Capítulo 1

1 Introducción

El reconocimiento de voz consiste en decodificar a partir de una señal de habla, una secuencia de palabras. La comunicación hombre-máquina a través de la voz es una meta común de muchos investigadores, debido a que el habla es una forma de comunicación muy natural y eficiente. En las últimas cuatro décadas se ha dedicado un gran esfuerzo a la investigación en el área de reconocimiento de voz. Se inició con sistemas de vocabulario limitado, dependiente del locutor y con reconocimiento de palabras aisladas. Estos sistemas únicamente reconocían un pequeño número de palabras, tal como dígitos, los cuales eran entrenados para un usuario en particular o un conjunto de usuarios, y requerían de pausas entre las palabras [Rabiner y Juang, 1993]. Actualmente, existen sistemas independientes del usuario, con reconocimiento de habla continua y con vocabularios grandes, capaces de reconocer decenas de miles de palabras [Markowitz, 1996; Rodman, 1999]. Con estos avances, los sistemas de reconocimiento de habla han incrementado su aplicación en redes telefónicas para automatizar servicios [Zue, 1997]. Los sistemas de hoy día han reducido dramáticamente los porcentajes de error de reconocimiento, pero a pesar del progreso en los últimos años, el reconocimiento de habla aún carece de perfección.

1.1 Definición del Problema

Los sistemas de reconocimiento de habla continua deben mejorar en términos de precisión y robustez. La falta de robustez es un problema muy grande al que se enfrentan los sistemas actuales. Para que un sistema de reconocimiento de habla sea robusto, debe tratar con habla espontánea de diferentes usuarios, habla que típicamente contiene titubeos, pausas, correcciones, palabras fuera del vocabulario, ruidos provocados por el locutor y ruidos en el medio ambiente, entre otros.

La importancia de esta tesis radica en que el habla continua frecuentemente contiene palabras y sonidos que no están definidos en el vocabulario activo del reconocedor, lo cual disminuye el desempeño de reconocimiento. Esta tesis presenta un estudio sobre el problema del habla fuera del vocabulario con el objetivo de hacer que el reconocedor encuentre palabras clave en las frases pronunciadas usando habla espontánea. Tal estudio debe encontrar un punto de equilibrio entre costo computacional y desempeño aceptable.

La tabla 1-1 muestra ejemplos de frases reales con habla fuera del vocabulario que ilustran la problemática de esta tesis.

FRASE	COMPONENTES
"El nombre es A*lcira Vargas*"	Palabras fuera del vocabulario + *palabra clave*
"[mm] *Oleg Starostenko* y me pasó para allá"	[Ruidos provocados por el locutor] + *palabra clave* + palabras fuera del vocabulario
"Este, me podría comunicar con con el maestro *Octavio Cabezut*"	Palabras fuera del vocabulario + *palabra clave*
"Me podría comunicar con la señorita no-*Nora Nunive* por favor"	Palabras fuera del vocabulario + *palabra clave* + palabras fuera del vocabulario
"Con con la señorita *Nancy Aguas* por favor"	Palabras fuera del vocabulario + *palabra clave* + palabras fuera del vocabulario
"Quiero hablar con el doctor este [mm] *Rogelio Dávila* por favor"	Palabras fuera del vocabulario + [Ruidos provocados por el locutor] + *palabra clave* + palabras fuera del vocabulario

Tabla 1-1 Frases pronunciadas en un sistema de conmutador automático donde la palabra clave es lo que se desea reconocer.

El problema de ruido en la línea telefónica, área que pertenece al procesamiento de señales, aunque también de gran importancia, está fuera del alcance de este trabajo. La siguiente sección describe la investigación previa a la problemática del habla fuera del vocabulario.

1.2 Trabajos previos

Antes de los 90's había poca investigación en el área de habla fuera del vocabulario. La mayoría de las investigaciones se enfocaron a reconocer correctamente habla dentro del vocabulario y dentro de una gramática predeterminada, sin preocuparse del desempeño de los sistemas fuera de estas condiciones. A partir de los 90's, se han reportado muchos proyectos de investigación en el desarrollo de sistemas de identificación de palabras clave (*Word Spotting*) para aplicaciones donde la detección de unas cuantas palabras es suficiente para realizar una transacción. Las aplicaciones más comunes que se han desarrollado son: servicios de operador automático [Sukkar y Wilpon, 1993], servicios de sección amarilla y servicios de asistencia [Chigier, 1992]. Estas aplicaciones son implementadas de manera similar, el cambio más significativo siendo el tamaño del vocabulario. Entre las técnicas que cuentan con la habilidad para mejorar el reconocimiento en presencia de habla fuera del vocabulario se encuentran los siguientes:

- Post-procesamiento
    - Niveles de Confianza
- Modelado explícito de habla fuera del vocabulario con
    - Palabras completas
    - Unidades inferiores a la palabra (fonemas y sílabas)
    - Reconocimiento de grandes vocabularios en habla continua (LVCSR)
- Análisis Robusto

1.2.1 Niveles de Confianza

En los sistemas de reconocimiento automático de habla, los niveles de confianza determinan cuando una frase debe ser aceptada, rechazada, o confirmada. Los niveles de confianza representan la probabilidad de que una hipótesis del reconocedor es correcta. Cuando una hipótesis tiene una probabilidad alta, la frase puede ser aceptada sin confirmar, dependiendo del diseñador de la aplicación. Si la confianza es muy baja, la frase debe ser rechazada, pero cuando una hipótesis tiene una probabilidad intermedia, la frase puede ser

confirmada para asegurar el resultado de reconocimiento. Considere un ejemplo de un sistema que ofrece el servicio de conmutador automático, y que hace la siguiente pregunta: "*Por favor diga el nombre de la persona con la que desea hablar*", y el usuario da como respuesta "*Si, me gustaría hablar con Alcira Vargas, por favor*".

Para efectos de este ejemplo, el reconocedor reporta un nivel de confianza de 85% en que la respuesta es "Alcira Vargas". Con el nivel de confianza obtenido se pueden tomar dos decisiones: Aceptar la frase a pesar de la falta de seguridad, o confirmar la frase para asegurar el reconocimiento. Por lo tanto, es claro que los niveles de confianza ayudan mucho a tener un comportamiento razonable para los usuarios. Al no tener esta información, un sistema siempre tendrá que aceptar la frase reconocida o confirmarla.

Los niveles de confianza se han implementado utilizando diferentes enfoques. En redes neuronales se ha implementado el perceptrón multicapa como estimador de niveles de confianza con buenos resultados usando una arquitectura como la que se muestra en la figura 1-1 [Colton, 1997, Weintraub, 1993]. Algunas de las técnicas implementadas para rechazo de frases incluyen el análisis discriminante descendente generalizado (GPD) y análisis discriminante lineal [Sukkar y Wilpon, 1993]. Otras investigaciones usan características a nivel frase para medir la confianza en sistemas conversacionales usando análisis discriminante para seleccionar el mejor conjunto de características [Pao y Schmid, 1998]. Estudios más recientes publican que el análisis discriminante de Fisher es uno de los métodos que presentan mejor desempeño para definir la confianza a un nivel acústico [Kamppari y Hazen, 2000].

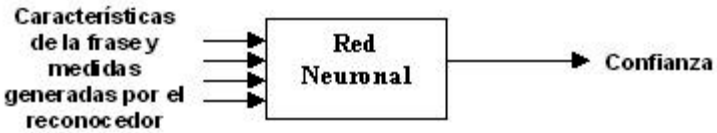

Figura 1-1 Arquitectura de una red neuronal para estimar niveles de confianza.

1.2.2 Identificación de Palabras Clave

El propósito de un sistema de identificación de palabras clave (*Word Spotter*) consiste en detectar palabras clave en habla continua. Existen dos enfoques principales de identificación de palabras clave. El primero consiste en hacer reconocimiento de habla continua y detectar la presencia de palabras clave en el habla reconocida, mientras que el segundo consiste en usar una red de palabras clave y complementarla de modelos *filler* que son entrenados a partir de palabras no clave y ruido de fondo (ver figura 1-2). Las palabras no clave pueden ser fonemas, sílabas o palabras completas.

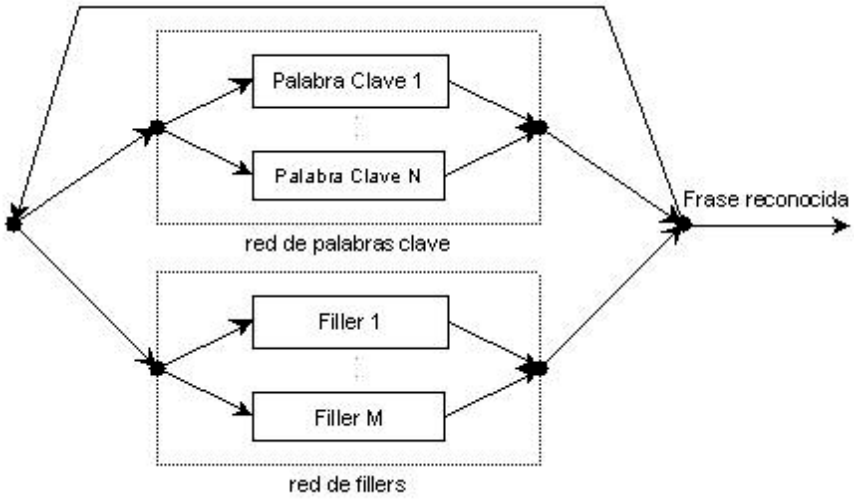

Figura 1-2 Red de reconocimiento para sistemas de identificación de palabras clave.

El primer enfoque es conocido como LVCSR (*Large Vocabulary Continuous Speech Recognition*) y es el enfoque que ha mostrado mejor desempeño a cambio de un costo computacional alto [Manos, 1996; Weintraub, 1993; Rose y Paul, 1990]. El segundo enfoque, aunque tiene un desempeño más bajo que el enfoque LVCSR, cuenta con la ventaja de que no presenta alto costo computacional. En este enfoque varias estructuras de reconocimiento pueden ser usadas, ya sea usando Modelos Ocultos de Markov (HMM) [Rohlicek et al, 1989], reconocedores basados en segmentos [SpeechWorks, 1999; Zue et al, 1990], redes neuronales [Rabiner y Juang, 1993; Álvarez et al, 1993] o una combinación de estructuras [Lippmann y Singer, 1993].

También existe la posibilidad de combinar LVCSR con modelos “filler”. Tal combinación no es muy prometedora cuando los modelos “filler” usan palabras completas, pero es más prometedora cuando se usan modelos “filler” de sub-palabras [Lau, 1998]. Por otro lado, [Chen et al, 1998] propone una estructura de identificación de frases clave (*Key-Phrase Spotting*). Este enfoque requiere poco conocimiento del dominio de interés y se subraya que es capaz de extraer fragmentos clave de las frases. Se remarca que este método sobresale en comparación a los sistemas de identificación de palabras clave comúnmente usados. En [Hetherington, 1995] se propone un enfoque diferente a los anteriores, en donde se examina la habilidad de los modelos acústicos, modelos de pronunciación y los modelos de lenguaje para incorporar nuevas palabras al vocabulario. Por último, [Chung y Seneff, 1997] proponen una técnica auxiliar para los sistemas de identificación de palabras clave, empleando duración como post-procesador para desambiguar palabras acústicamente similares, el cual mejora significativamente el desempeño.

1.2.3 Análisis Robusto

El análisis de lenguaje hablado es considerado más desafiante que el análisis de contexto escrito. A pesar del esfuerzo de investigación en las ultimas décadas, prácticamente ningún parser de texto no restringido ha sido desarrollado [Cole et al, 1997]. Un parser robusto debe resolver al menos los siguientes problemas, los cuales crean severas dificultades en parsers convencionales que usan algoritmos de parsing estándar.

Chunking: Consiste en la segmentación apropiada de texto en unidades sintácticamente analizables.
Desambiguation: Consiste en seleccionar una unidad de un gran número de unidades sintácticamente iguales.
Undergeneration: Consiste en tratar casos de entrada fuera del léxico del sistema.

El análisis robusto permite que una frase sea analizada por fragmentos con la finalidad de incrementar la habilidad de entender el habla espontánea que frecuentemente está fuera de la gramática de un reconocedor. En la década de los 80's, se desarrollaron varios enfoques de análisis robusto para resolver el problema de palabras fuera del vocabulario. La mayoría de los enfoques son de sintaxis relajada [Karlsson et al, 1995] y restricciones gramaticales [Menzel, 1995].

Estudios más recientes se basan en el enfoque basado en el algoritmo generalizado de Tomita (GLR*) [Hayes et al, 1986]. En [Lavie, 1995] se diseña un parser para ser robusto en situaciones donde las frases se encuentran fuera de la gramática por ruido en la entrada o errores en la gramática. GLR* intenta vencer estas situaciones al ignorar palabras no-analizables. En [Kaiser, 1999] se presenta un enfoque basado en máquinas de estado finito. Su enfoque es motivado por la creencia de que un parser de estado finito puede proveer un vehículo eficiente para integrar conocimiento lingüístico de alto nivel en el reconocimiento de habla. Otros enfoques combinan robustez sintáctica y semántica usando el formalismo de gramática restringida para expresar restricciones lingüísticas, sintácticas, semánticas y pragmáticas [Karlsson et al, 1995; Menzel, 1995].

1.3 Objetivos de la tesis

De acuerdo a los trabajos previos y de acuerdo a los recursos disponibles, en esta tesis se proponen los siguientes objetivos en un esfuerzo por modelar habla fuera del vocabulario.

Objetivos Generales

- Hacer que el reconocedor encuentre palabras clave en frases con habla fuera del vocabulario y a la vez rechace las frases que contienen únicamente habla fuera del vocabulario.
- Proponer una técnica que permita modelar habla fuera del vocabulario con un punto de equilibrio entre costo computacional y desempeño aceptable.

Objetivos Específicos

- Clasificar un corpus de desarrollo para realizar experimentos y afinar técnicas.
- Evaluar el corpus de desarrollo que servirá como punto de partida.
- Experimentar con las siguientes técnicas en el corpus de desarrollo:
    - Niveles de Confianza
    - Identificación de palabras clave modelando habla fuera del vocabulario con fonemas, sílabas y palabras
- Medir el desempeño y costo computacional de las técnicas experimentadas con el corpus de desarrollo.
- Evaluar el desempeño y costo computacional de las mejores técnicas en el corpus de prueba.

1.4 Resumen

En este capítulo se explica brevemente el reconocimiento de voz y su situación actual. Se expone que la problemática a la que se enfrenta esta tesis es el habla fuera del vocabulario. A partir de esto se describe la investigación previa sobre los trabajos relacionados en este tópico. Entre las técnicas que destacan como trabajos previos se encuentra la técnica de niveles de confianza, identificación de palabras clave y análisis robusto. Por último se describen de manera clara los objetivos generales y específicos de esta tesis, donde se especifica que las técnicas que se van a investigar y experimentar en esta tesis son: niveles de confianza e identificación de palabras clave.

El siguiente capítulo explica el reconocedor SpeechWorks y cada uno de sus componentes, haciendo énfasis en los componentes que afectan a esta tesis.

En el capítulo 3 se describe la estructura experimental y se explica la preparación del corpus de entrenamiento y del corpus de prueba. Además, se explica la metodología de evaluación usada para medir el desempeño de las técnicas a experimentar.

En el capítulo 4 se describe la técnica de niveles de confianza y se propone un método para definir los niveles de confianza más adecuados para un sistema específico. En el capítulo 5 se explica la técnica de identificación de palabras clave usando fonemas como fillers. Se hace una clasificación de los fonemas hablados del español hablado en México y se experimentan dos variantes con esta técnica. En el capítulo 6 se experimenta con la técnica de identificación de palabras clave usando sílabas como fillers. Se hace una clasificación de las sílabas usadas por esta técnica y también se experimentan diferentes variantes. En el capítulo 7 se explica la técnica de identificación de palabras clave usando palabras como fillers, también se experimenta esta técnica con diferentes variantes. En el capítulo 8 se evalúan los mejores experimentos con las técnicas investigadas usando el corpus de prueba. Esta tesis concluye con una discusión de los experimentos realizados y se comenta la dirección del trabajo a futuro.

Al final de esta tesis se incluye un glosario de la terminología usada comúnmente en el área de reconocimiento de voz, particularmente se definen los términos usados en este trabajo de tesis.