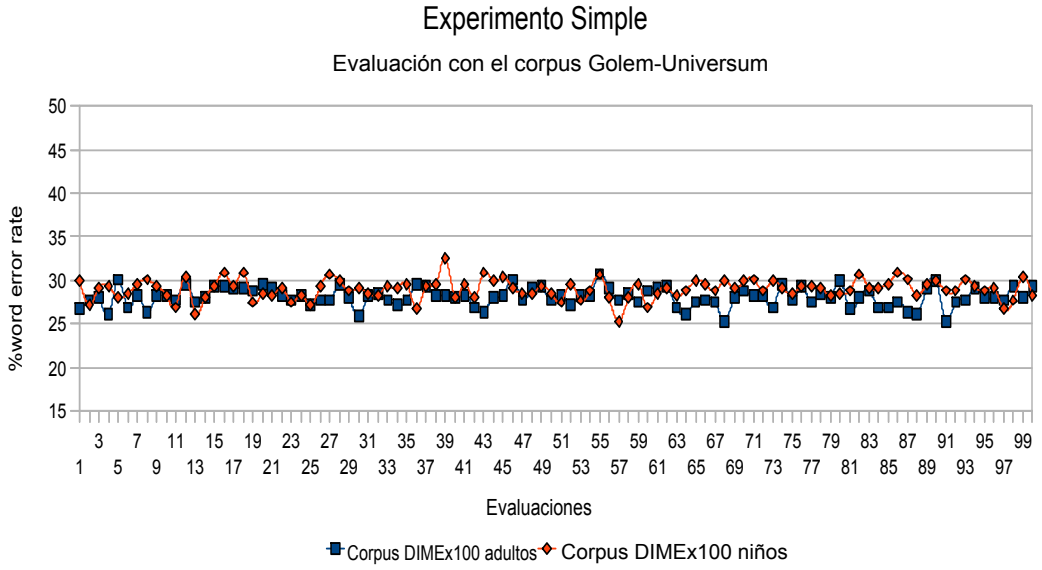


Figura 5.1: Gráfica del experimento Simple evaluado con el corpus *Golem-Universum*.

con voz espontánea se mejora considerablemente el reconocimiento; además que es bueno combinar los datos de entrenamiento, como se había mostrado con los datos de la evaluación en el capítulo 4 sección 4.2.2. En la figura de la gráfica 5.2 se observa el comportamiento de los modelos acústicos creados en el experimento Mixto balanceado cuando son evaluados con el corpus de habla espontánea *Golem-Universum*. Para observar mejor el comportamiento de los resultados que se observaron en la gráfica de la figura 5.2, se presenta el diagrama de la figura 5.3.

Los resultados que se obtuvieron al ser evaluados los modelos acústicos del experimento Mixto no balanceado que son los mismos que se utilizaron para las evaluaciones del capítulo 4 sección 4.1.3, han mostrado un buen desempeño durante el reconocimiento, lo cual indica

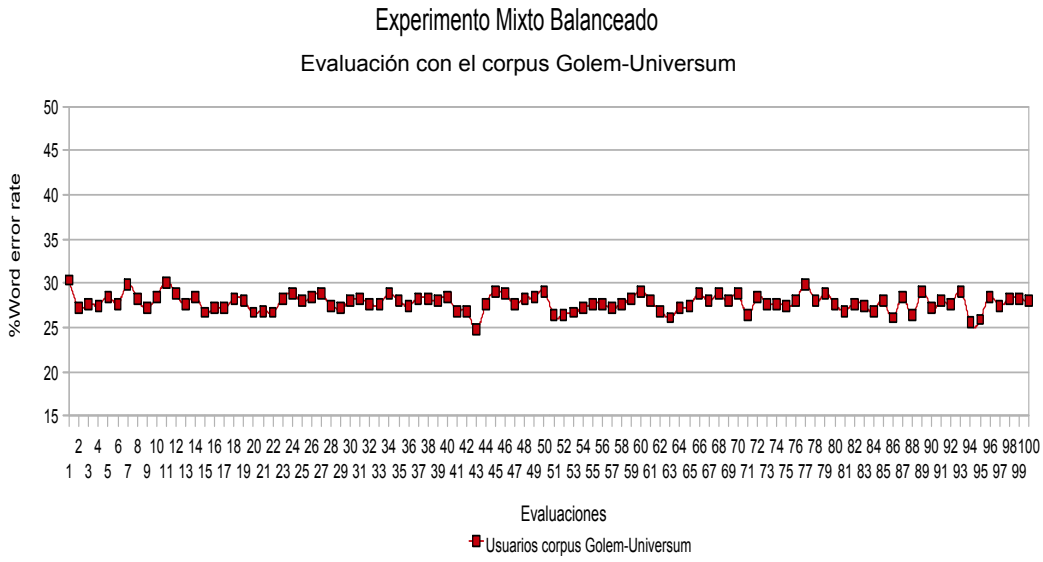


Figura 5.2: Gráfica del experimento Mixto balanceado evaluado con el corpus *Golem-Universum*.

que son de una calidad aceptable. La curva de aprendizaje donde se ilustra el resultado de estas evaluaciones para el caso donde se van agregando usuarios del corpus *DIMEx100 niños* se muestra en la figura 5.4, el caso donde se van agregando usuarios del corpus *DIMEx100 adultos* la curva de aprendizaje se muestra en la figura 5.5.

Como se puede observar en ambas gráficas de los dos casos que se tienen en este experimento, el porcentaje de *word error rate* de cada evaluación se mantienen sin cambios tan drásticos, casi constantes para la mayoría de las evaluaciones. Solo se observa para el caso donde se agregan datos del corpus *DIMEx100 niños* (ver figura 5.3) la onceaba evaluación presenta un aumento muy drástico con un *word error rate* de 46.1 %.

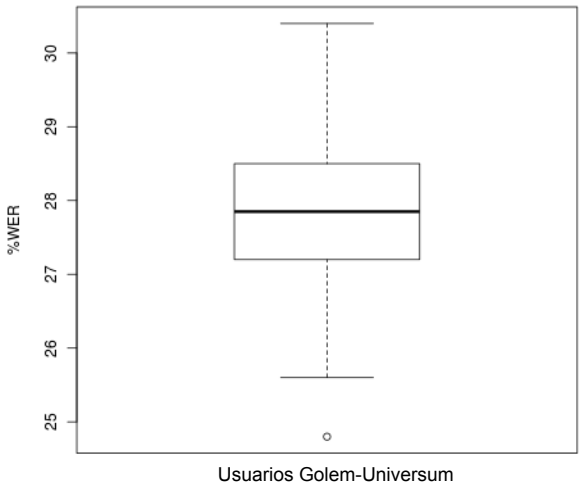


Figura 5.3: Diagrama del experimento Mixto balanceado evaluado con el corpus *Golem-Universum*.

Finalmente, el resultado del experimento Mixto controlado cuando es evaluado con voz espontánea se presenta en la figura 5.6, lo que podemos concluir de esta serie de evaluaciones tomando como referencia los resultados de las evaluaciones con voz leída los modelos acústicos han presentado un mejor reconocimiento cuando se evalúa con voz espontánea.

Para todos los experimentos que se presentan en esta sección en el apéndice A se muestran gráficas donde se muestran los resultados de todas las evaluaciones que se realizaron en cada experimento, utilizando los tres recursos lingüísticos que se tienen disponibles para esta investigación. En cada una de las graficas se puede comparar el desempeño de acuerdo al tipo de hablante con el que fue evaluado.

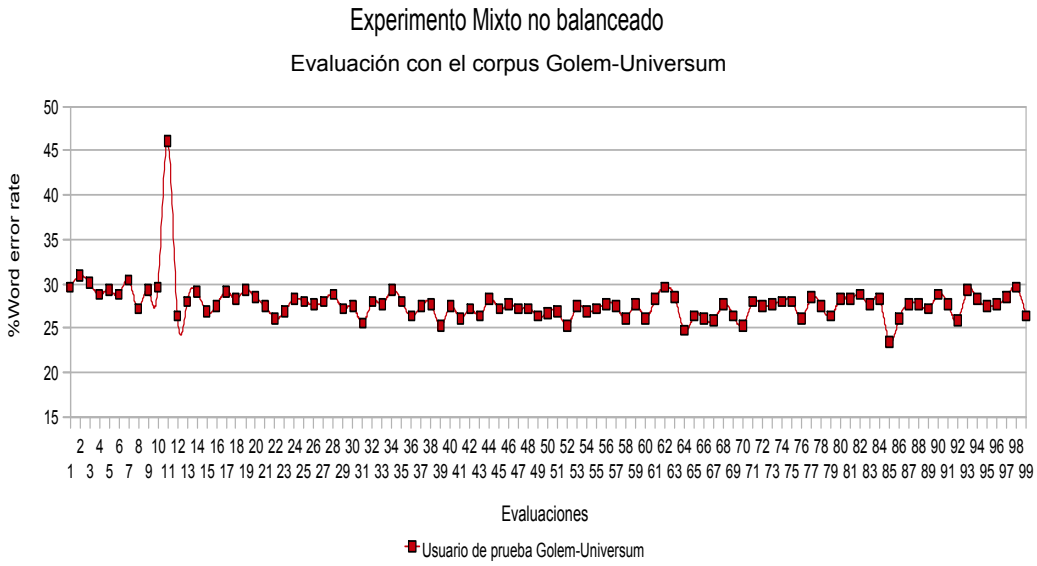


Figura 5.4: Curva de aprendizaje del experimento Mixto no balanceado agregando usuarios del corpus *DIMEx100 niños*.

### 5.2.2. Evaluaciones entrenado con voz espontánea

Los primeros resultados que presentamos de las evaluaciones al agregar datos de entrenamiento del corpus *Golem-Universum* se muestran en el cuadro 5.1, los cuales son nuestras evaluaciones de referencia para comparar y realizar las siguientes evaluaciones, además de tomar el criterio de selección de los mejores resultados del experimento de la sección 5.2.1. El porcentaje de *word error rate* para las evaluaciones es bajo en comparación con el que se presenta cuando no se ha entrenado con el corpus de habla espontánea. El mejor valor se presenta cuando el entrenamiento se realiza con datos de niños únicamente (corpus *DIMEx100 niños* + corpus *Golem-Universum*). Este resultado nos indica que un reconocedor entrenando y evaluando para un mismo tipo de hablante, en este caso niños, el reconocimiento mejora.

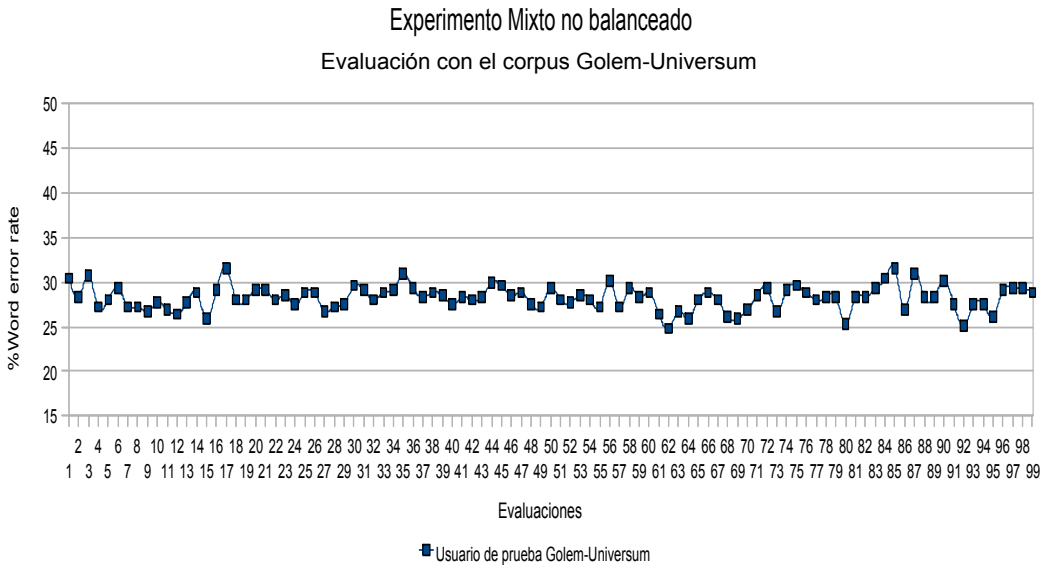


Figura 5.5: Curva de aprendizaje del experimento Mixto no balanceado agregando usuarios del corpus *DIMEx100 adultos*.

Estos experimentos son exploratorios para poder tener una mejor conclusión lo ideal sería realizar esta prueba para todos los 300 reconocedores de voz que se tienen disponibles con estas características.

Para conocer que tanto mejoró el porcentaje de *word error rate*, en la gráfica de la figura 5.7 se muestra los resultados con y sin complementar los datos de entrenamiento con el corpus *Golem-Universum*. En la gráfica se pueden apreciar que las tres evaluaciones que se realizaron agregado el corpus *Golem-Universum* como dato de entrenamiento la mejora fue mínima, con respecto a las evaluaciones sin haberse agregado información adicional, lo que indica que los modelos acústicos son más precisos durante el proceso de reconocimiento.

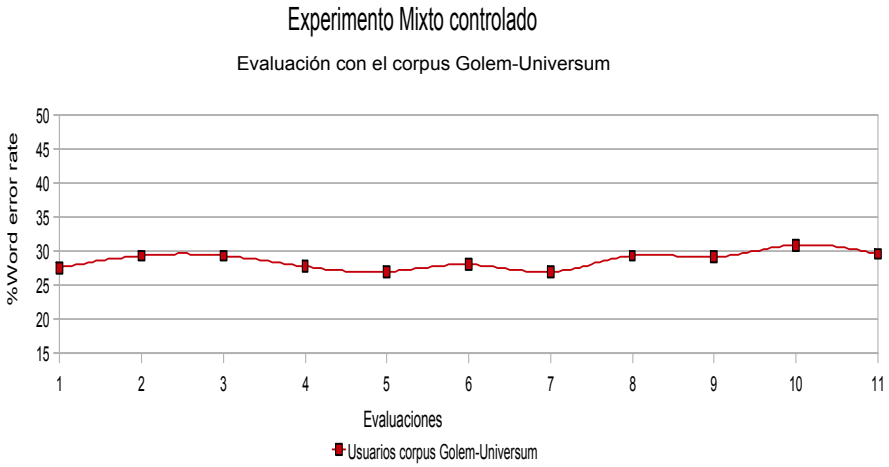


Figura 5.6: Gráfica del experimento Mixto controlado evaluado con el corpus *Golem-Universum*.

Una vez realizadas las evaluaciones anteriores ahora se procede con aquellas en donde se agregaron o se quitaron datos de uno u otro corpora. Cuando se entrenó con la siguiente configuración de datos 99 usuarios del corpus *DIMEx100 adultos* + 85 usuarios del corpus *DIMEx100 niños* + usuarios del corpus *Golem-Universum*, el valor de la evaluación del reconocedor de voz dio un porcentaje de *word error rate* a nivel de palabra de 21.9 %. Por el valor obtenido se puede decir que los modelos acústicos obtenidos al agregar información del corpus de habla espontánea son buenos. En la gráfica de la figura 5.8, se muestra una comparación entre la evaluación sin haber agregado el corpus *Golem-Universum* como dato de entrenamiento con la que se obtuvo en este experimento, con el objetivo de mostrar el comportamiento del reconocedor de voz cuando es evaluado con diferentes tipos de hablantes. En esta gráfica se puede observar que los modelos acústicos se vuelven más

| **Usuarios de entrenamiento** | **Evaluación con corpus *Golem-Universum*** |
|---|---|
| 99 niños + usuarios corpus *Golem-Universum* | 21.6 % |
| 99 adultos + usuarios corpus *Golem-Universum* | 23.7 % |
| 99 niños + 99 adultos + usuarios corpus *Golem-Universum* | 23.5 % |

Cuadro 5.1: Resultados de las evaluaciones de referencia con el corpus *Golem-Universum*

precisos cuando es agregada información del tipo de hablante con el que se está evaluando.

Si se desea conocer los resultados de las distintas evaluaciones que se realizaron al reconocedor de voz seleccionado para este experimento en el apéndice A se muestra un cuadro con todas las evaluaciones que se realizaron específicamente para esta configuración de datos.

El último de los experimentos que se realizó creando un reconocedor de voz con 59 usuarios del corpus *DIMEx100 adultos* + 40 usuarios del corpus *DIMEx100 niños* + usuarios del corpus *Golem-Universum* el porcentaje de *word error rate* que se obtuvo en esta evaluación fue de 22.9 %, lo que indica que los modelos acústico presentan mejoramiento con respecto a la evaluación anterior sin agregar datos del corpus *Golem-Universum* para entrenamiento donde el resultado fue de 26.9 %; en este experimento se puede decir que se tienen buenos modelos acústicos. En la figura 5.9 se muestra la gráfica donde se comparan el resultado de las dos evaluaciones que hemos mencionado

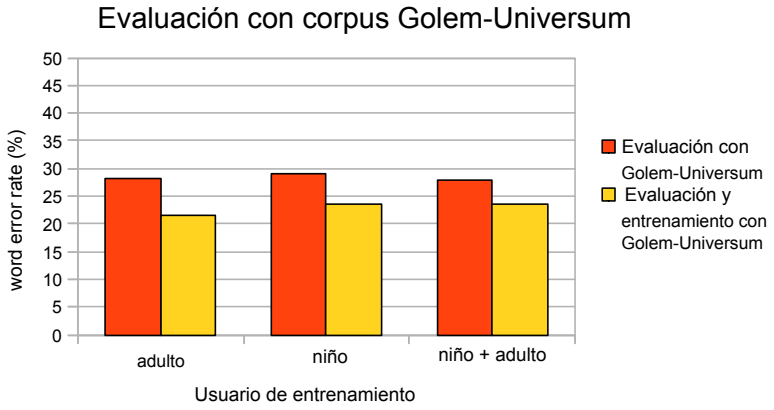


Figura 5.7: Resultados de las evaluaciones de referencia con el corpus *Golem-Universum*

donde se puede observar claramente que si se ha presentado mejoramiento en esta evaluación.

Todos los experimentos que se presentaron en esta sección fueron realizados con fines exploratorios, para poder obtener mejores conclusiones se recomienda realizar las evaluaciones para cada uno de los usuarios de los corpora.

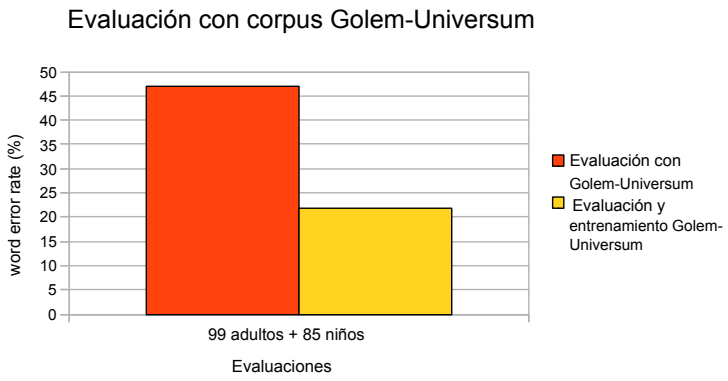


Figura 5.8: Gráfica comparativa de las evaluaciones realizadas con el corpus *Golem-Universum*, agregando 85 usuarios del corpus *DIMEx100 niños*.

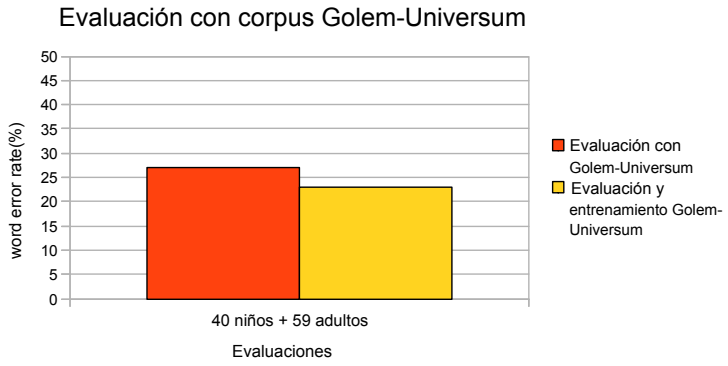


Figura 5.9: Gráfica comparativa de las evaluaciones realizadas con el corpus *Golem-Universum* complementando los datos de entrenamiento.

## 5.3. Ambiente de desarrollo

Los modelos acústicos que se utilizaron en las evaluaciones son creados durante el proceso de entrenamiento por lo tanto son únicos en cada reconocedor de voz, para el caso del modelo de lenguaje y el diccionario de pronunciación han sido tomados del reconocedor de voz que se encuentra implementado en el juego "Adivina la carta".

Las características del modelo acústico corresponden exactamente a los experimentos del capítulo 4, el orden del modelo de lenguaje del juego "Adivina la carta" corresponde a 3 (trigramas). El diccionario de pronunciación únicamente contiene palabras propias del contexto del juego.

De igual forma que los experimentos del capítulo 4, estos experimentos se realizaron bajo las mismas condiciones y utilizando el mismo equipo de computo, por tal motivo el tiempo de procesamiento para los experimentos de este capítulo consistió de aproximadamente 30 hrs. Las evaluaciones en este capítulo se realizaron nuevamente en modo *bach* y usando la herramienta *sclite*.

### 5.3.1. Discusión

En este capítulo se mostraron los experimentos donde se evaluó y se entreno con el corpus de voz espontánea *Golem-Universum* con la finalidad de realizar diversos entrenamientos dinámicos de los modelos acústicos para caracterizar los corpus que se tienen disponibles para esta tesis, además de conocer el desempeño de los modelos acústicos cuando son probados en condiciones reales ya que los experimentos fueron diseñados para que la mayor caga de trabajo durante el proceso de reconocimiento se llevara a cabo en los modelos acústicos.

El primer experimento que se presenta, los modelos acústicos que fueron evaluados han

sido los mismo que se evaluaron en los experimentos del capítulo 4, para estas evaluaciones se ha presentado una mejoramiento durante las pruebas, incrementando considerablemente el reconocimiento de las palabras, este resultado nos indica que los modelos acústicos creados con los corpora *DIMEx100 niños* y *DIMEx100 adultos*, son robustos, buenos y precisos durante el proceso de reconocimiento y lo mejor es utilizar proporciones de ambos corpora para crear los reconocedores de voz.

El experimento en donde es utilizando el corpus *Golem-Universum* como dato de entrenamiento muestra que agregar información del tipo de hablante con el que se esta evaluando como se pudo ver en el experimento Mixto no balanceado y Mixto Controlado descritos en el capítulo 4, para este caso la voz espontánea de niños, sí ayuda a mejorar los resultados del reconocimiento de las palabras, pero no es tan significativo e importante el mejoramiento que se obtiene de acuerdo a las pocas evaluaciones que se realizaron, pues el experimento únicamente es con fines exploratorios. Debe de tenerse en cuenta que las palabras que forman al corpus *Golem-Universum* son en su mayoría frases cortas todas pertenecientes al contexto del juego "Adivina la carta" pero la dificultad que presenta se tiene presente es que son frases espontáneas habladas por niños; por las características tan especificas que tiene este corpus es un recurso valioso para próximas investigaciones.

Por todo lo que hemos mencionado anteriormente y en base a los resultados obtenidos de las evaluaciones podemos decir que lo corpora *DIMEx100 niños* y *DIMEx100 adultos* son un excelente recurso lingüístico para la construcción de tecnologías del habla para el español de México, especialmente para dos tipos de hablantes: niños y adultos. En el siguiente capítulo se presentan las conclusiones generales de esta tesis así como el trabajo a futuro que es posible realizar.

# Capítulo 6

# Conclusiones

El desarrollo de tecnologías del habla particularmente para el español de México es muy importante para crear aplicaciones orientadas a reconocer la voz de distintos tipos de hablantes, además de impulsar la investigación para desarrollar este tipo de aplicaciones. Para la creación de estas tecnologías es primordial contar con buenos recursos lingüísticos, que proporcionen las características necesaria para que el reconocedor de voz presente un desempeño aceptable de acuerdo al tipo de hablante para el que fue diseñado. Por tal motivo este trabajo ha sido propuesto con la finalidad de caracterizar los corpora en español de México: *DIMEx100 niños* y *DIMEx100 adultos*, para conocer que tan buenos son estos recursos limitados por la cantidad de hablantes que tienen, pero ricos por las características fonéticas que poseen.

En este capítulo presentamos las conclusiones y contribuciones que se han logrado de el trabajo desarollado en esta tesis, así como el trabajo a largo plazo que es posible desarrollar partiendo de los resultados. En la sección 6.1 de este capítulo se presentan las conclusiones

y contribuciones de el trabajo; finalmente, en la sección 6.2 se presenta el trabajo a futuro que es posible llevar a cabo.

## 6.1. Conclusiones

En esta tesis hemos presentado una serie de experimentos que han sido diseñados con el objetivo de caracterizar corpora de habla leída en español de México: los corpora de habla leída *DIMEx100 niños* y *DIMEx100 adultos*. Para lograr el objetivo de esta tesis se plantearon una serie de preguntas que fueron contestadas diseñando distintos experimentos que nos ayudaron a observar el comportamiento de un reconocedor cuando es entrenado con diferentes proporciones y tipos de hablantes. A continuación se presentan las preguntas que fueron de gran ayuda para lograr el objetivo de caracterizar los corpora:

1. ¿Cúal es el comportamiento de un reconocedor de voz que ha sido creado para un tipo de hablante y que interactua con un tipo de hablante distinto?

2. ¿Cómo se comporta un reconocedor de voz cuando es entrenado para distintos tipos de hablantes?

3. Cuando es agregada información de entrenamiento poco a poco de un tipo de hablante distinto, ¿Cuál es el comportamiento de el reconocedor de voz?

4. De acuerdo a las características tan particulares que presentan ambos corpora de voz leída ¿pueden llegarse a complementar los datos de entrenamiento en un reconocedor de voz?

5. ¿Cuál es el comportamiento de un reconocedor de voz cuando la interacción ocurre en

un ambiente real?

Como se ha mencionado en capítulos anteriores, la inquietud de caracterizar estos corpora ha sido debido a las características tan peculiares que presentan, especialmente el hecho de que son paralelos pues los corpora contienen las mismas frases con la diferencia de que han sido grabadas por dos tipos de hablantes distintos. La caracterización de los corpora se ha realizado para saber si los recursos con que se cuentan son aptos para la creación de reconocedores de voz en español de México.

De acuerdo a los resultados obtenidos al evaluar los modelos acústicos creados con hablantes niños, hablantes adultos y con la unión de ambos tipos de hablantes, se ha observado que el mejor desempeño ocurre con los modelos creados con el corpus *DIMEx100 adultos*, este corpus presenta mayor estabilidad cuando es utilizado como dato de entrenamiento y de evaluación, esto se ha observado en la mayoría de las evaluaciones que se realizaron, por lo tanto, podemos decir que los modelos acústicos que son creados únicamente con este corpus son más robustos durante el proceso de reconocimiento a diferencia de los creados únicamente con datos del corpus *DIMEx100 niños* (ver figura 6.1). La explicación que se da a este comportamiento es por el tipo de hablantes y rango de edad que se encuentra constituido el corpus *DIMEx100 niños*, pues es el rango de edad donde se presentan mayores dificultades para obtener un buen desempeño en reconocimiento de voz [Russell and D' Arcy, 2007, Blomberg and Elenius, 2004].

Por otro lado se ha observado que crear modelos acústicos combinando dos tipos de hablantes, se obtienen mejores resultados que con un solo tipo de hablante (ver figura 6.1), esto ha sido observado en los experimentos Mixto balanceado y Mixto no balanceado, este último experimento nos ha ayudado a saber qué cantidades son las mejores para crear

modelos acústicos que ayuden a obtener un buen reconocimiento. De acuerdo a los resultados obtenidos en esta tesis, y de las investigaciones realizadas para otros idiomas mencionadas en el capítulo 3, se coincide en que los mejores resultados se obtienen al combinar los datos de los corpora para crear modelos acústicos, además de que el desempeño de un reconocedor de voz es mejor cuando el modelo acústico es creado con el mismo tipo de hablante con el que se evalúa, esto se ha mostrado en los resultados obtenidos de el experimento Simple y en las investigaciones realizadas que presentamos como en el capítulo 3 sección 3.2. La única diferencia que se presenta en las investigaciones y los resultados de esta tesis, es que las evaluaciones se realizaron para frases largas y palabras cortas, por el contrario las investigaciones trabajan únicamente con palabras cortas.

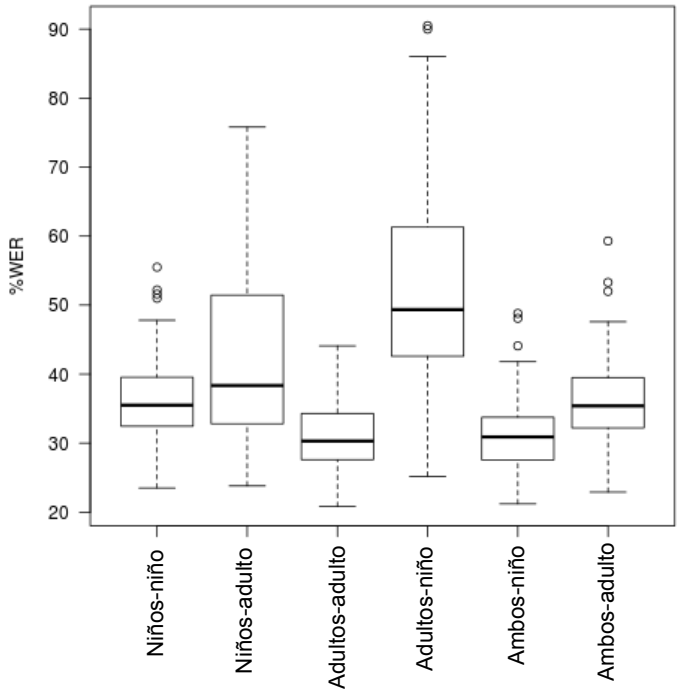


Figura 6.1: Diagrama de los resultados de las evaluaciones con los corpora de habla leída.

En general se puede decir que los modelos acústicos que se crearon han sido buenos, esto ha sido posible por las características de los corpora, del cuidado que se tuvo durante la creación de estos recursos; además de que se utilizaron transcripciones ortográficas creadas manualmente, las cuales ayudan pues se transcriben las palabras tal como son pronunciadas por el hablante.

Las evaluaciones realizadas con el corpus de habla espontánea que se diseñaron para evaluar los modelos acústicos bajo condiciones reales (habla del usuario en tiempo real y en condiciones normales), los modelos acústicos han dado muy buena respuesta presentando un reconocimiento aceptable, tal como se mostró en los resultados obtenidos de las evaluaciones en el capítulo 5, además de que una vez más se pudo observar que al agregar datos del corpus de habla espontánea *Golem-Universum*, ayuda a mejorar la calidad de los datos para la creación de los modelos acústicos disminuyendo el valor de *word error rate*; tal como se puede apreciar en el diagrama de la figura 6.2. Los porcentajes de *word error rate* que se obtuvieron de las evaluaciones no muestran una diferencia tan considerable pero si ayudan a mejorar el desempeño del reconocimiento. De igual manera que en las evaluaciones con los corpora de habla espontánea el mejor reconocimiento se obtiene cuando son ocupados como datos de entrenamiento los dos corpora de habla leída. Se debe de tener en cuenta que este corpus de habla espontánea *Golem-Universum* está formado en su mayoría por palabras cortas que pertenecen a un dominio especifico, en este caso del juego "Adivina la carta".

La construcción de un reconocedor de voz para el español de México no es una tarea fácil de llevar a cabo, principalmente por la falta de recursos lingüísticos; pero afortunadamente se cuenta con estos dos corpora de habla leída en español de México: el corpus *DIMEx100 niños* y *DIMEx100 adultos*, que han demostrado con los experimentos realizados en esta

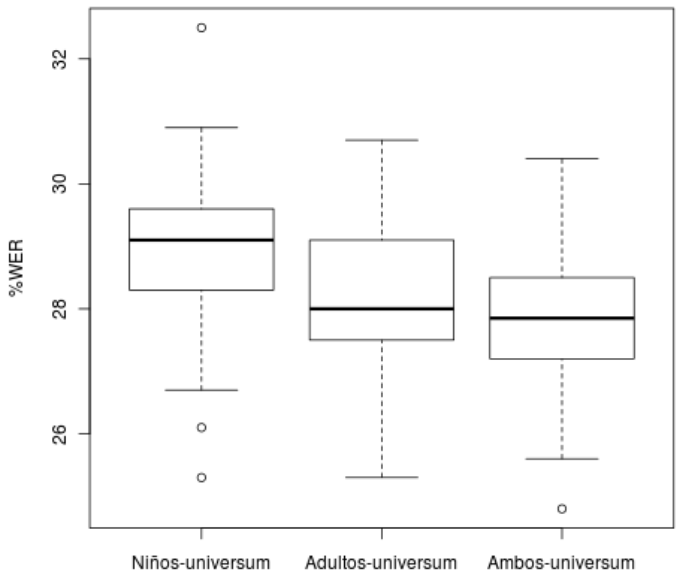


Figura 6.2: Diagrama de los resultados de las evaluaciones con el corpus de habla espontánea.

tesis, que son recursos que tienen datos con una buena calidad para crear modelos acústicos, los cuales dan una mayor certeza en el proceso de reconocimiento de voz, pues se han obtenido porcentajes aceptables del reconocimiento. Se debe de tener en cuenta que además de contar con buenos recursos lingüísticos, también una parte importante en la creación de un reconocedor de voz es el uso de la herramienta para crearlo, en este caso de *Sphinx3*, así como de el equipo utilizado para el procesamiento y manipulación de la información.

A continuación, enlistamos las contribuciones más sobresalientes que han surgido de este trabajo:

- Conocer más a detalle que tan buenos son los recursos fonéticos: corpora *DIMEx100 niños* y *DIMEx100 adultos*, de acuerdo a las características que presentan, para la creación de tecnologías del habla.

- Durante el proceso de caracterizar empíricamente los corpora de habla leída, se ha observado que lo mejor es combinar los dos tipos de hablantes para obtener un mejor desempeño del reconocedor de voz.

- Mejor desempeño del reconocedor de voz en español de México, cuando se interactúa simulando condiciones reales con hablantes niños.

- Los resultados de las evaluaciones reflejan que el corpus que presenta mejores características para crear reconocedores de voz es el corpus *DIMEx100 adultos*.

En http://code.google.com/p/acustics-models/ se encuentran los scripts que se implementaron para poder llevar a cabo los experimentos desarrollados en esta tesis.

## 6.2. Trabajo a futuro

De acuerdo a las evaluaciones que se realizaron en esta tesis, en un futuro es posible con los datos de los corpora que se tienen disponibles, realizar pruebas basadas en estudios de género de los hablantes siguiendo la misma metodología que se ha realizado en esta tesis. Este tipo de pruebas sería muy importante porque nos darían una idea más precisa de cuál es el la calidad de los modelos acústicos cuando son creados específicamente para un genero, además de poder realizar más combinaciones de los datos y conocer cuál es la calidad de los modelos acústicos cuando se crean bajo estas condiciones.

Además es posible aplicar técnicas de adaptación a los modelos acústicos tales como VTLN (por sus siglas en inglés, Vocal Tract Length Normalization), MAP (por sus siglas en inglés, Maximun a Posteriori) y MLLR (por sus siglas en inglés, Maximun Likelihood Linear Regresion), solo por mencionar algunas, y volver a realizar los experimentos propuestos en

esta tesis, para observar si presentan alguna mejora los modelos acústicos. Por el momento el aplicar estas técnicas de adaptación han quedado fuera del alcance de esta tesis.

Finalmente, es posible realizar todas las evaluaciones utilizando todos los usuarios de los corpora *DIMEx100 niños* y *DIMEx100 adultos* en los experimentos Mixto no balanceado y Mixto complementario a fin de observar el comportamiento en estos dos experimentos para todos los usuarios, ya que como mencionamos cuando se explicaron cada uno de estos experimentos en esta tesis, la evaluación que se realizo únicamente fue hecha con fines exploratorios.

# Apéndice A

# Resultados de evaluaciones con habla leída y espontánea

En este apéndice se presentan los resultados obtenidos de el conjunto de experimentos que se realizaron en esta tesis, los cuales han sido explicados en los capítulos 4 y 5. En cada una de las gráficas es posible observar la calidad de los modelos acústicos de acuerdo al tipo de hablante con el que se evaluó, además de poder realizar una comparación de acuerdo al corpora con el que se entreno y evaluó. Las tres evaluaciones que se presentan en cada una de las gráficas corresponden a: habla leída de niños, habla leída de adulto y habla espontánea de niños.

**Evaluaciones del experimento Simple**

La gráfica de la figura A.1 muestra los resultados de las tres evaluaciones que se realizaron

en el experimento Simple cuando el tipo de hablante con el que fue entrenado cada reconocedor de voz corresponde a niño. El caso contrario se presenta en la figura A.2. donde los reconocedores de voz están creados para reconocer voz de adultos.

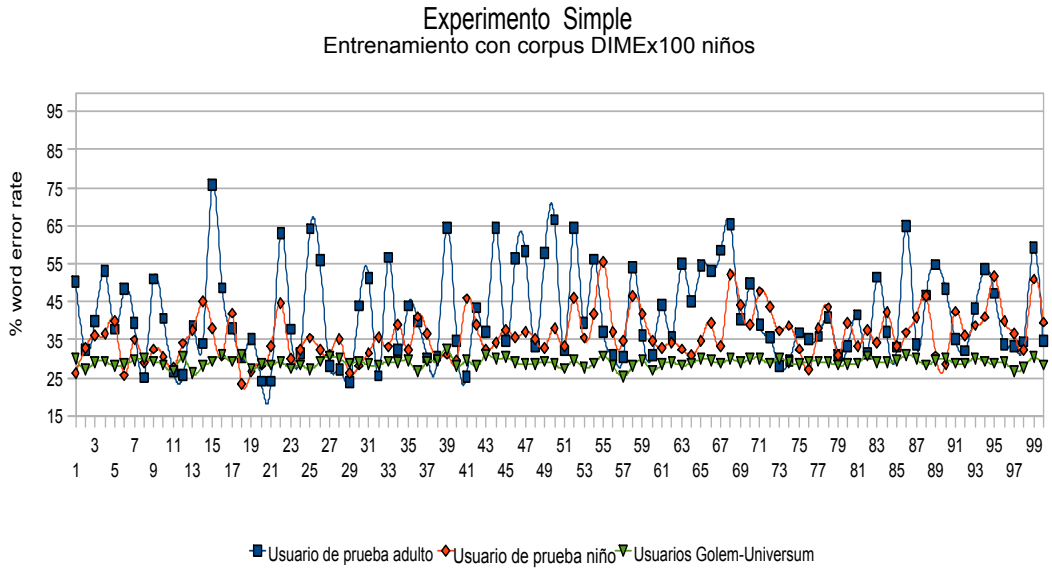


Figura A.1: Gráfica de evaluaciones del experimento Simple entrenado con corpus *DIMEx100 niños*.

**Evaluaciones del experimento Mixto balanceado**

La gráfica que se presenta a en la figura A.3, contiene los resultados de las tres evaluaciones que se realizaron en este experimento. En la gráfica se puede que los mejores resultados se obtienen cuando la evaluación es realizada con un usuario adulto, lo cual indica que la información de los ambos corpora ayudan a que los modelos acústicos sean más robustos para reconocer voz de adultos. Además con estos resultados se puede decir que el corpus *DIMEx100 adultos* es el que presenta mejores características fonéticas de los hablantes, lo

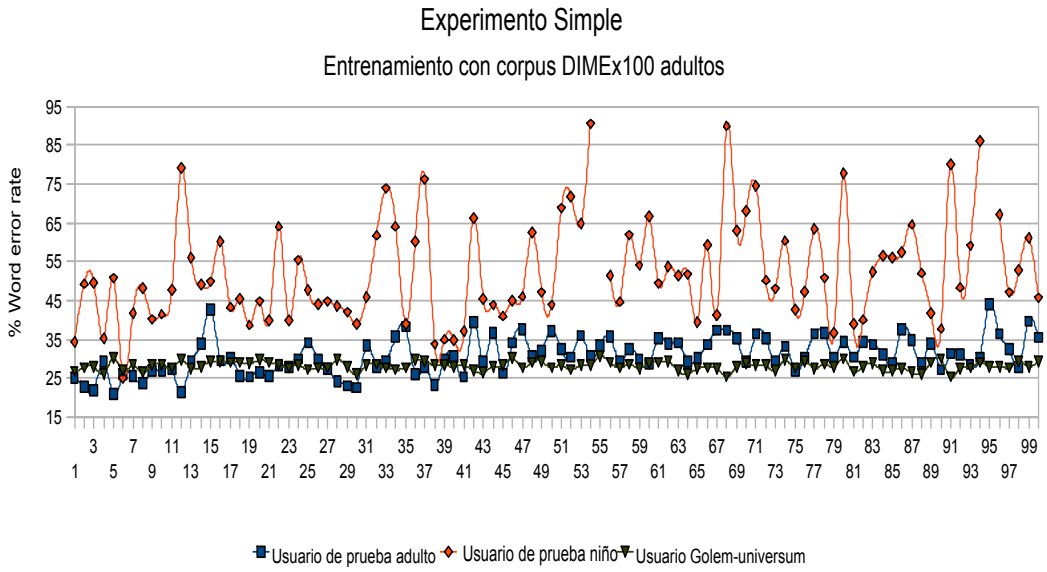


Figura A.2: Gráfica de evaluaciones del experimento Simple entrenado con corpus *DIMEx100 adultos*.

cual propicia un mejor reconocimiento.

**Evaluaciones del experimento Mixto no balanceado**

La figura A.4 muestra la curva de aprendizaje cuando se tiene como base usuarios del corpus *DIMEx100 adultos* para el entrenamiento, agregándose en cada iteración un usuario del corpus *DIMEx100 niño*. De igual manera en la figura A.5, se muestra el caso contrario de la gráfica A.4, donde ahora se tienen como base de datos de entrenamiento usuarios del corpus *DIMEx100 niño* y se van a agregando un usuario del corpus *DIMEx100 adultos* en cada iteración.

La curva de aprendizaje de la figura A.4 muestra claramente que a medida que se van agregando datos del tipo de usuario a reconocer se mejora el reconocimiento hasta llegar a

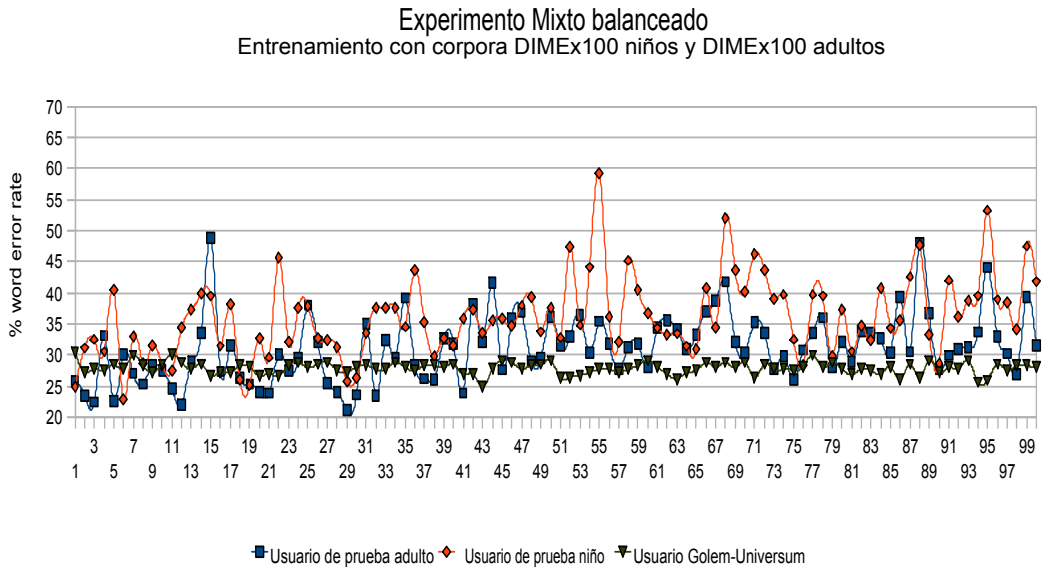


Figura A.3: Gráfica de evaluaciones del experimento Mixto balanceado.

comportarse estable; las otras dos evaluaciones (usuario adulto y usuarios *Golem-Universum*) el reconocimiento siempre se mantienen estable a lo largo de los ciclos de entrenamiento. De igual forma el resultado de las evaluaciones que se muestran en la figura A.5 siempre se mantienen constantes a medida que se agregan datos de entrenamiento.

Ambos experimento se han evaluado únicamente con los usuarios 15 de los corpora. Este experimento puede extenderse a ser evaluado con todos los usuarios de ambos corpora.

**Evaluaciones del experimento Mixto controlado**

Los resultados que se muestran en la gráfica A.6 complementa el resultado que se presento en el cap. 4, esta gráfica ahora presenta las tres evaluaciones que se realizaron en este experimento. Esta gráfica es muy interesante pues permite apreciar la diferencia de cada eva-

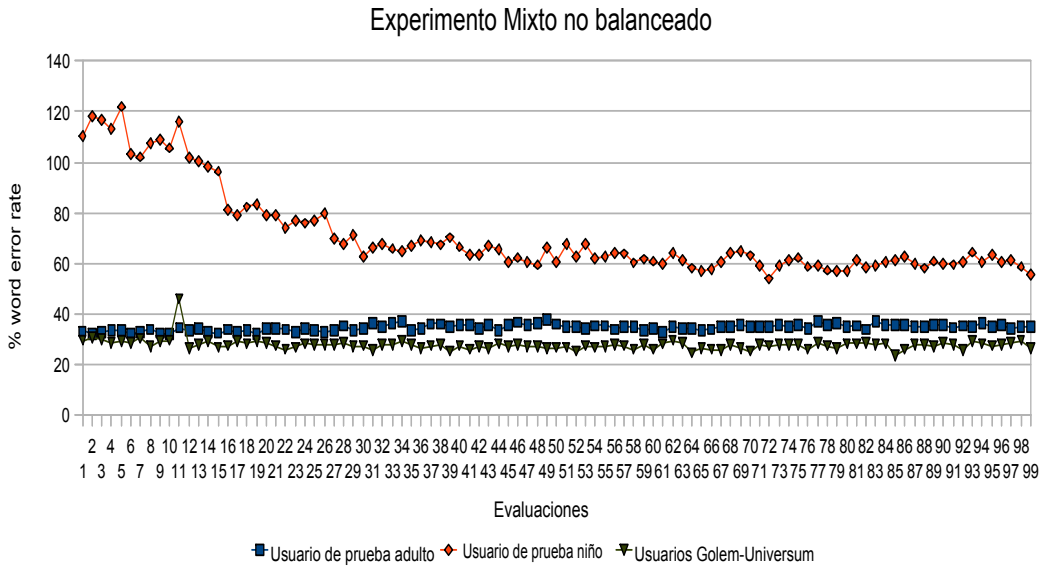


Figura A.4: Curvas de aprendizaje del experimento Mixto no balanceado (datos base de entrenamiento adultos).

luación, obteniendose peores resultados al ser evaluado con un usuario del corpus *DIMEx100 niños*, mientras que los mejores se obtuvieron con voz espontánea de niños; estos resultados pueden verse contradictorios pero debe de tenerse en cuenta que el corpus *Golem-Universum* está compuesta por frases en su mayoría corta y sin mucho grado de dificultad. La curva de aprendizaje muestra que a medida que los datos de entrenamiento se encuentran balanceados con los dos tipos de hablantes, el reconocimiento mejora. Este experimento como ya se ha mencionado únicamente es exploratorio, para sacar una mejor conclusión es necesario evaluar con todos los usuarios de los corpora *DIMEx100 niños* y *DIMEx100 adultos*.

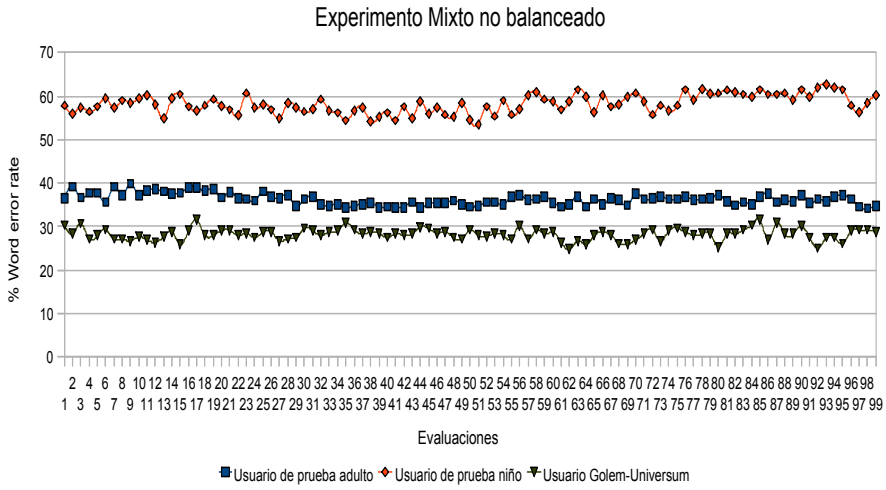


Figura A.5: Curvas de aprendizaje del experimento Mixto no balanceado (datos base de entrenamiento niños).

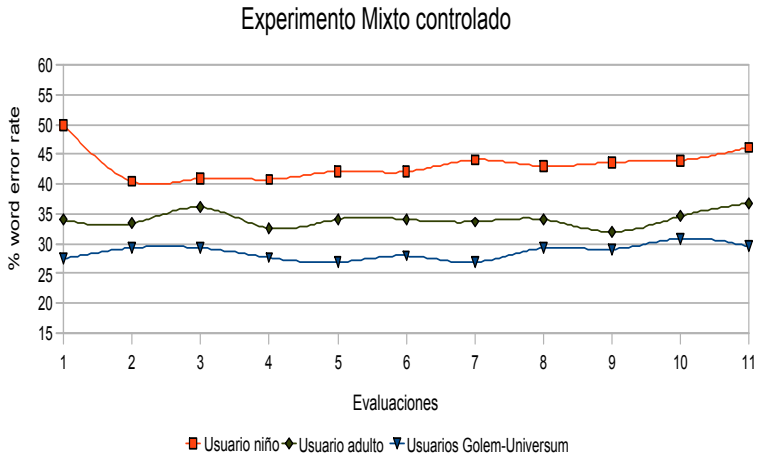


Figura A.6: Curvas de aprendizaje del experimento Mixto controlado.

# Bibliografía

[Allen et al., 2006] Allen, J. F., Ferguson, G., Blaylock, N., Byron, D. K., Chambers, N., Dzikovska, M., Galescu, L., and Swift, M. D. (2006). Chester: Towards a personal medication advisor. pages 500–513.

[Allen et al., 1994] Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., Martin, N. G., Miller, B. W., Poesio, M., and Traum, D. R. (Semptiembre 1994). The trains project: A case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI*, Vol. 7:7–48.

[Batliner et al., 2005] Batliner, A., Blomberg, M., DÁrcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., and Wong, M. (2005). The PF star children's speech corpus. In *Proc Interspeech 2005*.

[Blomberg and Elenius, 2004] Blomberg, M. and Elenius, D. (May 2004). Comparing speech recognition for adults and children. *Proc. of The XVIIth Swedish Phonetics Conference, Fonetik 2004*, pages 56–159.

[Elenius and Blomberg, 2005] Elenius, D. and Blomberg, M. (2005). Adaptation and normalization experiments in speech recognition for 4 to 8 year old children. In *Proc Interspeech*

*2005*.

[Giuliani and Gerosa, 2003] Giuliani, D. and Gerosa, M. (2003). Investigating recognition of children's speech. *Proc. ICASSP'03, Hong Kong, China*, Vol. 2:137–140.

[Heeman and Allen, 1995] Heeman, P. and Allen, J. (1995). Trains 93 dialogues. Technical report, Universidad de Rochester, Departamento de ciencias de la computación.

[Jelinek, 1999] Jelinek, F. (1999). *Statistical Methods for Speech Recognition*. Prentice Hall.

[Jurafsky and James H., 2008] Jurafsky, D. and James H., M. (2008). *Speech and Language Processing*. Prentice Hall.

[Juárez Vázquez, 2009] Juárez Vázquez, J. A. (2009). Construcción de un reconocedor de voz para el español de México con variación alofónica media. Tesis de Licenciatura, Universidad Nacional Autónoma de México.

[Llisterri and Machuca, 2006] Llisterri, J. and Machuca, M. J. (2006). *Los sistemas de diálogo*. Bellatera-Soria Universidad Autónoma de Barcelona.

[Meza et al., 2010a] Meza, I., Pérez, E., Salinas, L., Aviles, H., and Pineda, L. A. (2010a). A multimodal dialogue system for playing the game "Guess the card". IIMAS,UNAM.

[Meza et al., 2010b] Meza, I. V., Salinas, L., Venegas, E., Castellanos, H., Chavarría, A., and Pineda, L. A. (2010b). Specification and evaluation of a spanish conversational system using dialogue models. In Proceedings of Iberamia-2010, Kuri-Morales, A., editor, *In this volumen*.

[Miranda-Palma et al., 2007] Miranda-Palma, C., Camal-Uc, R., Cen-Magaña, J., Gonzalez-Segura, C., Gonzalez-Segura, S., García-García, M., and Lizzie, N.-D. (2007). Usabilidad

en un juego de memorama con reconocimiento de voz para niños. *conferencia IADIS Ibero-Americana*, pages 129–136.

[Pineda, 2008] Pineda, L. A. (2008). El proyecto DIME y el robot conversacional Golem: Una experiencia multidiciplinaria entre la computación y la lingüística. UNAM.

[Pineda et al., 2009] Pineda, L. A., Castellanos, H., Cuétara, J., Galescu, L., Juárez, J., Llisterri, J., Pérez, P., and Villaseñor, L. (2009). The corpus Dimex100: Transcription and evaluation. *Language Resources and Evaluation.*

[Pineda et al., 2004] Pineda, L. A., Villaseñor, L., Cuétara, J., Castellanos, H., and López, I. (2004). Dimex100: A new phonetic and speech corpus for Mexican spanish. In Advances in Artificial Intelligence, Iberamia-2004, C. L., Reyes, C. A., and Gonzalez, J. A., editors, *Lecture Notes in Artificial Intelligence*, volume 3315, pages 974–983. Springer-Verlag.

[Pérez Pavón, 2006] Pérez Pavón, E. P. (2006). Construcción de un reconocedor de voz utilizando Sphinx y el corpus Dimex-100. Tesis de Licenciarura, Universidad Nacional Autónoma de México.

[Russell and D' Arcy, 2007] Russell, M. and D' Arcy, S. (2007). Challenge for computer recognition of children's speech.

[Tapia et al., 2010] Tapia, G., Meza, I. V., and Pineda, L. A. (2010). Language models for name recognition in spanish sapoken dialogue system. In Proceedings of MICAI-2010, M.-.

[Wilpon and Jacobsen, 1996] Wilpon, J. G. and Jacobsen, C. (1996). A study of speech recognition for children and the elderly. *Proc. ICASSP'96*, Vol. 1:349–352.