Conclusiones

Dados los resultados en cada uno de los experimentos de redes neuronales y de modelos ocultos de Markov, podemos decir que hay dos características en las que nos podemos enfocar: la primera es que se puede tomar como un parámetro importante el tamaño del corpus y en segundo lugar un dominio específico para aplicaciones reales.

Los resultados de la red neuronal fueron satisfactorios para corpus de tamaño pequeño como el de dígitos, pero cuando se hicieron los experimentos con el corpus de teléfono, que es considerado de mediano tamaño, el nivel de reconocimiento bajó. Cabe señalar que el corpus de teléfono fue complementado con otro más pequeño para aumentar el número de ejemplos de cada unidad fonética, sin embargo dicho corpus planteaba varios problemas de etiquetado. Sólo se etiquetó con la herramienta *forced-alignment* y no se hizo un reajuste manual que siempre es necesario después de este proceso para tener mayor exactitud en los límites fonéticos. Debido a ello, se afectó negativamente el desempeño final, pero no de manera significativa, como se puede ver en las tablas de resultados mostradas en el capítulo 4.

Para modelos ocultos de Markov el nivel de reconocimiento mejoró respecto a las evaluaciones hechas con el reconocedor basado en redes neuronales. Debido a la naturaleza de los HMM el cual está basado en la probabilidad de reconocimiento entre un estado y otro, se pueden modelar mejor los efectos coarticulatorios especificando el número de contextos que se quieren manejar para cada unidad. (derecho, izquierdo o central). Esto nos da un mayor manejo de cada una de las unidades a reconocer sin tener tantas restricciones que pudieran alterar el reconocimiento que está basado en cada uno de los estados del modelo en general. También el CSLU-HMM permite hacer modelados sujetos a las especificaciones que nosotros propongamos y permitiendo configurar modelos que den una visión real del problema que se abstrae.

En tesis anteriores donde fueron desarrollados reconocedores basados en HMM se demostró que el nivel de reconocimiento era mayor al de un reconocedor basado en redes [ESPINOSA98] utilizando un corpus pequeño de prueba de propósito específico como el de *dígitos*. El objetivo de esta tesis fue el de desarrollar reconocedores con ambos enfoques e incluir un corpus más grande a los que se habían utilizado en proyectos anteriores, además de proponer un esquema de reconocimiento de propósito general.

Limitaciones

Al haber realizado los experimentos podemos ver en las tablas que el desempeño en los reconocedores basados en HMM fueron más altos que en redes neuronales, esto se debe a la naturaleza de un modelo oculto de Markov el cual no condiciona a una unidad (fonema o palabra) respecto a los demás contextos. A diferencia de otros experimentos realizados en TLATOA, los cuales han utilizado corpus más pequeños y por ende una menor complejidad en el entrenamiento, los experimentos que se muestran aquí requieren de un mayor esfuerzo en el área probabilística ya que un buen modelado será uno de los aspectos más importantes dentro del desarrollo de un reconocedor basado en HMM. Por esta razón algunos de los limitantes en esta tesis fueron los siguientes:

- La inexperiencia al tratar de modelar un HMM para cierta gramática establecida, sin tener antecedentes que pudieran ayudar a un mejor desarrollo. Esto debido a que en tesis anteriores no documentaron cada uno de los pasos a seguir para hacer el entrenamiento de HMM.
- La falta de un corpus más completo, es decir, que el etiquetado haya sido lo mejor posible a nivel fonema y con una gran diversidad de frases, no sólo de una región sino de todo un país, para hacerlo más general. En este caso se puede considerar que la base de datos de archivos de voz es de la parte centro-sur de la República Mexicana, pero se añadieron archivos de la parte norte pero se desearía tener de todo el país.
- Tener un dominio específico con suficientes unidades para que el sistema cumpla con los objetivos propuestos y sea menos complejo de manejar. Aquí la cantidad de

unidades fue tanta que el tiempo de entrenamiento duró mucho dando pie a retrasos en cuanto al desarrollo.

Perspectivas

Dado que el corpus de teléfono es muy grande y demasiado difícil para entrenar en el sentido de tener una complejidad computacional considerable, se propone hacer una selección de datos de tal manera que se tenga un dominio específico para aplicaciones ya estudiadas. En caso de que se requiera hacer el entrenamiento con un corpus de esas dimensiones, es necesario aumentar el desempeño de la máquina donde haremos el experimento, ya que de no hacerlo el tiempo requerido será muy largo.

También se propone construir un reconocedor con un enfoque híbrido. Dicho reconocedor puede alcanzar mejores niveles de reconocimiento ya que utilizaría características esenciales del entrenamiento basado en redes neuronales y en modelos ocultos de Markov, limitando aquellas características de ambos enfoques que hacen tener menores niveles de reconocimiento. Estos métodos generalmente usan redes neuronales para estimar la emisión de probabilidades las cuales son usadas por los HMM. Para esto la red neuronal requiere un número de unidades de salida iguales al número de estados en los HMM. Un problema que se presenta en este punto es el tamaño de la red neuronal la cual puede llegar a ser muy cara en cuestión del tiempo de entrenamiento y requerimientos de memoria. Dos de los sistemas híbridos que han tenido éxito son los desarrollados por el departamento de ingeniería de Cambridge (CUED) el cual usa redes neuronales recurrentes, y el otro desarrollado por Instituto Internacional de Ciencias Computacionales en Berkeley, el cual está basado en perceptrones multicapa muy grandes. El problema con estos sistemas es que todos ellos requieren hardware especial que procesen en paralelo para entrenar las redes neuronales en una cantidad de tiempo razonable, además de que requieren una gran cantidad de opciones de parámetros para ser estimados.

Así mismo, se propone implementar un nuevo modelo de red neuronal tipo cascada en paralelo. Este nuevo enfoque mencionado en el capítulo 2, puede mejorar el reconocimiento. Anteriormente, una red de cascada en paralelo fue usada y probada para un vocabulario en japonés. El trabajo a futuro consistiría en sustituir la red actual del CSLU Toolkit (*feed forward*) por la nueva red es de tipo recurrente. Se ha identificado donde se harían los cambios dentro del Toolkit, y se propone seguir utilizando la técnica de procesamiento de señales MFCC para obtener los vectores de características y sólo modificar el algoritmo de entrenamiento *nntrain.c* y librerías asociadas.

Como ya mencionamos antes, el grupo de reconocimiento automático de voz de la UDLA-P cuenta con el CSLU-Toolkit el cual además de permitir el entrenamiento y desarrollo de sistemas con enfoques como redes neuronales y de modelos ocultos de Markov, también permite el entrenamiento de reconocedores con enfoques híbridos. A diferencia de los primeros sistemas, este último no necesita de hardware especial que procese en paralelo (aunque se recomienda) ni grandes cantidades de parámetros de entrada. Un tutorial para entrenar reconocedores con este tipo de enfoque es proporcionado con el ambiente de desarrollo del CSLU-HMM.

Muchas serían las propuestas para cambiar los enfoques que a menudo se han venido manejando dentro del grupo TLATOA. Sin embargo la experiencia de algunos de nuestros colaboradores harían que estas propuestas llegarán a realizarse sin mayor problema. Contando con documentación y código tanto de la red neuronal de cascada paralela como del modelo híbrido [KIRSCHNING98].