Capítulo V

Mejoras y prueba final

Los experimentos que se exponen aquí son mejoras realizadas al experimento base presentado en la sección anterior. Este tipo de mejoras radica en la optimización de los límites de duraciones, y la incorporación de modelado dependiente del contexto. Como ya se mencionó en el capítulo uno, este tipo de modelado requiere de la creación de subconjuntos de fonemas que compartan características lingüísticas. Es importante aclarar que en el experimento base del capítulo anterior no existe una clasificación de los fonemas en estas categorías, debido a que esa prueba se llevó a cabo independiente del contexto.

Mas adelante se muestra a detalle todos los pasos que se siguieron para llevar a cabo la prueba final, desde la preparación del *corpus* hasta el reporte de resultados. También se presenta un experimento que muestra la manera en que las mejoras hechas en un reconocedor fonético influye sobre el reconocimiento a nivel palabras.

5.1 Experimentos independientes del contexto

Como se vio en la sección anterior, los resultados obtenidos con el experimento base dejaban mucho que desear. Una de las cosas que estaba afectando de manera directa el desempeño del reconocedor era el gran número de inserciones que se estaban generando en la salida. Las inserciones por lo regular ocurren cuando hay fonemas de duraciones muy cortas, de tal manera que el reconocedor intenta meterlos dentro de una frase cuyo reconocimiento no fue muy bueno.

5.1.1 Modelos de duraciones

En la búsqueda Viterbi, la longitud de un fonema en particular determina que tan importante es éste para el *score* total. Debido a que estos son solo un componente de todo el *path*, los fonemas muy cortos influyen menos en el *score* que los fonemas más largos, lo cual puede generar errores en el reconocimiento. Para reducir esta fuente de errores, se agregan duraciones mínimas y máximas para cada fonema.

Además de aplicar duraciones absolutas mínimas y máximas, los reconocedores bajo la filosofía del CSLU *toolkit* agregan por cada *frame* un *penalty* en el *score* para segmentos que son muy largos o muy cortos. Esto le da al reconocedor flexibilidad en casos, por ejemplo, cuando hay una mala articulación donde se puede perder todo un segmento.

Para minimizar el problema de las inserciones se aumentó las duraciones mínimas de los fonemas a reconocer. De esta manera, se restringía la posibilidad de que el reconocedor metiera fonemas pequeños que no correspondían con la frase real.

Durante la etapa de entrenamiento se genera un archivo .olddesc (ver apéndice A) que contiene el vocabulario a reconocer y las duraciones mínimas y máximas de cada fonema en el vocabulario. Es en este archivo donde se aumentaron las duraciones mínimas usando un script de tcl llamado update_descdur.tcl. Los cambios más significativos se dieron al aumentar las duraciones mínimas de los fonemas en 4%, 6% y 8%.

5.1.2 Duraciones 4 % en experimento base

El criterio que seguía el *script* para aumentar el 4% a las duraciones mínimas era el siguiente: Por cada fonema, primero calculaba la diferencia de la duración máxima y mínima, a este resultado le calculaba el 4% y esto a su vez se lo sumaba a la duración mínima del fonema. Una vez modificado el archivo .olddesc se evaluaba de nuevo sobre el mismo subconjunto de evaluación. Como resultado, el porcentaje de inserciones bajo de un 57.3% a un 15.17%, y la precisión global subió de un 22.9% a un 47.23%.

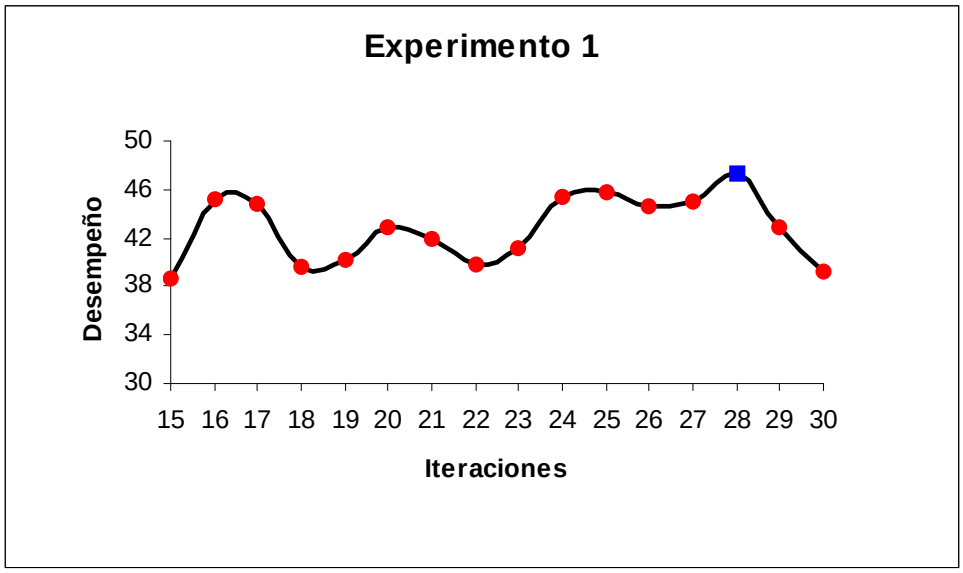

Figura 5.1 Iteraciones del aumento de duraciones del 4%, mejor red 28 con desempeño de 47.23%, inserciones 15.17%, substituciones 22.2% y eliminaciones 15.4%.

5.1.3 Duraciones 6% en experimento base

Como se vio que el desempeño del reconocedor mejoraba experimentando con los límites mínimos de duración, se aumentó otro 2% más; es decir, esta vez se evaluó incrementando en 6% los límites mínimos de duración de cada fonema. Aquí se nota que el porcentaje de inserciones siguió bajando, ahora de un 15.17% hasta 13.2%, con una precisión total de 53.6%.

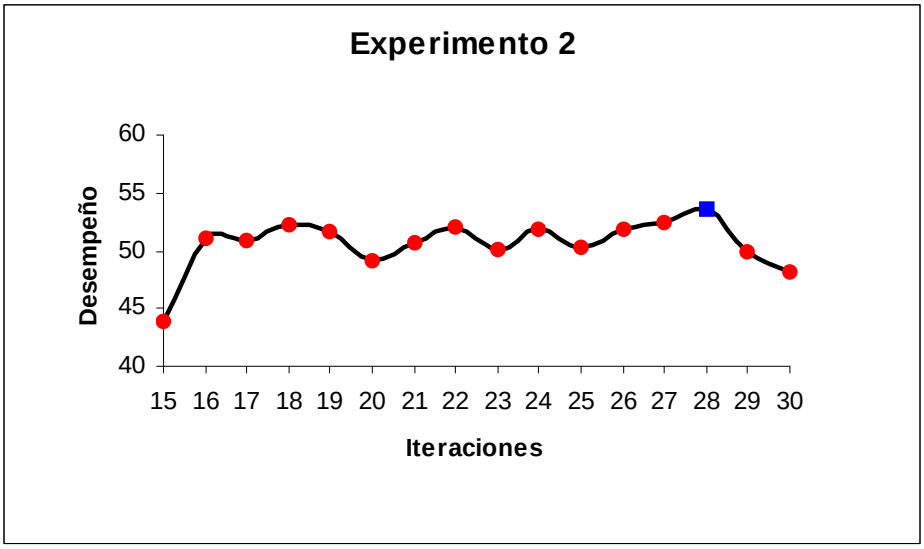

Figura 5.2 Iteraciones del aumento de duraciones del 6%, mejor red 28 con desempeño de 53.6%, inserciones 13.2%, substituciones 18.4% y eliminaciones 14.9%.

5.1.4 Duraciones 8% en experimento base

Se hizo una tercera prueba aumentando aún más las duraciones mínimas en el archivo .olddesc; esta vez se aumentaron un 8%. Cabe señalar que en este experimento el rendimiento fue más bajo, debido a que al aumentar las duraciones otro poco se penaliza al reconocedor al meter fonemas cortos pero legítimos. Esto ocasiona que el número de errores por eliminación aumente, resultando contraproducente para el desempeño del reconocedor.

Este experimento alcanzó los siguientes resultados:

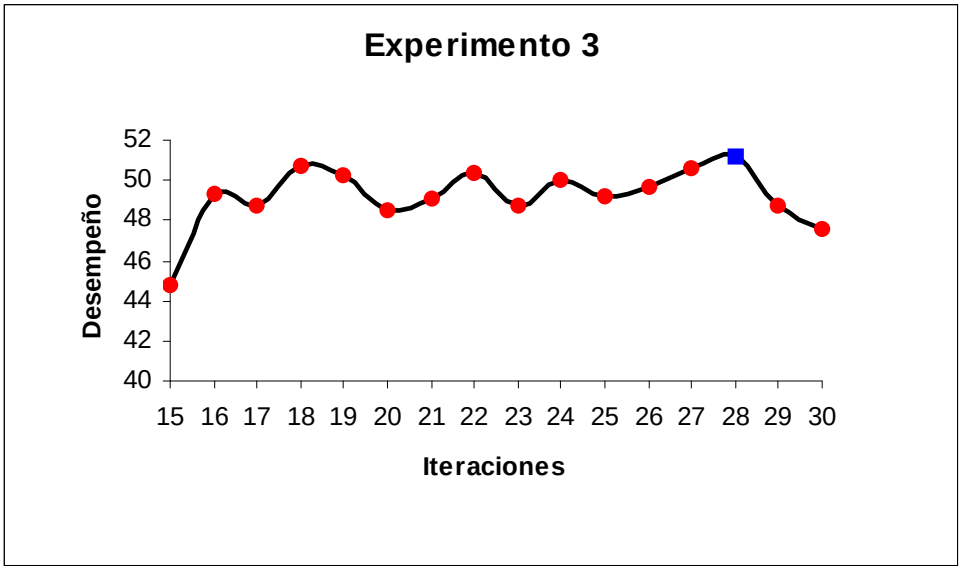

Figura 5.3 Iteraciones del aumento de duraciones del 8%, mejor red 28 con desempeño de 51.2%, inserciones 9.3%, substituciones 19% y eliminaciones 20.5%.

Notar el incremento en eliminaciones que se explicaba con anterioridad. Hasta aquí se ha mejorado el experimento base solamente en el aspecto de las duraciones; en la siguiente sección se incorporará además modelos dependientes del contexto. A continuación se muestra una tabla comparativa de los experimentos hasta ahora.

Red	Exp	Duraciones	Desempeño	S	E	I
28	base	-	**22.9%**	15.7%	4.1%	57.3%
28	1	4%	**47.2%**	22.2%	15.4%	15.17%
28	2	6%	**53.6%**	18.4%	14.9%	13.2%
28	3	8%	**51.2%**	19.0%	20.5%	09.3%

Tabla 5.1 Tabla comparativa de los resultados obtenidos en esta sección.

5.2 Categorías dependientes del contexto

Con base al conjunto de fonemas, las clases generales, las partes y el vocabulario definido, se genera el conjunto de categorías. Las categorías corresponden a los fonemas y sus variaciones en cada contexto del vocabulario. Si no fuera por las clases generales, se tendría un número muy grande de categorías, en el capítulo 2 se explica este tipo de modelado.

5.3 Número de partes

Como ya se vio en el capítulo dos, para efectos de modelado los fonemas se pueden dividir en partes.

- Las vocales y semivocales se dividen en tres partes porque son los fonemas más afectados por el contexto y porque son fonemas largos.
- Las consonantes oclusivas se dividen en una parte, contexto derecho (r), porque son demasiados cortos para dividirlos y también porque a la izquierda tienen siempre el mismo contexto (el closure).
- Los fonemas restantes se les dejo con una o dos partes.

Los subconjuntos de fonemas y sus partes obtenidas se observan en la siguiente tabla.

No. Partes	Fonemas
1	Silencios, *closures,* fricativas, africativas
2	Consonantes sonoras
3	Vocales, semivocales, diptongos
r	Oclusivas

Tabla 5.2 Partes de un fonema según su manera de articulación.

De acuerdo al modelo que se utilize para entrenar un reconocedor será la representación interna que maneje el *toolkit*, la siguiente tabla muestra los diferentes niveles en los que se puede representar una frase.

Nivel	Frase
Texto	no gracias
Fonemas independientes del contexto	.pau n o gc g r a s i a s .pau
Fonemas dependientes del contexto sin clases	<.pau> .pau>n <n> n>o <o> gc g>r <r> r>a <a> a>s <s> s>i <i> i<a <a> a>s <.pau>
Fonemas dependientes del contexto con clases	<.pau> $pau>n <n> n>$bck <o> <gc> g>$vib <r> r>$mid <a> a>$fri <s> s>$frnt <i> i<$mid <a> a>$fri <.pau>

Tabla 5.3 Niveles de representación fonética

5.4 Experimentos dependientes del contexto

Esta sección presenta la manera en que se mejoró el reconocedor incorporando modelos dependientes del contexto. En la sección 5.2 se explicó la manera en que se procede a crear las categorías. Estas categorías se definen en el archivo de partes que se usó para entrenar de manera dependiente del contexto (ver apéndice A, archivo telecd.parts)

A continuación se presenta un ejemplo que muestra las categorías en las cuales se agruparon los fonemas:

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

Tabla 5.4 Categorías generales

Los experimentos dependientes del contexto se dividieron en dos grupos, uno que se entrenó con un máximo de 10,000 muestras (frames) por cada categoría, y otro con un máximo de 50,000 muestras por categoría. Y a la vez se hicieron mejoras con duraciones para cada uno de ellos.

5.4.1 Dependiente del contexto 10,000 muestras

Se decide tomar el número máximo de categorías para realizar el entrenamiento dependiendo del corpus que se ocupe para entrenar. Normalmente, entrenar con 10,000 categorías para un corpus de las características de *Tlatoa Common Questions Corpus* esta bien; por ello se decidió hacer esto como primer experimento. Cabe señalar que los únicos fonemas que lograron sobrepasar las 10,000 muestras son las vocales y los silencios. Esta prueba inicial con 10,000 muestras como máximo por categoría generó los siguientes resultados.

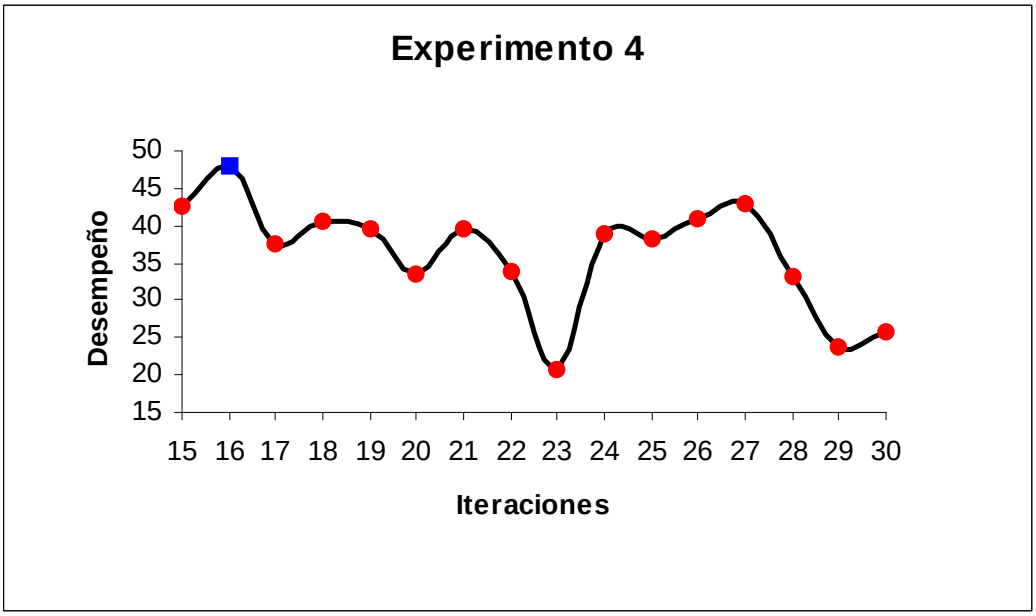

Figura 5.4 Iteraciones del experimento dependiente del contexto con 10,000 muestras, mejor red 16 con desempeño de 47.9%, inserciones 27.3%, substituciones 18.3% y eliminaciones 6.5%.

Si comparamos el desempeño de este sistema contra el obtenido en el experimento base (22.9%), donde no existen modelos dependientes del contexto, se nota una diferencia de más del doble. Esto muestra la ventaja de entrenar de manera dependiente del contexto.

5.4.2 Dependiente del contexto 10,000 muestras con duraciones 6%

Una vez que ya se tenía un sistema dependiente del contexto entrenado con 10,000 muestras máximas por cada categoría y los resultados de evaluación habían reportado un desempeño de 47.9%, se volvió a evaluar nuevamente, pero esta vez aumentando las duraciones mínimas de los fonemas. Se aplico nuevamente el 6% de aumento, ya que se notó que esta cantidad resultó mejor en experimentos anteriores. De esta manera se obtuvieron los siguientes resultados para esta etapa:

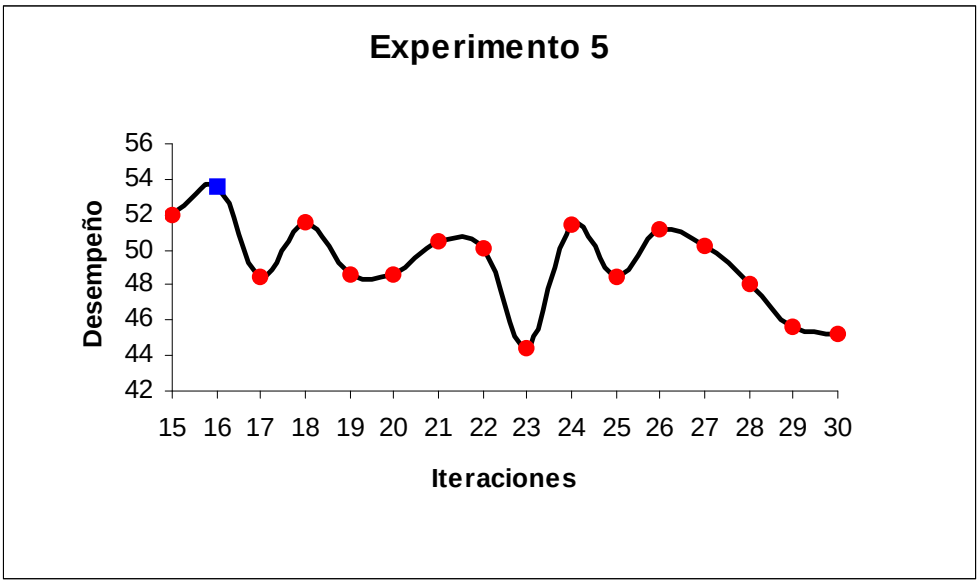

Figura 5.5 Iteraciones del aumento de duraciones del 6% al experimento 4, mejor red 16 con desempeño de 53.6%, inserciones 8.4%, substituciones 22.5% y eliminaciones 15.4%.

Casualmente este resultado concuerda con el obtenido en el experimento 2, el cual era independiente del contexto y mejorado con modelo de duraciones en 6%. Se puede observar que mientras las eliminaciones permanecieron similares en ambos experimentos, en este experimento el porcentaje de substituciones aumentó y el de inserciones bajó con respecto al experimento 2.

5.4.3 Dependiente del contexto 50,000 muestras

Aparte del experimento con 10,000 muestras se hizo otro con 50,000 muestras como máximo para cada categoría. Hay que destacar que, en general, si se aumenta el número de muestras para categoría, el desempeño de cualquier reconocedor se debe incrementar. La razón por la cual se limita el número de muestras proviene del deseo de reducir el tiempo de entrenamiento y el uso de espacio en disco duro. Así que los resultados que se consiguieron en esta sección se muestran a continuación.

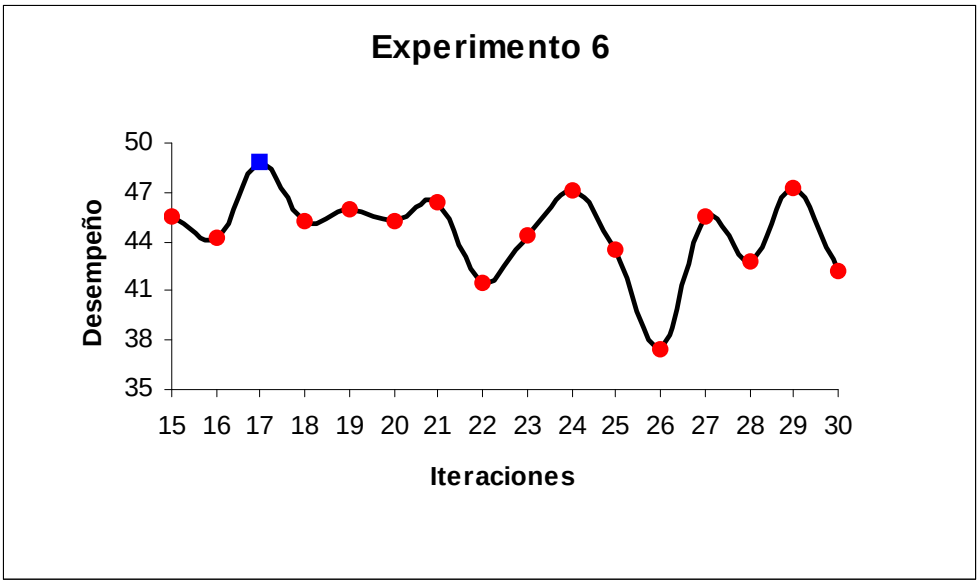

Figura 5.6 Iteraciones del experimento dependiente del contexto con 50,000 muestras, mejor red 17 con desempeño de 48.9%, inserciones 25.4%, substituciones 17.7% y eliminaciones 8%.

Cabe señalar que esta red es 1% mejor que la red del experimento 4. Esto se debe a que con este experimento el reconocedor mejoró, “aprendiendo” de las vocales y los silencios que sobrepasaban las 10,000 muestras.

5.4.4. Dependiente del contexto 50,000 muestras con duraciones 6%

Ya que se tenía el sistema dependiente del contexto entrenado con 50,000 muestras máximas por cada categoría y los resultados de evaluación habían reportado un desempeño de 48.9%, se volvió a evaluar nuevamente, pero una vez más aumentando las duraciones mínimas de los fonemas. Los resultados quedaron de la siguiente manera:

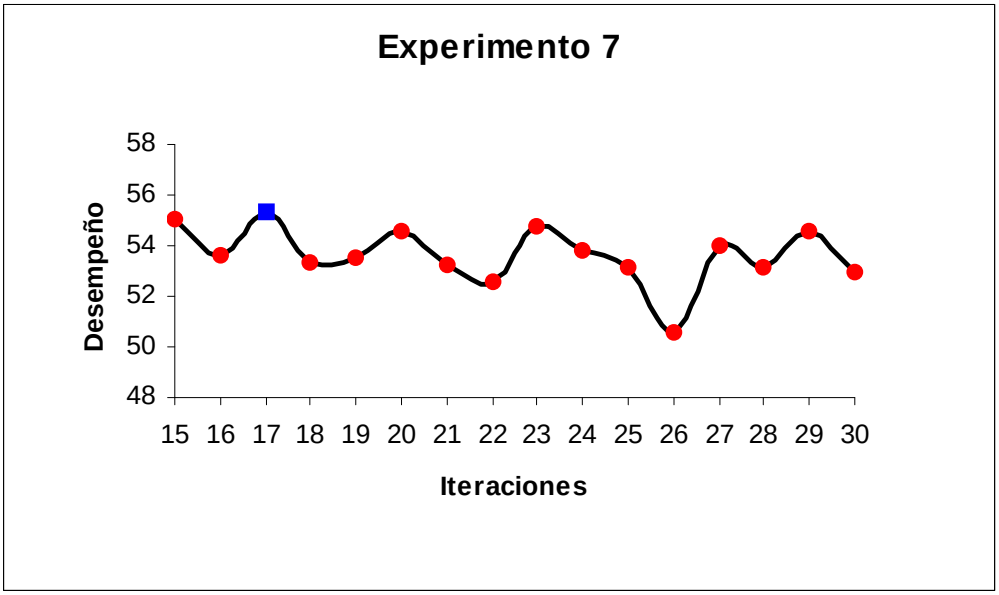

Figura 5.7 Iteraciones del aumento de duraciones del 6% al experimento 6, mejor red 17 con desempeño de 55.4%, inserciones 8.3%, substituciones 20.1% y eliminaciones 16.2%.

Este resultado fue el mejor que se consiguió dentro de la categoría de modelos dependientes del contexto, esta red es casi 2% mejor que la red del experimento 5.

A continuación se muestra una gráfica del desempeño alcanzado por los diferentes experimentos comentados en los capítulos cuatro y cinco.

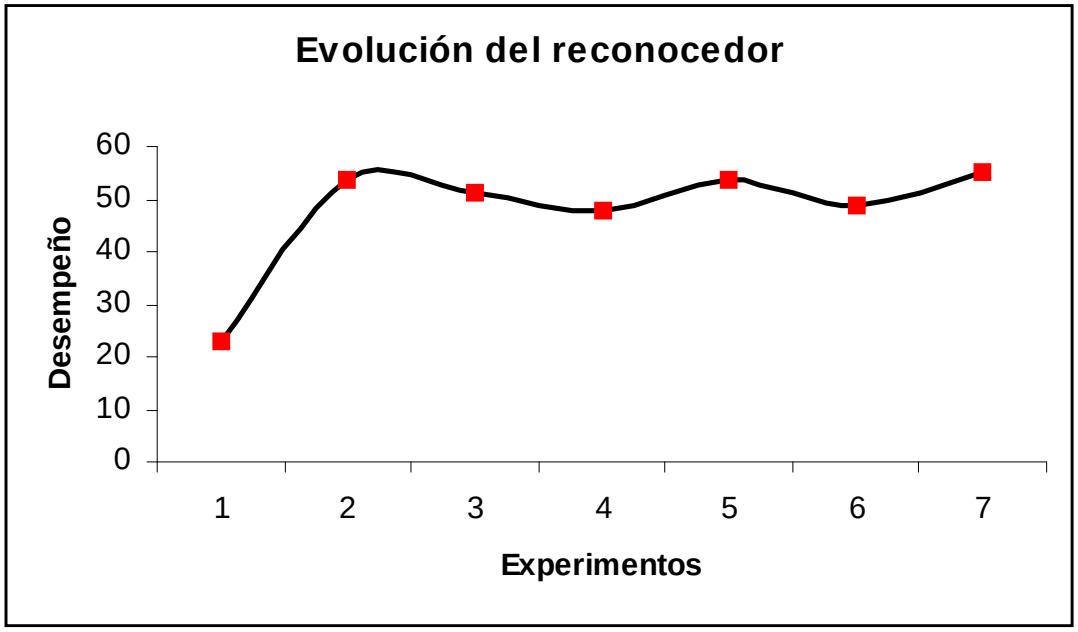

Figura 5.8 Evolución del desempeño del reconocedor

Otra manera de mostrar el avance sobre cada experimento es en términos de la reducción de error de cada experimento sobre el experimento base. Esta información se resume en la siguiente tabla:

Experimento	Reducción del error
1	31.55%
2	39.81%
3	36.70%
4	32.42%
5	39.81%
6	33.72%
7	42.02%

Tabla 5.5 Reducción del error en cada experimento

5.5 Fase de prueba final

Una vez que se han evaluado y mejorado los sistemas, viene la prueba final, en la cual se tomó la mejor red de todos los experimentos dependientes del contexto y su similar para el caso de los independientes del contexto. Para la siguiente fase es necesario probar con datos totalmente desconocidos para las redes, las cuales se han optimizado sobre los datos de evaluación. Cabe mencionar que el desempeño sobre datos de evaluación no refleja la realidad de lo que sucederá en el mundo real.

Finalmente es necesario efectuar una evaluación final de la redes neuronales para medir su desempeño. Para realizar este proceso se prueban sobre el último conjunto de datos las redes que obtuvieron los mejores resultados en la fase de desarrollo. Este grupo de datos tiene la transcripción .fone de cada archivo, hecha por el script, por lo que se sabe con anterioridad la salida esperada del reconocedor.

Como ya se dijo anteriormente el conjunto de prueba es habla espontánea o *stories*, que son frases donde el locutor puede decir cualquier cosa que se le venga a la mente en ese momento. Generalmente estas frases son muy largas, y contienen muchas partes dentro de la grabación que no es habla, las cuales no nos interesa reconocer. Por ejemplo, un *story* típico podría ser el siguiente.

EH YO AHORITA ESTOY ESTUDIANDO EN LA UNIVERSIDAD DE LAS AMERICAS ESTUDIO **(silencio largo**) **EH** ADMINISTRACION DE HOTELES Y RESTAURANTES ME GUSTA MUCHO MI CARRERA **(aclarando voz)** SOY DE CHIAPAS PERO ESTOY ESTUDIANDO AQUI AHORITA **EM** ESTUVE UN AN~O DESPUES DE LA PREPARATORIA ESTUDIANDO INGLES EN ESTADOS UNIDOS EN UTAH **EM** ESTO ES COMO…

Ejemplo 5.2 Un *story* típico

Como se puede ver, hay varias cosas que no nos interesan reconocer y que minimizan el desempeño de un reconocedor, como por ejemplo los eh, em, esteee, am, um, o ruidos como estornudos, garraspeo, tos, risas, ruido de fondo, etc..

Debido a lo anterior y a que los sistemas fueron entrenados con frases no tan largas, se hicieron dos scripts para depurar el subconjunto de prueba. Uno de ellos, validawrd.tcl (ver apéndice B) leía las etiquetas de los archivos .wrd e iba clasificando las palabras en habla, silencio o basura.

El otro *script* wavecut.tcl (ver Apéndice B), tomaba las etiquetas generadas por validawrd.tcl e iba cortando la señal de voz, dejando solamente aquellos segmentos de voz que correspondían a habla. Posteriormente los cortes fueron revisados manualmente.

Cuando ya se tuvo listo el subconjunto de prueba, se probó con la mejor red independiente del contexto y con la mejor dependiente del contexto, que correspondían a las redes 18 del experimento 2 y la 17 del experimento 7, respectivamente. Para este conjunto de 65 locutores y 495 frases, cuyo número promedio de frases por locutor fue 7, se obtuvieron los siguientes resultados:

Red	Experimento	Desempeño	S	E	I
28	2	**43.2%**	23.9%	26.1%	6.8%
17	7	**47.7%**	22.6%	26%	3.7%

Tabla 5.6 Resultados finales

Notar que el reconocimiento es más bajo de lo que se logró en el conjunto de evaluación, debido a que son datos no conocidos anteriormente, y debido a las diferencias en el estilo de habla de los *stories*. En particular, notar el alto número de eliminaciones, quizás debido a los aumentos de duración mínima que perjudican al reconocimiento de habla espontáneo, la cual es más rápida en general. No se puede optimizar sobre datos de prueba final, pero se puede agregar los demás *stories* a los datos de entrenamiento y evaluación para tener datos más similares en los tres conjuntos. Esto quedaría como un trabajo a futuro.

5.6 Prueba a nivel palabras

Regresando a la motivación original de esta tesis, una de las metas del reconocimiento de voz basado en fonemas es que este tipo de reconocedores sean independientes del vocabulario y puedan ser usados para diferentes tareas. Las demandas de independencia de vocabulario y del hablante requieren de un examen basado en fonemas como unidades. Si los modelos fonéticos son mejores, el sistema entero tendrá un mejor desempeño en cualquier aplicación y con cualquier vocabulario. Para comprobar esta hipótesis, se probaron estas dos redes sobre un corpus de dígitos usando el siguiente vocabulario:

PALABRA	PRONUNCIACION
cero	{s e r o}
uno	{u n o}
dos	{[dc d | D] o s}
tres	{tc t [e r | r] e s}
cuatro	{kc k w a tc t [o r | r] o}
cinco	{s i N kc k o}
seis	{s e i s}
siete	{s i e tc t e}
ocho	{o tSc tS o}
nueve	{n w e V e}
diez	{[dc d | D] i e s}

Tabla 5.7 Vocabulario de dígitos

Para este conjunto de 20 locutores, 393 frases y de 2 a 9 palabras por frase, el número promedio de frases por locutor fue de 19, y se consiguieron los resultados que a continuación se muestran:

Detalle		**Desempeño**		**Errores**		
Red	**Experimento**	**Fonético (Red. Error)**	**Dígitos (Red. Error)**	**S**	**E**	**I**
28	base	**22.9%**	**66.2%**	2.1%	0%	31.7%
*	tlatoa	**53.8% (40.0%)**	**94.1% (82.0%)**	2.1%	0.1%	3.6%
28	2	**53.6% (39.81%)**	**93.1% (79.58%)**	2.9%	0.0%	4.0%
17	7	**55.4% (42.02%)**	**95.12% (85.5%)**	2.43%	0.05%	2.4%

Tabla 5.8 Resultados finales en dígitos, (*) se incluye la mejor red conseguida hasta ahora por el grupo Tlatoa.

Como se puede apreciar en la tabla anterior, una mejora pequeña a nivel fonético causa una mejora de mayor magnitud en reconocimiento a nivel de palabras. Existe una correlación entre la reducción del error a nivel fonético y a nivel de palabras.

En este capítulo se presentaron los resultados obtenidos al aplicar varias pruebas para evaluar el nivel de desempeño del reconocedor. Todos los resultados aquí mostrados se encuentran también disponibles en internet, en la siguiente dirección: http://info.pue.udlap.mx/~sistemas/tlatoa/rec_fonetico.html.