Capítulo II

El CSLU *toolkit*

El *Center for Spoken Language Understanding* (CSLU) del *Oregon Graduate Institute* (OGI), localizado en la ciudad de *Portland*, *Oregon*, ha desarrollado una caja de herramientas conocida como el CSLU *toolkit*.

El *toolkit* soporta varios enfoques para el reconocimiento de voz, incluyendo redes neuronales artificiales (RNA) y modelos ocultos de Markov (HMM). El *toolkit* viene con un reconocedor de voz independiente del vocabulario, y contiene varios reconocedores de vocabulario específico (todos ellos para el idioma inglés). También incluye todos los tutoriales y herramientas necesarias para entrenar nuevos reconocedores RNA y HMM.

Este capítulo tiene como objetivo mostrar las herramientas que se utilizarán para la realización de este proyecto de investigación. Primero se presenta la estructura del *toolkit* y la metodología en el proceso de reconocimiento, y después se presenta la relación de las redes neuronales con el *toolkit*.

2.1 Arquitectura del CSLU *toolkit*

El CSLU *toolkit* es un conjunto de herramientas que provee un ambiente poderoso y flexible para el desarrollo de sistemas de voz. El CSLU *toolkit* está diseñado para facilitar el desarrollo rápido de sistemas de lenguaje hablado para una

variedad de aplicaciones, así como también para proveer una plataforma para realizar investigación de tecnologías de voz [Serridge98].

El *toolkit* está formado por dos componentes principales:

- el CSLUsh, un *shell* de programación formado por algoritmos y librerías escritas en C y Tcl, y
- el *Rapid Application Developer* (RAD), un generador rápido de prototipos.

Los prototipos creados en en RAD son traducidos a un conjunto de scripts que se ejecutan dentro del CSLUsh. Debido a que el objetivo principal de esta tesis es evaluar un reconocedor, solo se trabajará en el CSLUsh, ya que no hay necesidad alguna de crear una aplicación en el RAD.

2.2 El proceso de reconocimiento basado en frames

Los avances actuales en reconocimiento de voz tienden a caer en dos categorías, reconocimiento basado en *frames* y reconocimiento basado en segmentos. El reconocimiento basado en *frames* es la técnica utlizada actualmente por la mayoría de los sistemas de reconocimiento de voz. En un sistema basado en *frames*, cada frame de observación en la secuencia $O = \{o_1,\ldots,o_T\}$ recibe un *score* para cada modelo fonético. No existe pre-segmentación de la señal en unidades mas largas. Más aún la segmentación viene implicitamente como una consecuencia del *scoring frame* por *frame.* En un sistema basado en segmentos los principios y los finales de unidades mas largas son hipotetizadas con la señal como un paso distinto en el proceso de *scoring.* Estas unidades mas grandes generalmente representan unidades fonéticas individuales de voz.

En este proyecto todos los sistemas desarrollados estan basados en *frames* ya que es el tipo de enfoque con el que cuenta el CSLU *toolkit.* El proceso del reconocimiento incluye varias acciones: Primero, se digitaliza la señal de voz a reconocer. El siguiente paso es obtener una representación que contenga características de la señal. Estas características son procesadas en secciones llamadas *frames*, donde cada *frame* es una pequeña sección de la señal de voz que contiene un número igual de ejemplos en la señal. En esta tesis los *frames* son de 30 ms en intervalos de 10 ms.

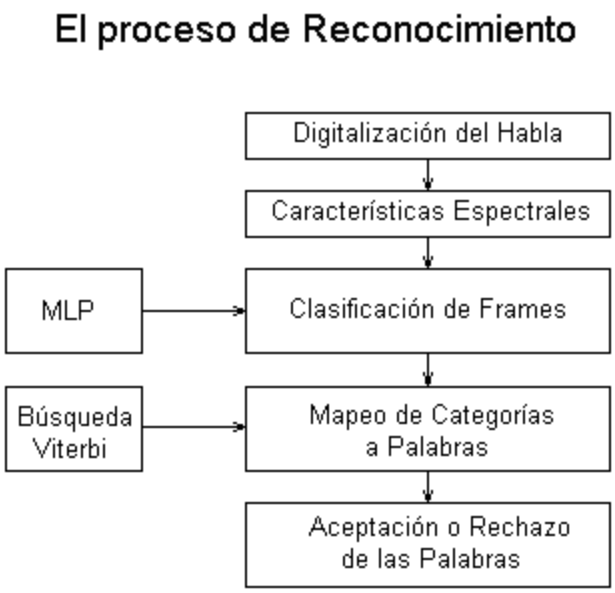

Figura 2.1 El proceso de reconocimiento de voz en el CSLU *toolkit*

Después, se usa una red neuronal para asignar a cada *frame* una probabilidad por cada una de las unidades a reconocer. Se usa búsqueda Viterbi para encontrar la secuencia de unidades con mayor probabilidad de reconocimiento. En las secciones que siguen se describe cada paso con más detalle.

2.2.1 Extracción de Características

De la señal digitalizada se debe obtener una representación espectral para representar las características de cada *frame*. Estas características o parámetros normalmente describen el ambiente espectral en ese *frame* y en un número pequeño de *frames* cercanos.

El proceso de obtención de frames consiste en:

- Elegir una porción de la señal de aproximadamente 30 ms, la cual pertenece al *frame.*
- Crear un vector de parámetros que aproxime el espectro de esa porción.
- Avanzar 10 ms y repetir el proceso hasta haber extraído las características de la señal completa.

La duración del *frame* debe ser lo suficientemente larga para permitir la extracción de las características, pero a su vez lo suficientemente corta para que la señal permanezca estable dentro del *frame.*

Una de las más importantes áreas en el proceso del reconocimiento de voz es el procesamiento de la señal, la cual convierte la forma de onda de voz en algún tipo de representación paramétrica. Esta información paramétrica es usada para análisis posteriores. Entre las técnicas de procesamiento de señales para la obtención de parámetros que soporta el CSLU *toolkit* se encuentran: *Perceptual Linear Prediction* (PLP), *Mel Frequency Cepstral Coefficients* (MFCC), normalización de la energía y supresión de ruido [Schalkwyk96]. Para el presente trabajo la técnica que se implementó fue la de MFCC´s, sobre las cuales hay una explicación mas detallada más adelante.

2.2.1.1 Vectores MFCC de características

Una de las técnicas más utilizadas para la obtención de los parámetros es la codificación MFCC. Al aplicarla, se reduce un número determinado de muestras de la señal de voz a un conjunto de coeficientes que representan las concentraciones de energía y anchos de frecuencia de la señal. Esto se realiza con base a las características de los procesos de producción y percepción del habla. Empleando esta técnica, se obtienen 12 coeficientes MFCC, más una medida de la energía, para cada intervalo de 10 ms. Además, se calcula una medida de el "delta" de estos 13 parámetros para obtener al final un vector de 26 muestras por cada *frame*.

El procesamiento de la señal en este trabajo es típico de la mayoría de los sistemas de reconocimiento de voz. El cepstrum es la transformada inversa del logaritmo del poder espectral de una señal [Rabiner78]. Los MFCC proveen un alto grado de reducción de datos mediante el uso directo de la densidad del poder espectral, a partir de que el poder espectral en cada *frame* puede representarse con relativamente pocos parámetros.

2.2.1.2 Ventanas de contexto

Debido a la naturaleza continua de la voz, una porción de ésta no depende solamente de las características espectrales de un instante en el tiempo, sino que depende de la variación de las características a manera que el tiempo transcurre. Con el objeto de lograr una clasificación más precisa, se toma como entrada al clasificador una ventana de contexto, la cual contiene un *frame* en particular así como los *frames* que se localizan a -60, -30, 30 y 60 milisegundos del *frame* de interés, para tomar en cuenta la naturaleza dinámica de la voz [Hosom98]. Una ventana de contexto se representa con un total de 130 parámetros, 26 por cada uno de los 5 frames incluidos en la ventana.

2.2.2 Clasificación de frames

Esta etapa tiene la función de clasificar las características de cada *frame* en categorías fonéticas usando una red neuronal o modelos ocultos de Markov. El clasificador a usar en este trabajo de investigación es una red neuronal, por lo que mas adelante se profundizará más sobre redes neuronales y su relación con el *toolkit*.

Las salidas de la red neuronal son usadas como una estimación de probabilidad para cada categoría fonética contenida por el *frame*.

2.2.3 La búsqueda Viterbi y la red de pronunciaciones

Una vez que la clasificación fonética termina, se obtiene una matriz de probabilidades para cada uno de los fonemas y para cada *frame*. A esta matriz se le aplica el algoritmo de búsqueda Viterbi para encontrar la secuencia de fonemas con mayor probabilidad de reconocimiento.

El algoritmo Viterbi determina la secuencia de fonemas más probable a partir de la matriz anterior y de las reglas de gramática. Además usa información estadística de las duraciones máximas y mínimas de cada fonema con el objetivo de restringir las opciones. A medida que los límites de duraciones para cada fonema se afinan, el nivel de reconocimiento se incrementa [Rabiner93].

2.3 Modelado dependiente del contexto

Para determinar las categorías que la red clasificará se necesita determinar la pronunciación para cada una de las palabras que serán reconocidas. Entre más precisa sea la pronunciación mayor será el índice de reconocimiento.

En el proceso de producción del habla cada fonema se ve afectado por su contexto. Los fonemas no se producen independientemente debido a que los órganos articulatorios que producen los sonidos se encuentran en constante movimiento y no pueden cambiar instantáneamente de una posición a otra [Vargas, Munive97]. Por ejemplo, si se observa el espectograma de una misma vocal en diferentes contextos, la posición de las formantes varía dependiendo de los fonemas vecinos.

Lo que se pretende mas adelante al usar este tipo de modelado es optimizar el desempeño del reconocedor enriqueciendo los datos de entrada con información contextual.

2.4 División del fonema en partes

Otra cosa muy unida al modelado dependiente del contexto es la división del fonema en partes. Para modelar el dinamismo de los fonemas, dentro del CSLU *toolkit* se les divide de la siguiente manera:

- Una parte: El fonema es independiente del contexto.
- Dos partes: La primera mitad del fonema depende del contexto izquierdo y la segunda del contexto derecho.
- Tres partes: El primer tercio del fonema es dependiente del contexto izquierdo, la parte central es independiente del contexto y el tercio restante depende del contexto derecho.

Cabe señalar que un fonema se podría dividir en partes y todavía podría ser independiente del contexto, o se podría hacer un fonema dependiente del contexto sin dividir en partes.

2.5 Redes neuronales y el CSLU *toolkit*

Una red neuronal es una interconección de elementos simples de procesamiento, cuya funcionalidad se encuentra basada en el modelo del neurón natural. La habilidad de procesamiento de la red radica en la fuerza de intercomunicación, o pesos, obtenidos por un proceso de adaptación, aprendiendo de un conjunto de patrones de entrenamiento [Hosom98].

Los pasos generales para crear un reconocedor basado en redes neuronales son:

- Especificar las categorías fonéticas que la red va a reconocer.
- Encontrar muchos ejemplos de cada una de las categorías en el corpus.
- Entrenar redes iterativamente para reconocer éstas categorías.
- Evaluar las redes en el conjunto de desarrollo
- Evaluar el desempeño de la mejor red utilizando un conjunto de prueba.

Cuando ya se cuenta con un corpus balanceado foneticamente, se debe dividir en partes, una parte para entrenar la red, otra para determinar cual iteración ha sido la que ha aprendido mejor y otra para hacer la prueba final. Estos conjuntos son llamados de entrenamiento, desarrollo y prueba respectivamente.

2.6 Entrenamiento de una red neuronal

Una vez ya obtenidos los 130 parámetros para cada *frame*, éstos deberán ser introducidos en un clasificador. La ventana de contexto se envía como entrada a la red neuronal, y la salida de la red neuronal es una clasificación de cada frame de entrada medido en términos de probabilidades de categorías basadas en fonemas**.**

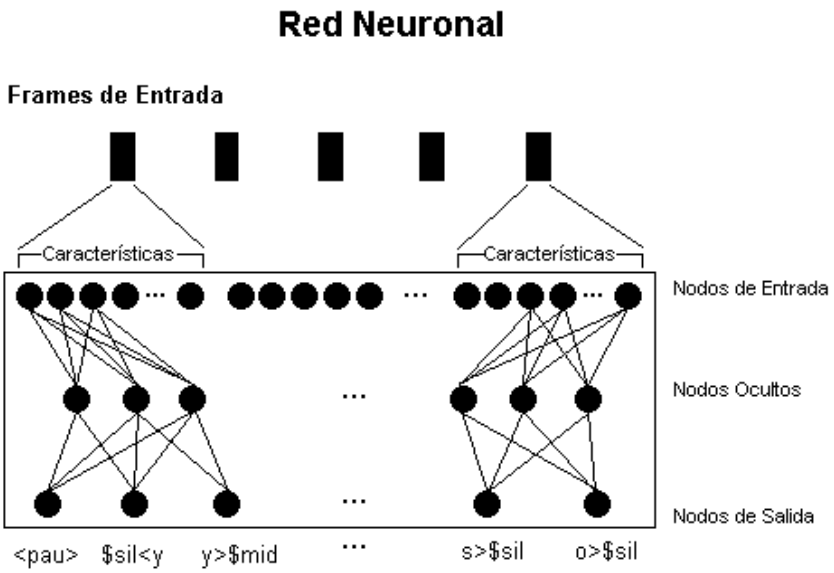

Figura 2.2 Ejemplo del funcionamiento de una red neuronal

Las redes que se utilizan en el *toolkit* son redes de tipo *feed-forward* de tres etapas, las cuales se entrenan utilizando el algoritmo de *back propagation*. El objetivo de este algoritmo es modificar los pesos de la red neuronal por medio de un proceso de entrenamiento iterativo, de tal manera que cada vez se obtengan valores de salida más confiables, la red resultante de cada iteración es guardada en un archivo.