4. Reconocimiento de Voz de Niños para el Español Hablado en México Usando SONIC

Los objetivos principales de esta tesis son construir un reconocedor robusto para niños y minimizar los errores que afectan el desempeño del mismo, utilizando métodos de normalización de la longitud del tracto vocal, y adaptación de los parámetros del modelo (como las medias y las varianzas) o utilizando la probabilidad de un evento antes de considerar nuestro conocimiento adicional sobre los datos de entrenamiento. Cada uno de los métodos utilizados ataca problemas en específico que pueden ocurrir en nuestros datos de entrenamiento o evaluación provocando una degradación en los porcentajes de reconocimiento.

En los capítulos anteriores se han dado las bases para poder entender el proceso en general de cada una de las tareas de un reconocedor de voz. Hemos utilizado este conjunto de herramientas llamado SONIC porque nos permite la investigación y el desarrollo de nuevos algoritmos para el reconocimiento de voz continua, y la creación de aplicaciones para su uso en tiempo real con rapidez y eficiencia. SONIC está basado en tecnología sobre modelos ocultos de Markov de densidad continua, además incorpora dentro de sus herramientas métodos de normalización y adaptación para un mejor desempeño del reconocedor. En el presente capítulo se explicarán cada uno de los pasos y la configuración necesaria para el entrenamiento de nuestro reconocedor de voz de niños.

Es necesario puntualizar que el reconocimiento automático de voz de niños no es nuevo, tampoco lo son las técnicas empleadas en esta tesis. Lo importante de este trabajo es que no se han realizado estudios de este tipo teniendo como fuentes corpus nativos de México y en

especial de niños, utilizando sistemas de reconocimiento de voz de última generación y aplicando algunos métodos probados para mejorar el desempeño los modelos obtenidos en dicho reconocimiento.

El proceso de entrenamiento y evaluación así como el preproceso de los datos se llevaron a cabo en computadoras con procesador Pentium IV, 512 MB en RAM y con suficiente capacidad de almacenamiento. Además se contó con micrófonos y tarjetas telefónicas que permiten el desarrollo y evaluación de sistemas de lenguaje hablado. Además se utilizaron como herramientas de apoyo Emacs, scripts en Tcl y el CSLU Toolkit principalmente. La mayoría de los scripts están hechos en C-shell el cual permite un procesamiento por lotes para cada una de las rutinas que necesitamos [7].

4.1 Modelado Acústico

SONIC utiliza modelos ocultos de Markov de densidad de continua que son estimados basándose en un árbol de decisión de agrupamiento de estados [45]. Esto significa que cada fonema puede ser una secuencia de estados de un modelo oculto de Markov. Se utilizan 3 estados para caracterizar las partes de inicio, mitad y fin de un fonema. En la figura 4.1 podemos ver un ejemplo de un HMM de 3 estados para la palabra *uno*. Cada fonema consiste en 3 estados de modelos ocultos de Markov. Se ve la secuencia de observaciones $O=\{O_1,...,O_{18}\}$ que consiste en 18 vectores de características. En el HMM se muestra la transición entre estados, emitiendo 18 vectores de observación S = {S1,S1,S2,S2,S2,S3,S3,S4,S5,S5,S6,S6,S6,S7,S8,S8,S8,S9} .

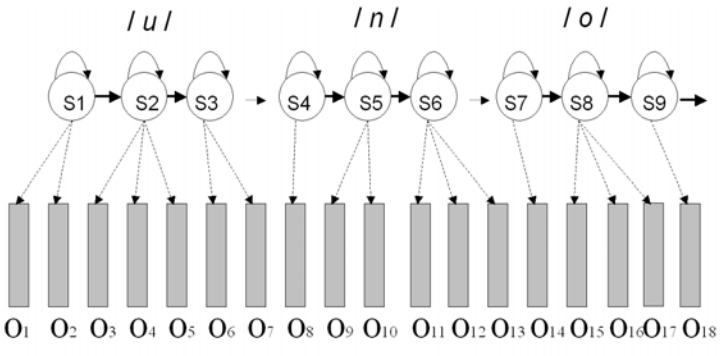

Figura 4.1 Secuencia de un modelo oculto de Markov para la palabra "uno"

El sistema maneja entre 6 y 32 densidades Gausianas por estado. Cada Gausiano está representado por un peso (w_m), un vector de medias (μ_m), y una matriz de covarianzas ($\sum_m$). Datos más específicos sobre el modelado se pueden encontrar en [6].

El objetivo principal del proceso de entrenamiento del modelo acústico es estimar los parámetros de cada estado del modelo oculto de Markov (medias, varianzas, pesos compuestos) mientras se condiciona las distribuciones en el contexto fonético que los rodea (modelado del contexto del trifonema). Muchos de estos trifonemas no se visualizan en los datos de entrenamiento.

4.2 Panorama del Proceso de Entrenamiento del Modelo

El proceso de entrenamiento del modelo consiste en tres pasos: extracción de características, alineamiento de Viterbi basado en estados, y estimación del modelo. El resultado del proceso de entrenamiento es un conjunto de HMM de árboles de decisión y agrupados por estado. La alineación y estimación del modelo se repiten hasta que los errores del conjunto de evaluación sean minimizados. El corpus que utilizamos necesita

tener un conjunto de archivos de audio (con un muestreo de 16 bits lineales PCM sin encabezado) y transcripciones de texto junto con un diccionario de pronunciación en español. Los archivos de voz de niños difieren del formato especificado por el sistema y fue necesario escribir un script en C-shell que utiliza la herramienta "sox" para convertir cada uno de los archivos de audio a dicho formato.

En resumen, para esta tesis, una vez mapeados cada uno de los fonemas del inglés al español, se utilizaron el CSLU Toolkit junto con un script en Tcl para obtener el léxico [53]. Además se eliminaron algunos fonemas extras de este léxico y se cambió el formato al requerido por SONIC. La siguiente etapa consistió en hacer los archivos de configuración de los fonemas en español. Se modificaron los scripts de entrenamiento y evaluación usados en [6] cambiando el rango de muestreo, así como los archivos de configuración de los fonemas.

En la figura 4.2 se puede ver el diagrama del proceso de entrenamiento del modelo usado por SONIC. Los pasos incluyen la extracción de características y alineación fonética de los archivos de voz con su transcripción correspondiente. Durante esta fase se utilizan modelos acústicos iniciales para hacer una alineación de Viterbi a nivel de estados del modelo oculto de Markov. Los vectores de características y sus alineaciones de estado son agrupados en un archivo MLF (Master Label File). Este archivo se procesa para generar conjunto de datos binarios antes de entrenar el modelo acústico. Los nuevos modelos son usados para realinear los datos y se repite el proceso (obteniendo cada vez mejores modelos).

4.3 Preparación de los Datos

La frecuencia de muestreo del corpus de niños es de 16 kHz. Cada directorio está etiquetado con un número para identificar el locutor. Dentro de cada directorio, tenemos los archivos de audio (.raw) con un archivo correspondiente que contiene la transcripción de lo que ha dicho el niño (.txt). En caso de querer identificar el género del locutor es necesario un archivo adicional. Estos serán nuestros datos de entrenamiento para el sistema.

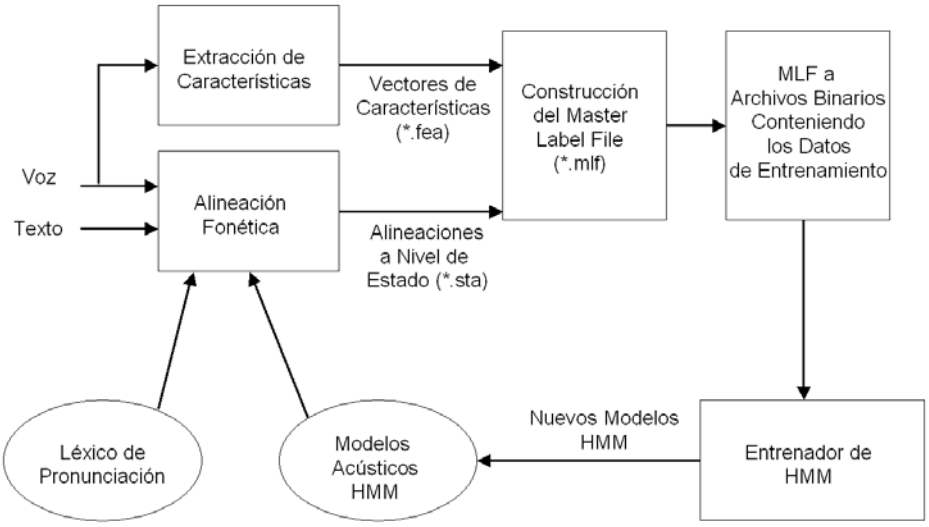

Figura 4.2 Diagrama del proceso de entrenamiento del modelo acústico

4.4 Alineación y Extracción de Características

El objetivo de este paso es asociar los vectores de características a cada uno de los tres estados de cada fonema. SONIC originalmente trabaja con fonemas en Inglés y se tuvo que hacer un mapeo de dichos fonemas a fonemas en Español. Las guías que se utilizaron para hacerlo fueron tomadas de [52,53]. En la tabla 4.1 podemos ver la lista de los símbolos

fonéticos del Español y su equivalente en Inglés, así como un ejemplo para cada uno. Con estos fonemas también se genera el archivo de configuración utilizado por el sistema. Los archivos de configuración se pueden ver en el Apéndice A.

La alineación inicial se puede hacer de dos maneras, teniendo modelos en español o tomar los modelos en Inglés que el sistema provee. Para esta tesis la primera alineación de Viterbi se hizo con modelos en Inglés.

El proceso de alineación en el tiempo genera un archivo para cada pronunciación de entrenamiento que contiene el marco (*frame*) inicial y final para cada estado del HMM, los fonemas asignados a los marcos y la posición de los fonemas dentro de una palabra.

En esta etapa se usa el algoritmo de Viterbi para determinar la segmentación más probable del archivo de audio dada la secuencia de las características extraídas y la transcripción conocida a nivel palabra. Cabe destacar que la precisión que tenga la alineación impactará directamente en el desempeño de los modelos entrenados. Muchas veces se puede iterar realineando hasta que haya convergencia o se alcance el desempeño esperado en el reconocedor. Detalles sobre este punto se darán en el siguiente capítulo.

El siguiente paso antes de entrenar, es construir un archivo llamado MLF (Master Label File). El contenido de este archivo de texto es el nombre del archivo de características (contiene vectores de características extraídos de los archivos de audio) así como información de la alineación de cada estado. Se puede generar un archivo mlf para todos los datos de entrenamiento (mlf.si), o sólo para los datos donde los locutores son hombres (mlf.gm) o cuando son mujeres (mlf.gf).

Como nuestros experimentos están enfocados a reconocimiento de voz de niños y dado que las características de la voz en esas edades (8 a 11 años) son similares, hemos decidido emplear solamente el archivo que reúne las características de todo nuestro conjunto de entrenamiento, es decir, usar modelos independientes del locutor.

MEX	EU	Ejemplo
a	AA	p**a**pá
b	B	**b**oca
tS	CH	**ch**ica
d	D	**d**entro
D	D	ca**d**a
dZ	DH	**ll**ama
e	EY	b**e**bé
f	F	**f**avor
g	G	**g**oma
G	G	a**g**uda
j	G	**h**ierba
x	HH	**j**ota
I	IY	ni**ñ**o
k	K	**c**oma
l	L	**l**ana
m	M	**m**atar

MEX	EU	Ejemplo
n	N	**n**adar
N	N	ci**n**co
nj	NG	ni**ñ**o
o	OW	b**o**da
p	P	**p**oco
r	R	pa**r**a
rr	R	**r**osa
s	S	**s**ol
t	T	ces**t**a
u	UW	d**u**da
V	V	sa**b**io
w	W	h**u**eso
z	Z	de**s**de
SIL	SIL	silencio

Tabla 4.1 Conjunto de fonemas del Español usado en SONIC y sus equivalentes en Inglés

4.5 Entrenamiento

Una vez completada la alineación para cada estado de los datos de entrenamiento, extraídos los vectores de características para los archivos de audio, y generado el MLF para describir el conjunto de datos de entrenamiento, se puede empezar con el entrenamiento. La idea principal es extraer las características y la información contextual del archivo MLF reescribiendo la información en un directorio temporal. Los elementos que sirven de entrada para la rutina de entrenamiento son: un directorio de salida de datos, el archivo de

configuración del conjunto de fonemas mapeados en una de las secciones anteriores, el archivo MLF, el número de estados en el modelo HMM final para cada fonema base, y la longitud del vector de características.

4.6 Realineación y Reentrenamiento

Ya que se tienen los modelos acústicos entrenados para niños, se pueden refinar las alineaciones usando estos nuevos modelos. Como se ha dicho, la realineación y reentrenamiento se repiten muchas veces. La teoría dice que para corpus limpios se puede repetir este proceso de 3 a 4 veces, para nuestros corpus fue necesario repetirlo hasta 7 veces para alcanzar cierta convergencia.

4.7 Entrenamiento del Modelo del Lenguaje

Este paso es implementado directamente con la ayuda del conjunto de herramientas CMU/Cambridge[1]. Primero, se copian las transcripciones de referencia en un solo archivo que contiene las oraciones entre dos marcas <s>,</s>, que se usan para denotar el inicio y fin de una oración. Se llaman a las funciones del conjunto de herramientas para calcular el vocabulario de la tarea a realizar con los datos proporcionados en nuestro conjunto de entrenamiento. Después colectamos estadísticas sobre las secuencias de palabras y estimamos el trigrama del modelo del lenguaje. Y por último se comprime el modelo del lenguaje en un formato binario, el cual será utilizado en la evaluación.

[1] www.eng.cam.ac.uk/~prc14/toolkit.html

En el siguiente capítulo se detallará la metodología utilizada para llevar a cabo cada uno de los experimentos y analizar los resultados de cada uno.

Esta tesis no trata de proponer un nuevo sistema de reconocimiento de voz, sino que toma tecnología moderna en esta área y desarrolla un nuevo reconocedor de voz a partir del corpus de niños recolectado por el laboratorio de Tecnologías de Voz TLATOA para su utilización en los tutores de lecto-escritura de Colorado [14]. Por otra parte permite una mayor comprensión de la tarea de reconocimiento de voz de niños para el Español y el desarrollo de nuevos métodos que ayuden a mejorar el desempeño de nuevos reconocedores.