Capítulo 2

2 El Reconocedor SpeechWorks

El reconocedor SpeechWorks es un sistema de reconocimiento de habla continua basado en segmentos e independiente del usuario. El software de SpeechWorks permite el desarrollo de aplicaciones de reconocimiento de habla para la automatización de servicios a través del teléfono. En este capítulo, se describe el reconocedor SpeechWorks en cada uno de sus componentes.

2.1 Introducción

La mayoría de sistemas actuales usan modelos ocultos de Markov (HMMs) para modelar características de las medidas acústicas sobre una secuencia de frames de longitud fija. Estos frames usualmente son muy cortos en duración (entre 5 y 20 ms). Esta duración es mucho más corta que la mayoría de las unidades fonéticas individuales. En el sistema SpeechWorks se usa un conjunto de segmentos de longitud variable que son relacionados a las unidades fonéticas individuales. El reconocedor SpeechWorks está basado en el reconocedor SUMMIT desarrollado en MIT [Zue et al, 1990].

El proceso de reconocimiento tiene tres componentes principales. El primer componente transforma la señal de entrada en una representación acústico-fonética. El segundo ejecuta una expansión de las pronunciaciones dentro de una red léxica, y el tercer componente provee restricciones lingüísticas en la búsqueda a través de la red léxica. En la figura 2-1 se muestra un diagrama a bloques del reconocedor SpeechWorks. Las siguientes secciones dan una breve explicación de los componentes del sistema.

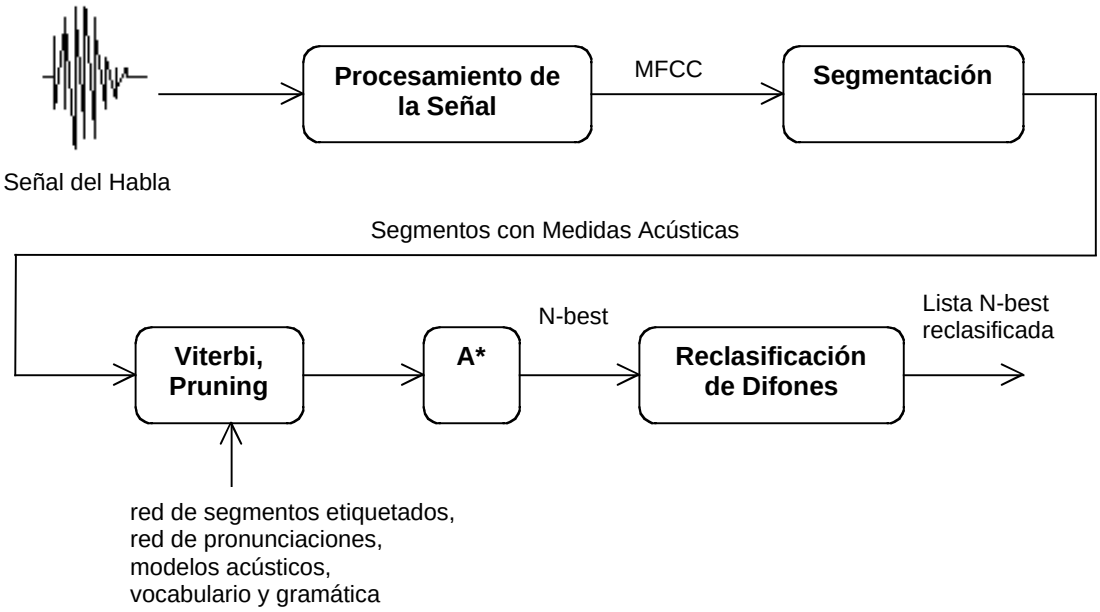

Figura 2-1 Diagrama a bloques del reconocedor SpeechWorks.

2.2 Procesamiento de la Señal

En esta etapa, la señal de entrada es transformada a una representación de coeficientes cepstrales (MFCC). Los MFCCs reciben como entrada la señal y la salida es una representación de como coeficientes cepstrales que representan características relevantes de la señal.

2.3 Segmentación

El proceso de segmentación tiene como meta la identificación de regiones de habla llamadas segmentos, con los que se busca restringir el espacio de búsqueda del reconocedor. El enfoque basado en segmentos es conceptualizado a partir de la representación visual del habla presentada en un espectrograma (ver figura 2-2), donde se aprecia claramente las divisiones entre las regiones relativamente constantes del habla. En las representaciones basadas en segmentos, cada segmento se representa con un solo vector de características.

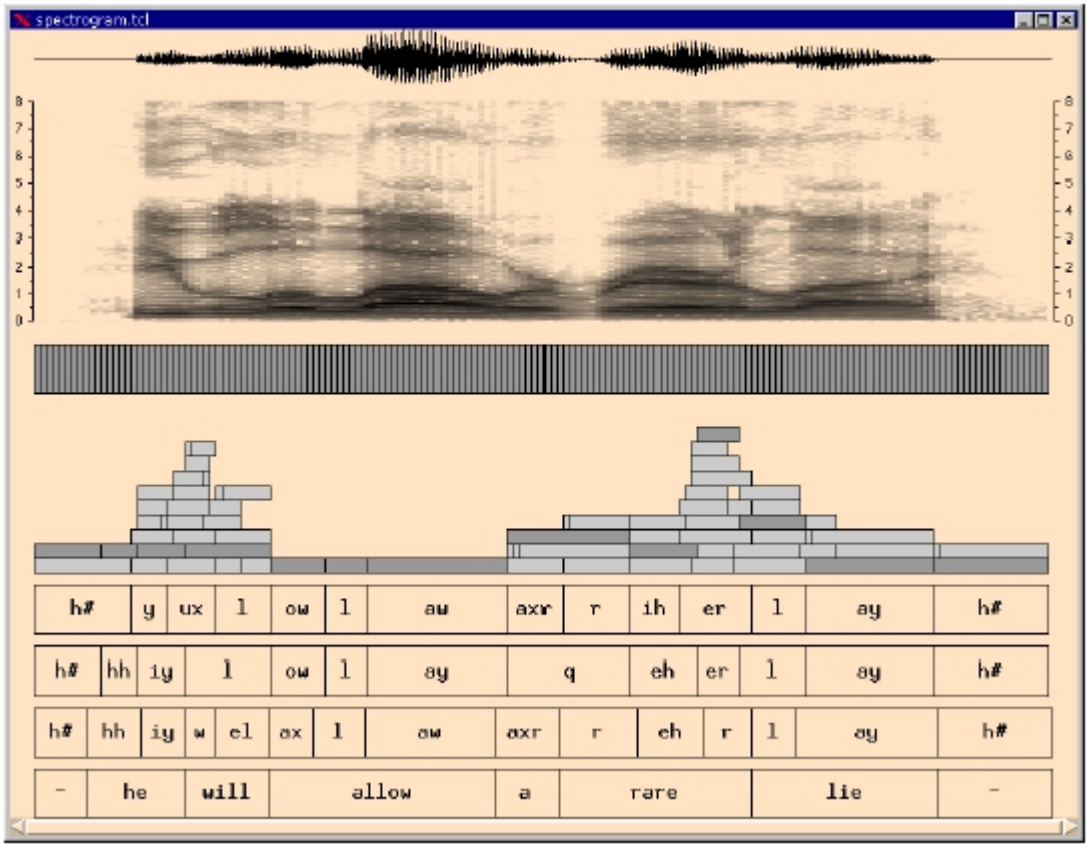

Figura 2-2 Ejemplo de habla, considerando (de arriba hacia abajo) la señal, el espectrograma, la representación basada en frames y segmentos, los posibles caminos, y las transcripciones de fonos y palabras del mejor camino. Este ejemplo fue extraído de [Chang, 1998].

El enfoque basado en frames representa la señal con una secuencia de vectores de características, donde un vector es extraído cada 10 ms. La representación basada en segmentos es un grafo de vectores de características, donde cada vector corresponde a la hipótesis de un fono o segmento. La meta del reconocimiento es la de encontrar el mejor camino a través del grafo de segmentos. En la figura 2-2, los mejores caminos se identifican a través de rectángulos con un gris más oscuro y sus representaciones lingüísticas se muestran en la parte de abajo.

La mayoría de los segmentos son definidos usando la siguiente estrategia:

- Se buscan los cambios acústicos significativos (las fronteras).
- Se conecta cada frontera con las fronteras más prometedoras para formar un conjunto de segmentos.
- Para minimizar el riesgo de omitir un segmento, se postulan más fronteras usando reglas heurísticas, con lo que se produce una red basada en segmentos que será usada posteriormente por los algoritmos de búsqueda Viterbi y A*.

La salida del proceso de segmentación es el conjunto de segmentos encontrados con sus respectivas medidas acústicas, además de la definición de las posibles secuencias de segmentos que pueden representar la señal.

2.4 Modelos Acústicos

Los modelos acústicos de SpeechWorks están basados en mezclas de distribuciones probabilísticas Gaussianas [SpeechWorks, 1999; Rabiner y Juang, 1993] y tienen la tarea de asignar la máxima probabilidad *a posteriori* a cada modelo (fono) para cada segmento. Los modelos acústicos de SpeechWorks están definidos por la función de probabilidad normal descrita por la ecuación 2.1, donde α denota una unidad fonética, M denota el número de mezclas en el modelo y x representa el vector de medidas de un segmento.

$$p(x/\alpha) = \sum_{i=0}^{M} w_i \, p_i(x/\alpha) \qquad (2.1)$$

2.5 Red de Pronunciaciones

La red de pronunciaciones define las posibles pronunciaciones de cada palabra en términos de unidades fonéticas, y también define las posibles transiciones de una palabra a otra. Las pronunciaciones son expresadas como un grafo dirigido, en donde los círculos representan los estados de las pronunciaciones y los arcos representan unidades fonéticas (fonos). En la figura 2-3 se ilustra un ejemplo de una red de pronunciaciones.

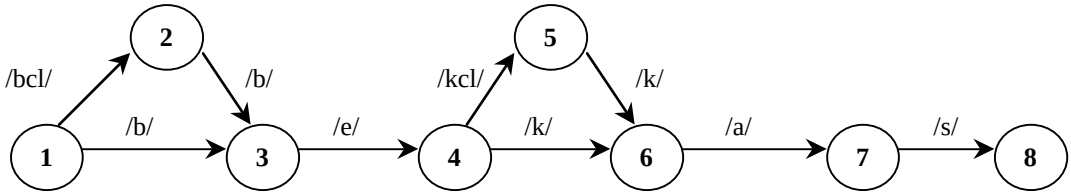

Figura 2-3 Red de pronunciaciones para la palabra “becas”.

2.6 La Búsqueda

En la tarea de búsqueda se usan dos tipos de búsqueda, hacia adelante y hacia atrás. La búsqueda hacia delante usa el algoritmo Viterbi para encontrar el mejor camino en la red de segmentos. La búsqueda hacia atrás usa el algoritmo A* para proveer las *N* mejores hipótesis de la señal de entrada. La razón por la que se usan dos tipos de búsqueda es porque se desea generar eficientemente las N mejores hipótesis (*lista N-best*), y el algoritmo Viterbi solo obtiene la mejor hipótesis. Existen muchas razones por las que se desea obtener más que solo la mejor hipótesis. Considere un ejemplo práctico en donde se desea reconocer un número con n dígitos que debe ser verificado en una base de datos. Suponga que la mejor hipótesis no está en la base de datos, pero la segunda hipótesis sí se encuentra en la base de datos. En este caso, sería mejor confirmarle la segunda mejor hipótesis que decir “*lo siento el número que me dijo no se encuentra en la base de datos*”.

2.6.1 La Búsqueda Viterbi

La búsqueda Viterbi debe encontrar caminos a través de la red de segmentos, asignando a cada segmento una etiqueta fonética, tal que la secuencia de etiquetas forme una sentencia válida respecto a la red de pronunciaciones. De todos los caminos debe ser encontrado un sólo camino con la probabilidad más alta, donde la probabilidad de un camino es la combinación de la probabilidad individual de los segmentos con sus etiquetas fonéticas y la probabilidad de la secuencia completa de acuerdo al modelo del lenguaje [Rabiner y Juang, 1993; Forney, 1973; Ryan y Nudd, 1993].

2.6.2 Viterbi con Podado de Ramas

La búsqueda Viterbi puede ser extendida para ofrecer un eficiente método de podado de ramas (*pruning*). Hay varias formas de ejecutar pruning. Típicamente, en cada punto de tiempo de la búsqueda (frontera en la red de segmentos), se debe tomar una decisión para retener solamente un subconjunto de los nodos en ese punto de tiempo para extensiones de caminos posteriores. Este subconjunto puede ser obtenido al retener los nodos con la probabilidad más alta, o al retener todos los nodos dentro de algún umbral del nodo con la probabilidad más alta en ese punto [Phillips y Goddeau, 1994]. En la practica, el método de pruning disminuye un poco el desempeño, pero a cambio ofrece una disminución significativa de tiempo computacional.

2.6.3 La Búsqueda A*

La búsqueda A* provee las *N* mejores hipótesis de la señal de entrada. El espacio de búsqueda de A* es definido por el mismo espacio de búsqueda usado en Viterbi. Sin embargo, a pesar de que Viterbi es una búsqueda síncrona, la búsqueda A* es una búsqueda primero-mejor (*best first*) que en este caso utiliza los resultados de la búsqueda Viterbi para estimar la distancia mínima que queda a cada camino [Winston, 1992].

2.7 Reordenamiento de la Lista N-best

Aún con algoritmos eficientes como Viterbi, A* y pruning, la búsqueda es una tarea con un alto costo computacional. Miles de caminos deben ser considerados y la cantidad de esfuerzo gastado para cada camino debe ser mantenido lo más bajo posible. Por lo tanto, SpeechWorks incorpora un proceso de post-búsqueda que consiste en reordenar la lista N-best a través de una reclasificación de difones (pares de fonos) [SpeechWorks, 1999], evitando la computación involucrada en los modelos de difones durante la búsqueda. Entonces, el resultado final de una hipótesis de reconocimiento es una combinación entre la búsqueda original y la reordenación de la lista N-best, la reclasificación de difones reduce significativamente el error a un pequeño costo.

2.8 Medidas de Desempeño en SpeechWorks

Un resultado de reconocimiento está compuesto de tres partes: la frase, el resultado de reconocimiento y el resultado de rechazo. Las frases reconocidas se clasifican como dentro del vocabulario o fuera del vocabulario, aunque en algunos casos esta clasificación no es fácil de definir. Una frase es considerada dentro del vocabulario cuando puede ser entendida como parte de un contexto de reconocimiento. Por otro lado, una frase es considerada fuera del vocabulario cuando no forma parte de un contexto de reconocimiento; es decir, todo lo que no está dentro del vocabulario. En la tabla 2-1 se definen las medidas de desempeño para frases dentro del vocabulario del reconocedor y en la tabla 2-2 se definen las medidas de desempeño para las frases fuera del vocabulario.

MEDI-DA	RECONOCI-MIENTO	RECHAZO	DEFINICION
RR_IN	CORRECTO	NO APLICA	Porcentaje de frases reconocidas correctamente antes de aplicar rechazo.
CA_IN	CORRECTO	ACEPTADO	Porcentaje de frases que fueron reconocidas correctamente aceptadas sin confirmación.
CC_IN	CORRECTO	CONFIRMADO	Porcentaje de frases que fueron reconocidas correctamente aceptadas con confirmación.
FA_IN	INCORRECTO	ACEPTADO	Porcentaje de frases que fueron reconocidas incorrectamente aceptadas sin confirmación.
FR_IN	NO APLICA	RECHAZADO	Porcentaje de frases rechazadas incorrectamente.
FC_IN	INCORRECTO	CONFIRMADO	Porcentaje de frases que fueron reconocidas incorrectamente y confirmadas.

Tabla 2-1Medidas de desempeño para las frases dentro del vocabulario.

Una frase es reconocida correctamente si el valor de la salida del reconocedor es igual al valor de la transcripción. Por ejemplo, sí un locutor dice: "*con Alcira Vargas*" y la salida del reconocedor es "*Alcira Vargas*", la frase es reconocida correctamente. En caso contrario, la frase es reconocida incorrectamente.

MEDIDA	RECONOCI-MIENTO	RECHAZO	NOMBRE
CR_OUT	NO APLICA	RECHAZADO	Porcentaje de frases rechazadas correctamente sin confirmación.
FA_OUT	NO APLICA	ACEPTADO	Porcentaje de frases que fueron reconocidas incorrectamente y aceptadas sin confirmación.
FC_OUT	NO APLICA	CONFIRMADO	Porcentaje de frases que fueron reconocidas incorrectamente y confirmadas.

Tabla 2-2 Medidas de desempeño en SpeechWorks para frases fuera del vocabulario.

Las frases dentro del vocabulario (*IV*) usan la regla definida por la ecuación 2.2, y las frases fuera del vocabulario (*OOV*) usan la regla definida por la ecuación 2.3.

$$IV : ca\_in + cc\_in + fa\_in + fr\_in + fc\_in = 100\% \tag{2.2}$$

$$OOV : cr\_out + fa\_out + fc\_out = 100\% \tag{2.3}$$

En la tabla 2-3 se muestran ejemplos de las medidas descritas en las ecuaciones anteriores. Para efectos de estos ejemplos, se supone que las frases "*Nora Munive*" y "*Colegio Murray*" son consideradas dentro del vocabulario y la frase "Nancy Aguas" es considerada fuera del vocabulario.

Medida	Ejemplo
CA_IN	El usuario dice "*Nora Munive*" y el sistema reconoce "*Nora Munive*" sin confirmar.
CC_IN	El usuario dice "*Nora Munive*" y el sistema le responde "*Creo que usted dijo Nora Munive?*".
FA_IN	El usuario dice "*Nora Munive*" y el sistema reconoce "*Colegio Murray*" sin confirmar.
FR_IN	El usuario dice "Nora Munive" y el sistema le responde, "*Lo siento no le entendí*".
FC_IN	El usuario dice "*Nora Munive*" y el sistema le responde, "*Creo que usted dijo Colegio Murray*".
CR_OUT	El usuario dice "*Nancy Aguas*" y el sistema le responde "*Lo siento no le entendí*".
FA_OUT	El usuario dice "*Nancy Aguas*" y el sistema reconoce "*Nora Munive*" sin confirmar.
FC_OUT	El usuario dice "*Nancy Aguas*" y el sistema le responde "*Creo que usted dijo Nora Munive?*"

Tabla 2-3 Ejemplos de las medidas de desempeño.

2.9 Resumen

En este capítulo se explica el reconocedor SpeechWorks y cada uno de sus componentes. Además, se describen las medidas de desempeño usadas por el reconocedor.

El siguiente capítulo describe la estructura experimental usada en esta tesis. Además, se describe un método de evaluación para sistemas ASR de diálogo dirigido, particularmente para el reconocedor SpeechWorks.