En este capítulo se presenta el desarrollo y resultados de seis experimentos realizados; tres usando la metodología de redes neuronales y tres con modelos ocultos de Markov. De cada metodología tenemos un experimento independiente del contexto y dos dependientes del contexto. Para realizar una comparación de los diferentes reconocedores desarrollados se tomaron características como tiempo de entrenamiento, desarrollo y pruebas, cantidad de memoria y velocidad de proceso par cada red y HMM, así como errores de inserción, omisión y sustitución. Así mismo se presenta la descripción de los corpus utilizados, la metodología de evaluación y finalmente las tablas de desempeño para cada experimento.

4.1 El Corpus

Para construir los reconocedores se utilizaron dos corpus, uno para desarrollo (*teléfono)* y otro corpus de propósito específico (*dígitos)* para realizar pruebas a nivel de palabra.

El corpus *teléfono* consiste de voz grabada por teléfono por más de 500 locutores. Se diseñó para cubrir vocabularios comunes como dígitos, números naturales, si/no, horas, días, fechas, etc... El corpus se formó de la unión de dos subcorpus: el *Tlatoa Common Questions* y *norte*. Una vez unidos se sacaron estadísticas del corpus sobre el número de ejemplos que se tienen para cada fonema así como duraciones promedio, duraciones mínimas y máximas. Véase la *tabla 4.1*, los datos completos se muestran en el apéndice D. Las condiciones de grabación incluyen formato RIFF y una frecuencia de muestreo de 8000 Hz. La base de datos de voz consta aproximadamente de 11.69 hrs de grabación continua y ocupan un total de 763 MB en disco.

El corpus *dígitos* fue grabado por 50 locutores y con un micrófono a 8000Hz. Las series contienen los números del 0-9 más el número 10.

Fonema	Milisegundos	Porcentaje en el corpus	Duración promedio	Duración mínima-máxima
a	25174	8.17%	88.79	10-907
b	2495	0.81%	17.06	6-70
c	3354	1.09%	17.57	5-99
e	36843	11.95%	80.99	8-1370

Tabla 4.1 Estadísticas del corpus

4.2 Diseño de Experimentos

Para desarrollar los reconocedores se utilizaron dos enfoques. El primer enfoque en redes neuronales y el segundo en modelos ocultos de Markov. Los datos en la fase de entrenamiento, desarrollo y prueba fueron obtenidos del corpus de *teléfono.* El tipo de dominio fue general para el desarrollo de este reconocedor (*teléfono*).

En el proceso de producción del habla cada fonema se ve afectado por su contexto. Para modelarlo es necesario dividir los fonemas en partes, y así modelar el dinamismo de los fonemas. El Toolkit permite dividirlo de la siguiente manera:

- Una parte, el fonema es independiente del contexto,
- Dos partes, la primera mitad depende del contexto izquierdo y la segunda del derecho.
- Tres partes, el primer tercio del fonema es dependiente del contexto izq., la parte central es independiente del contexto y la última depende del contexto derecho.
- También se provee la funcionalidad de modelar a un fonema como tres partes y una

Siguiendo la metodología estándar, el proceso de reconocimiento consta de tres fases: entrenamiento, desarrollo y evaluación. Por eso se necesitan datos para cada una de estas fases. Para esto se dividió el corpus para entrenamiento, desarrollo y prueba. Para

entrenamiento se asignaron 300 locutores, para la partición de desarrollo se usaron 100 locutores y para prueba se usaron todos los locutores del corpus *dígitos*. De los datos de entrenamiento se toman las muestras para que la red aprenda y es importante que sean suficientes, para asegurar de este modo un mejor modelado y reconocimiento. Los datos de desarrollo se utilizan para escoger la mejor iteración, es importante señalar que éstos datos sean diferentes a los de entrenamiento.

En los experimentos con redes neuronales se realizan un total de 30 iteraciones y se evalúan las últimas 15 redes escogiéndose solamente la que obtiene el mejor desempeño en el reconocimiento sobre los datos de la fase de desarrollo. En los experimentos de Modelos Ocultos de Markov la fase de entrenamiento consta de 10 iteraciones y se ocupa *Vector Quantization* y la realineación de *Viterbi*. De los 10 modelos se escoge el que obtiene el mejor reconocimiento.

Finalmente, para la etapa de evaluación final, se toma la red o el modelo con mayor desempeño resultante de la etapa de desarrollo, también son datos no usados en ninguna de las dos etapas anteriores.

4.2.1 Evaluación NIST

Los reconocedores fueron evaluados con el software NIST (*National Institute of Standards and Technology*) el cual aplicamos de forma manual para la evaluación del reconocedor basado en redes neuronales e implícito en el de modelos ocultos de Markov. Este software tiene dos propósitos: Primero, alentar a los investigadores a usar medidas estadísticas para resumir sus mejoras y conclusiones; y segundo, proveer una manera estándar para medir el porcentaje de reconocimiento, asegurando de esta forma que las diferencias en resultados publicados por diferentes grupos de investigación son debido al desempeño de sus reconocedores y no a sus algoritmos para calcular los resultados [CASTILLO99].

El software de NIST para evaluar un reconocedor, primero necesita ejecutar un algoritmo de alineación de cadenas para minimizar el número de diferencias (sustituciones, inserciones y eliminaciones) entre los resultados del reconocedor y lo correcto.

La forma de evaluación del CSLU Toolkit, genera unos archivos que muestran el resultado de la red, comparado con el correcto. El formato de estos archivos que se utilizan para evaluar el reconocedor es diferente del usado por NIST, por lo que fue necesario usar la documentación y los scripts desarrollados en el trabajo de tesis de Omar Castillo [CASTILLO99].

A continuación se dará la descripción y características de cada uno de los reconocedores y los resultados obtenidos en los diferentes experimentos que se diseñaron.

4.3 Resultados de los Experimentos

El primer experimento está basado en modelos acústicos independientes del contexto. Los experimentos independientes del contexto proveen un experimento base estándar sobre el cual se pueden realizar mejoras. Después se muestran los experimentos dependientes del contexto. En ellos, las clases dependientes del contexto se modelan de la siguiente manera:

Tabla de clases generales

$back	=	o o_h o_x u u_h u_fp u_x w ;
$dip	=	ae ai ei ia ie io oi wa we ;
$flap	=	r rZ rr ;
$fric	=	s s_v x x_fp f ;
$front	=	e e_h e_fp e_x i i_h i_fp i_x ;
$lat	=	l ;
$mid	=	a a_h a_fp a_x ;
$nasa	=	m m_fp n N nj ;
$noise	=	\.unk \.bn \.ln ;
$pau	=	\.pau \.br ;
$semi	=	D G V ;
$uclosu	=	tc tSc pc kc ;
$ustop	=	k p t tS ;
$vclosu	=	dc dZc bc gc ;
$vstop	=	b d dZ g ;

Nivel	Frame
Texto	dos
más independiente del contexto	dc d o s
más dependientes del contexto con clases	<dc>o>$fric

Niveles de representación fonética

4.3.1 Experimento I

Todas las unidades fonéticas se modelaron independientes del contexto (una parte). Los resultados se ven en la *Tabla 4.2*. Los resultados que se muestran se obtuvieron al evaluar los reconocedores sobre 1883 frases de 49 locutores.

	Exactitud	Palabras correctas	Error inserción	Error supresión	Error sustitución
NN2	98.2%	98.2%	5.5%	0%	1.4%
HMM2	71.19%	75.5%	4.3%	3.72%	20.78%

Tabla 4.2 Resultados Experimento I

4.3.2 Experimento II

Cabe mencionar que este segundo experimento se realizó dividiendo los fonemas en 3, 1 y r partes.

	Exactitud	Palabras correctas	Error inserción	Error supresión	Error sustitución
NN1	98%	98%	0	0	1.8
HMM	92.14%	93.29%	1.10	2.10	4.59

Tabla 4.3 Resultados Experimento II

4.3.3 Experimento III

Este experimento se realizó tomando sólo la mitad de los datos que se ocuparon en el primer experimento. Los resultados del tercer experimento se observan en la *Tabla 4.4.*

	Exactitud	Palabras correctas	Error inserción	Error supresión	Error sustitución
NN2	99.1%	99.1%	2.6	0	.9
HMM2	93.0%	93.97%	.9	1.9	4.06

Tabla 4.4 Resultados Experimento III

4.4 Uso de Recursos

A continuación se muestran los requerimientos en memoria y tiempo para cada experimento. En general redes neuronales utilizó más recursos de memoria y tiempo de entrenamiento.

4.4.1 Requerimientos de Memoria

En la *Tabla 4.5* se muestra los recursos que se utilizaron respecto a la memoria de cada uno de los experimentos. En general los reconocedores desarrollados utilizando redes neuronales ocuparon el doble de memoria.

Experimento	Memoria (MB)
NN I	700
NN II	650
NN III	600
HMM I	300
HMM II	224
HMM III	200

Tabla 4.5 Memoria

4.4.2 Requerimientos de Tiempo

El tiempo que se muestra en la *Tabla 4.6* es aproximado, ya que algunos procesos varían de acuerdo al número de datos que se estén manejando y a la complejidad computacional en cada uno de los pasos de desarrollo de los reconocedores. Todos los experimentos se realizaron en una máquina con 128 MB en ram. Al realizar una prueba en una máquina con 64 MB en ram, el tiempo de entrenamiento de los modelos de redes neuronales sobrepasaba las 36 hrs.

Reconocedor	Entrenamiento	Desarrollo	Prueba
NN I	4 hrs.	30 min.	1 hr.
NN II	3.5 hrs.	30 min.	1 hr.
NN III	2 hrs.	20 min.	45 min.
HMM I	2 hrs.	50 min.	40 min.
HMM II	1.5 hrs.	40 min.	30 min.
HMM III	1.25 hrs.	35 min.	25 min.

Tabla 4.6 Tiempo

Una vez dadas a conocer las características de cada uno de los experimentos, como el tiempo, cantidad de memoria para cada una de las etapas como son desarrollo, entrenamiento y prueba, se darán las conclusiones de lo que fue dicha tesis relacionando toda la teoría estudiada de cada uno de los enfoques con los resultados reales obtenidos en los experimentos.

4.5 Interpretación de Resultados

Los resultados de los experimentos muestran un mayor desempeño de reconocimiento al utilizar redes neuronales.

La contribución más significativa es la construcción de un primer reconocedor de voz multilocutor de propósito general usando modelos ocultos de Markov. Aunque éste mostró ligeramente un desempeño menor a redes neuronales, se observaron ventajas. Se notó además que los errores de sustitución en modelos de Markov son muy elevados en comparación a redes neuronales y repercutieron significativamente en los resultados finales. La dificultad de reconocimiento tiene varias causas entre las que están los efectos coarticulatorios, fronteras entre palabras así como ruido ambiental y ruido producido por el hablante (tos, risa, respiración). Sin embargo es necesario hacer más pruebas de evaluación con corpus grandes y diferentes al utilizado para el desarrollo, esto con el efecto de tener un mejor conjunto de evaluación..

En este capítulo se mostraron los resultados obtenidos en los diferentes experimentos utilizando los enfoques basados en redes neuronales y modelos ocultos de Markov. En el siguiente capítulo se presentarán las conclusiones, perspectivas y trabajo a futuro.