CAPÍTULO

5

Reconocimiento de voz usando redes neuronales artificiales del CSLU

Una vez teniendo los datos listos para ser usados, el paso siguiente es entrenar un reconocedor de voz usando alguna técnica conocida de reconocimiento, para ello empleamos redes neuronales, que son herramientas que también forman parte del Toolkit y de las cuales se hablará en este capítulo.

5.1 El proceso de Reconocimiento

En este apartado se expondrá una de las partes más importantes de este trabajo de tesis, que es el proceso de reconocimiento. Se hablará de como se realiza este proceso en el CSLU.

Hay cinco pasos básicos para que el reconocimiento funcione. Cada paso será explicado enseguida:

Figura 5.1: El proceso de reconocimiento

Primero: Digitalizamos la voz que queremos reconocer.

Segundo: Las características que representan el contenido del dominio espectral de la voz (regiones de fuerte energía en frecuencias particulares) son calculadas. Estas características son calculadas cada 10 milisegundos, con una sección de 10 milisegundos llamada FRAME.

Tercero: Un perceptrón multicapa (o red neuronal) es usado para clasificar un conjunto de estas categorías en categorías basadas fonéticamente de cada frame.

Cuarto: Una búsqueda Viterbi es usada para comparar los promedios de la red neuronal con las palabras objetivo (palabras que posiblemente estén contenidas en la entrada de voz) para determinar la palabra que fue mas probablemente completa.

Quinto: La detección de la palabra se realiza y la confianza que tenemos en la palabra con mayor promedio es medida. La palabra es aceptada si la confianza es suficientemente alta.

La onda digitalizada es convertida a una representación espectral y se hace un proceso para eliminar algunos de los efectos de ruido. Doce coeficientes de frecuencia son calculados, y después se calculan doce características delta que indican el grado de cambio ocurrido en este frame. Las últimas características son energía y energía delta, para un total de 12 + 12 +2= 26 características por frame.

Nos gustaría entonces clasificar cada frame en categorías basadas fonéticamente, pero antes de hacer eso tomamos una ventana de contexto. Esto significa simplemente tomar los frames de interés así como los frames que están a –60, -30, 30 y 60 milisegundos lejanos del frame de interés. Esto es hecho tomando en consideración la naturaleza dinámica de la voz: la identidad de un fonema a menudo dependerá no solamente de las características espectrales a un punto en el tiempo, también dependerá en cómo las características cambian en todo momento.

Mandamos las características en una ventana de contexto a una red neuronal para su clasificación . La salida de la red neuronal es una clasificación del frame de entrada, medido en términos de las probabilidades de las categorías basadas en fonemas. Mandando las ventanas de contexto para todos los frames de voz a una red neuronal, podemos construir una matriz de probabilidades de categorías basadas en fonemas sobre tiempo. En este ejemplo de la salida de la red neuronal, la palabra a ser reconocida es two, y las

regiones obscuras en la t, t&ltu, y las categorías u indican la mayor posibilidad de aquellas clases al tiempo indicado.

Las palabras objetivo son entonces combinadas con la gramática (si hay) para producir una lista de cadenas legales o categorías. Una búsqueda Viterbi es usada para encontrar la mejor ruta a través de la matriz de probabilidades para cada cadena legal. La cadena con la mayor probabilidad es la salida de la búsqueda Viterbi.

5.2 Modelado dependiente del contexto

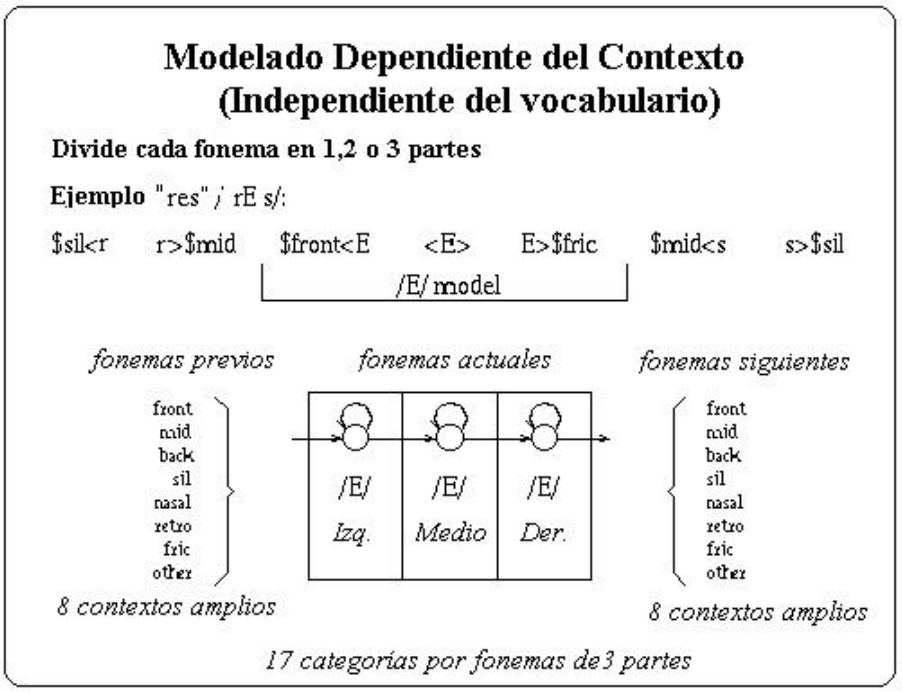

Figura 5.2: Modelado dependiente del contexto

Por lo tanto, ¿qué categorías debería clasificar la red neuronal?. La opción obvia es un fonema por categoría, así que una palabra como “res” debería tener tres categorías: /r/ , /E/, y /s/. Los primeros reconocedores fueron construidos de esta manera, pero fue encontrado que los fonemas tienen como una gran influencia en fonemas vecinos que la /E/ que sigue a la /s/ pueda verse diferente que una /E/ que sigue, dicho de otra manera, a /b/. Por otro lado, para considerar estos efectos coarticulatorios, la estrategia actual es dividir cada

fonema en uno, dos o tres partes, dependiendo en cuanto ese fonema será influenciado por los fonemas de alrededor. El fonema /E/, por ejemplo, puede ser influenciado por los fonemas de la izquierda y derecha, pero puede ser también bastante largo en duración que la parte media no es muy influenciada por fonemas que lo rodean. Como un resultado, separamos /E/ en tres partes.

Si tuviéramos que modelar cada fonema de la izquierda y derecha de cada fonema, hablaríamos de un orden de 5000 diferentes categorías que necesitamos clasificar. Para simplificar esta situación, agrupamos categorías que son similares en uno de ocho amplios contextos. Por ejemplo, /r/ es un fonema vibrante, y así es asignada al contexto amplio "$front". El fonema /s/ es un fricativo, asignado con /f/ y /sh/.

Al contexto amplio "$fric". El símbolo de pesos es una notación utilizada cuando nos referimos a un contexto amplio en lugar de un simple fonema. Así, nuestro modelo de /E/ en "res" llega a ser "$front&ltE" (/E/ en el contexto de proceder de una vocal frontal), &ltE> (/E/ en medio sin efectos contextuales), y E>$fric (/E/ en el contexto de un fricativo siguiente).

Para un sistema independiente del vocabulario, /E/ será dividido en 17 categorías: 8 categorías para cada contexto amplio precedente, 8 categorías por cada contexto amplio siguiente, y una para categorías independientes del contexto. Este método resulta en 543 categorías para todos los fonemas del inglés americano, el cual es computacionalmente más factible.

5.3 Redes Neuronales

La red neuronal tiene 130 nodos de entrada, uno por cada característica en los 5 frames de la ventana de contexto. Hay 200 nodos escondidos, y 545 nodos de salida (un nodo de salida por cada categoría basada fonéticamente, y un nodo extra llamado "basura" que se explicará mas adelante).

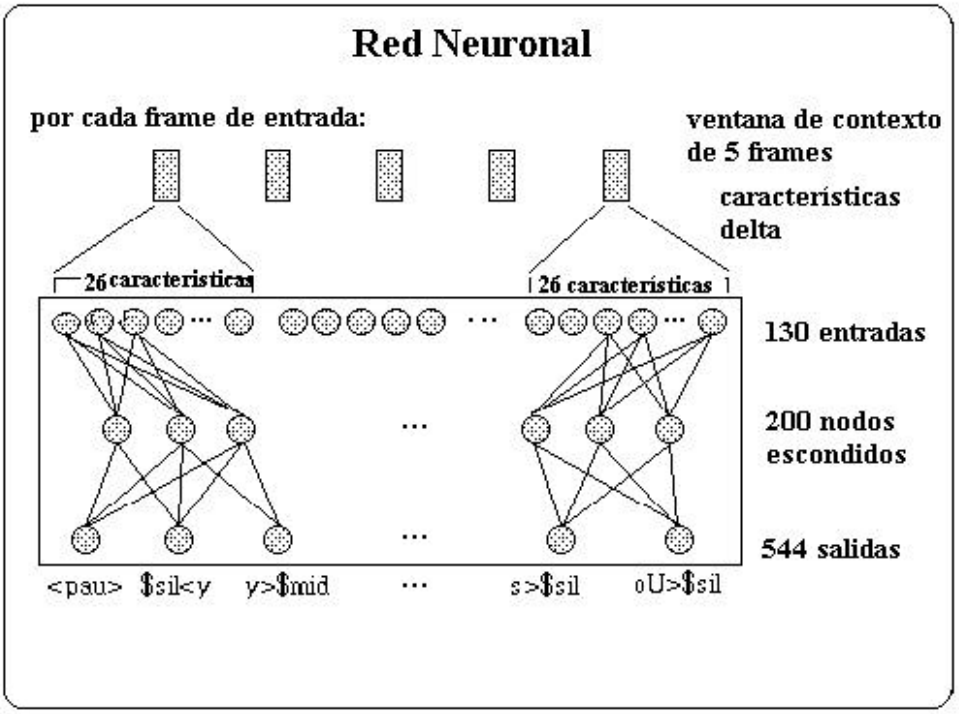

Figura 5.3: Red Neuronal

La red de propósito general usada en OGI fue entrenada usando 1500 ejemplos de cada categoría, con los datos tomados de los números de OGI, Yes/No, Apple, y un corpora de Historias, también como de un corpora llamado NYNEX. Una cosa importante de notar es que las salidas de la red neuronal son usadas como estimados de probabilidades posteriores, en otras palabras, no tomamos justo las categorías con el mas alto promedio y decir , está bien, en el frame 42 encontramos la categoría r>$mid. En cambio, la red es usada para

estimar las probabilidades de cada categoría, así que podemos decir, está bien, en el frame 42 encontramos un 82% de probabilidad de r>$mid, un 7% de probabilidad de r>$front, un 7% de probabilidad de r>$mid, un 3% de probabilidad de &ltiy>, y un porcentaje cercano a cero para todas las otras categorías.

5.4 Búsqueda Viterbi

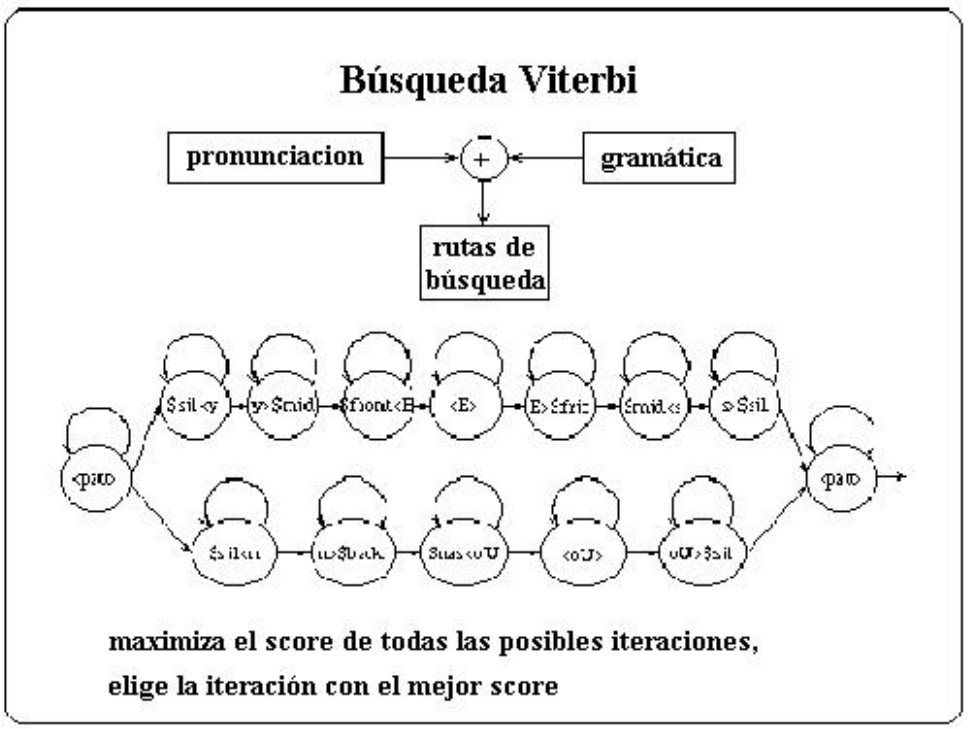

Figura 5.4: Búsqueda Viterbi

Una vez que tenemos una matriz de cómo las probabilidades fonéticas cambian con el tiempo, queremos buscar la mejor palabra. Antes de hacer esto, necesitamos calcular el conjunto de cadenas legales de categorías fonéticas. Este conjunto es dependiente de las palabras que queremos reconocer y el posible orden de las mismas, por eso, combinamos modelos de pronunciación por cada una de nuestras palabras, (por ejemplo, “res” = $sil&lty, r$mid, $front&ltE, &ltE>, E>$fric, $mid&lts, s>$sil) con una gramática si

permitimos que múltiples palabras lleguen en más de un orden. En el ejemplo que se muestra aquí, tenemos una simple ruta de búsqueda que puede reconocer solo uno de los dos, “res” o “no” los cuales tiene que ser precedidos y seguidos de un silencio. En la búsqueda, cuando vemos un nuevo frame, hacemos una transición a un nuevo estado si la probabilidad de la nueva categoría es mayor que la probabilidad de la categoría actual. También realizamos esta búsqueda a través de todas las ramas de la ruta de búsqueda, encontrando las rutas a través de la matriz que maximice el promedio para cada palabra en nuestro vocabulario. Al final de la búsqueda, tenemos promedios de todas las palabras en nuestro vocabulario, y la palabra con el más alto promedio es la palabra mas conveniente para el dato de entrada, y es por lo tanto la palabra con la mas alta probabilidad de ser la palabra que fue pronunciada.

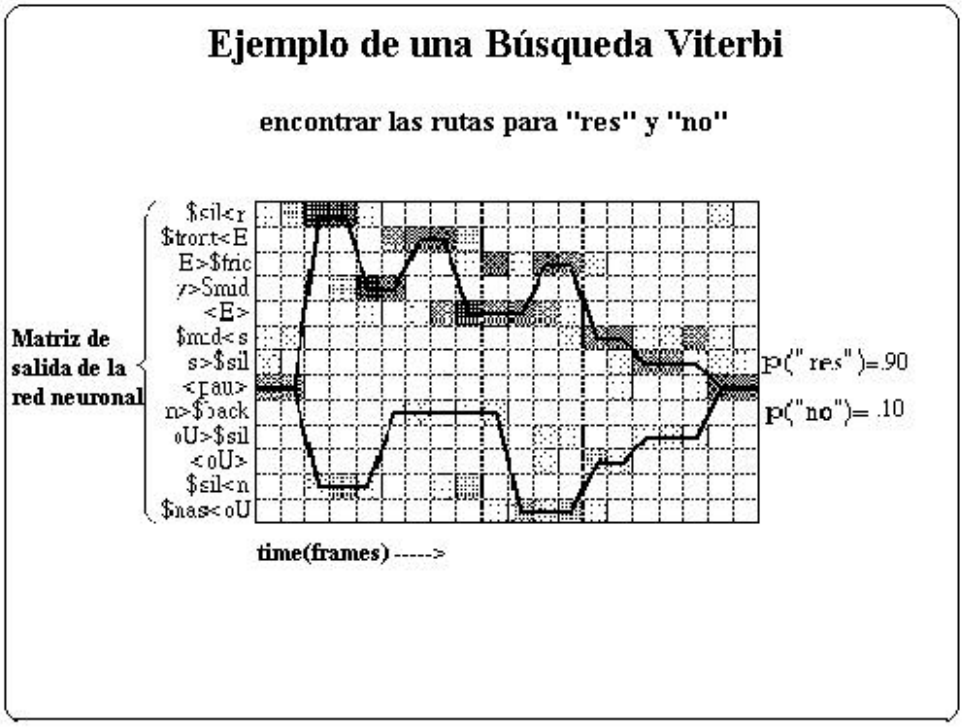

Figura 5.5: Ejemplo de una búsqueda viterbi

5.5 Desarrollo del reconocedor usando redes neuronales y el CSLU Toolkit

Una vez visto el tema de redes neuronales, el siguiente paso es cómo utilizar esta tecnología para el desarrollo de un reconocedor utilizando las herramientas del Toolkit.

5.5.1 Organización de los datos

En el capítulo IV se mencionó cómo se obtuvieron los datos para formar el corpus que se necesitó para el desarrollo de este proyecto. Una vez que se tiene una base de datos recolectada es necesario organizarla en grupos específicos para llevar a cabo el proceso de entrenamiento.

5.5.2 Distribución de los datos

Siguiendo la metodología del Toolkit el proceso de reconocimiento de voz consta de tres fases principales:

- ?? Entrenamiento
- ?? Desarrollo
- ?? Evaluación

Para esto es necesario dividir los datos que se obtuvieron para formar el corpus de niños en tres grupos:

- ?? 60% para entrenamiento (train)
- ?? 20% para desarrollo (dev)
- ?? 20% para evaluación (test)

El primer conjunto de datos se utiliza para entrenar la red neuronal. Es importante que se tengan suficientes muestras de cada categoría para asegurar un buen modelado de los fonemas y por tanto un mejor reconocimiento, ya que con estos datos la red neuronal deberá aprender.

Los datos para la etapa de desarrollo se emplean para evaluar el nivel de reconocimiento de la red en cada iteración. Es necesario que este conjunto de datos sea diferente al de entrenamiento porque se requiere una generalización; es decir, se espera que el reconocedor obtenga un desempeño adecuado para voces diferentes a las del conjunto de entrenamiento, estos datos son elegidos al azar.

5.5.3 Generación de vectores de características

Una vez obtenidas las muestras de cada categoría ya se cuenta con los datos suficientes para generar los vectores de características. Estos vectores corresponden a los parámetros representativos de la señal de voz. El proceso que se lleva a cabo para su obtención es:

- segmentar la señal de voz
- extraer las características esenciales de cada segmento de la señal
- generar los vectores MFCC (coeficientes de multi-frecuencia)
- verificar la información obtenida

Este procedimiento debe realizarse para los datos de entrenamiento, desarrollo y evaluación

Es importante que una vez obtenida la cantidad de muestras se verifique que el número de vectores generados para cada categoría corresponda a la cantidad especificada anteriormente. Esto es, porque cuando los vectores creados exceden la cantidad especificada se ignora el resto; por ejemplo, si fueron especificadas 200 muestras y se obtuvieron 238, se ignoran las 38 sobrantes.
Por otro lado, cuando los vectores no alcanzan el nivel indicado se pueden tener dos casos:

1. El número de vectores es mayor a cero pero menor al límite especificado o,
2. El número de vectores es cero.

En los dos casos, cuando las muestras obtenidas por categoría no son suficientes, significa que esa categoría no es muy común en el corpus por lo que se puede tomar una de tres decisiones:

1. Modificar la división de partes de los fonemas para generar más ejemplos
2. Eliminar esa categoría
3. Atar la categoría con un número pequeño de muestras a otra de similares características pero con suficientes muestras

Si se opta por la segunda alternativa, puede representar un problema, pues significa que no se tienen ejemplos de esa categoría , por lo que la red no va a "aprender" sus características y es posible que en los datos de prueba exista uno o varios ejemplos, y por consiguiente no la podrá reconocer.

Una vez obtenidos los vectores de características (frames) estamos listos para llevar a cabo el entrenamiento de la red neuronal.

5.5.4 Fase de entrenamiento

Esta fase tiene como objetivo que el clasificador "aprenda" las características esenciales de aquello que desea reconocer. Por lo que, es importante que los datos empleados en esta fase sean representativos del medio ambiente. Durante el proceso de entrenamiento, como se mencionó anteriormente, los pesos se van ajustando gradualmente hasta encontrar aquellos que generen la salida deseada.

El proceso de entrenamiento es iterativo, es decir la red será entrenada cierto número de veces el cual puede variar según sea necesario. La red neuronal que se utiliza para el entrenamiento consta de 3 capas:

?? La red neuronal, que tiene 130 nodos de entrada, uno por cada característica en los 5 frames de la ventana de contexto.

?? 200 nodos ocultos. El número de nodos ocultos puede variar, sin embargo se ha observado que este número de nodos es suficiente para obtener un buen resultado. Generalmente no se sabe con precisión el número de nodos ocultos que debe tener una red pero se recomienda que sea el doble al número de categorías que se desean clasificar.

?? Los nodos de salida, que representan las categorías que se desean clasificar. Este valor puede variar dependiendo de las unidades que se desean reconocer. Por ejemplo, en el caso de que se decida tomar a cada fonema como una unidad y se tiene un conjunto de 22 fonemas, se tendrían por lo tanto 22 nodos de salida.

Una vez realizadas las 30 iteraciones, con las redes correspondientes, continuamos con la fase de desarrollo.

5.5.5 Fase de desarrollo

Esta etapa consiste en determinar cuál fue la red que obtuvo el desempeño más alto. En este experimento, la iteración que obtuvo el mejor desempeño fue la número 16 (véase el Apéndice C) con un porcentaje de 96.4% de reconocimiento. Para esto se evalúa el nivel de error alcanzado en cada iteración haciendo uso de los vectores de características generados para esta etapa. Se asume que, cuando se alcanza un nivel mínimo de error, la red está preparada para reconocer las características generales de la voz para la cual fue entrenada.

Una de las técnicas, y la utilizada por el Toolkit, consiste en determinar el nivel de reconocimiento en términos del grado de error generado. Para realizar este cálculo se aplica la siguiente fórmula:

$$E = [(S + I + D) / N] * 100$$

Donde:

N es el número total de palabras en el conjunto de pruebas

S el número de substituciones
I el número de inserciones
D el número de impresiones

A medida que el porcentaje de error disminuye, significa que se tiene un buen nivel de reconocimiento. Este método es aplicado a nivel de palabras y frases.
Al término de esta etapa se obtiene cuál fue la red que tuvo el mejor desempeño y que se utilizará en la última fase.

5.5.6 Fase de Evaluación

Finalmente es necesario efectuar una evaluación final de la red neuronal para medir su desempeño. Para realizar este proceso se emplea el conjunto de datos que se eligió (test), el cual arrojó un porcentaje de reconocimiento de 96.4%.

Este grupo de datos sólo debe tener la transcripción a nivel de texto de cada archivo, por lo que se sabe anticipadamente la salida esperada

Esta fase consiste en probar con el nuevo conjunto de datos la red que obtuvo los mejores resultados en la fase anterior e ir comparando con la salida esperada para obtener el nuevo porcentaje de reconocimiento, de la misma manera que en la fase anterior.