UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA
INGENIERÍA ELÉCTRICA – PROCESAMIENTO DIGITAL DE SEÑALES

RECONOCEDOR DE PALABRAS CONTINUAS Y LINGÜÍSTICAMENTE CONFUSAS PARA EL ESPAÑOL HABLADO EN MÉXICO

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN INGENIERÍA

PRESENTA:
OSCAR FRANCISCO NAVARRETE TOLENTO

TUTOR PRINCIPAL
DR. JOSÉ ABEL HERRERA CAMACHO, FACULTAD DE INGENIERÍA

MÉXICO, D. F. OCTUBRE 2013

JURADO ASIGNADO:

Presidente:	DR. GERARDO EUGENIO SIERRA MARTINEZ
Secretario:	DR. BORIS ESCALANTE RAMÍREZ
Vocal:	DR. JOSÉ ABEL HERRERA CAMACHO
1 er. Suplente:	DR. FELIPE ORDUÑA BUSTAMANTE
2 d o. Suplente:	DR. PABLO ROBERTO PEREZ ALCÁZAR

Lugar o lugares donde se realizó la tesis: LABORATORIO DE PROCESAMIENTO DE VOZ, FACULTAD DE INGENIERÍA, UNAM.

TUTOR DE TESIS:

DR. JOSÉ ABEL HERRERA CAMACHO

_____

FIRMA

(Seg u n d a ho j a)

Agradecimientos

A mis padres, Oscar Armando Navarrete y Maya y María Estela Tolento Maciel, quienes con su ejemplo, sus consejos y su apoyo, me han motivado para superarme y para ser una mejor persona.

A mis hermanos, Eduardo y Diana Navarrete Tolento, que siempre están ahí cuando los necesito.

A mi alma mater, la Universidad Nacional Autónoma de México, que me ha permitido tener una formación integral.

A la Facultad de Ingeniería, por brindarme sus instalaciones para tener un mejor desempeño en mi profesión.

A mi tutor, el Dr. José Abel Herrera Camacho, por su paciencia, su apoyo y todos sus consejos al realizar este trabajo.

A mis profesores, por brindarme sus conocimientos y consejos.

Al CONACYT, por la beca otorgada para realizar mis estudios de maestría.

Al laboratorio de procesamiento de voz de la FI de la UNAM, por facilitarme sus instalaciones para la realización de este trabajo.

A mis abuelos, tíos y primos, por su ejemplo y su apoyo cuando lo necesito.

Al M.I. Jaime Alfonso Reyes Cortés, por su amistad, su compañerismo y el apoyo que me brindó para realizar este trabajo.

Al Ing. Carlos Andrés Acosta Ramos, por su amistad y apoyo, durante nuestra estancia en la maestría y en el laboratorio de procesamiento de voz.

A mis compañeros de la maestría y del laboratorio de procesamiento de voz, por su amistad a lo largo de mi estancia.

A todos mis amigos, que, aún en la distancia, me han brindado su apoyo y afecto.

Índice General

Índice de Figuras

Índice de Tablas

Resumen del trabajo

El desarrollo de este trabajo tuvo como objetivo central conocer el sistema Sphinx y con base en este, y en otras técnicas existentes, diseñar un sistema propio de reconocimiento de voz continua. Ya que, aunque se ha trabajado con este en México, aún no se ha diseñado un sistema de reconocimiento de voz continua en el idioma español.

Con la finalidad de conocer mejor las señales de voz, primeramente, se presentan algunos de los fundamentos de producción de la voz, haciendo un análisis de los mecanismos anatómicos y fisiológicos que intervienen, y mencionando como va cambiando con respecto a la edad de las personas; se establece una clasificación de los sonidos y se muestra el estándar del alfabeto fonético internacional.

Para el desarrollo del proyecto se empleó el corpus del laboratorio de procesamiento de voz de la FI UNAM y se empleó en el sistema Sphinx 3. Posteriormente, con la finalidad de mejorar el desempeño del reconocedor, se conformó otro corpus realizando nuevas grabaciones, considerando el diccionario de lenguaje existente y buscando aumentar el número de repeticiones de cada una de las palabras. Se consiguió que el desempeño del sistema Sphinx 3, usando el nuevo corpus en las pruebas, mejorara con respecto al desempeño empleando el corpus original. De una tasa de error de frase de 100% se redujo a 87.5% y de una tasa de error de palabra de 90.3% se redujo a 41.5%.

Posteriormente, se buscó mejorar el resultado haciendo uso de las redes neuronales como método adicional para realizar el reconocimiento, particularmente, de las palabras fonéticamente confusas y continuas, se comparó el resultado con el método por cuantización vectorial, y se probaron varias configuraciones y modelos. De acuerdo con los resultados obtenidos y su análisis, se constató que las redes neuronales en conjunto con los parámetros empleados en este desarrollo no conllevan una mejora en el índice de reconocimiento, especialmente comparando con los resultados obtenidos empleando cuantización vectorial, de manera que, queda como posible trabajo a futuro realizar pruebas considerando parámetros de diversa naturaleza para observar el desempeño del sistema y con ello determinar si alguno de estos nos lleva a un mejor desempeño.

Introducción

Definición del problema y justificación

A lo largo de las últimas décadas hemos observado el gran avance que han tenido los desarrollos relacionados con el reconocimiento de voz; desde los años 60’s en los que ya se podía reconocer pequeños vocabularios (del orden de 10 - 100 palabras) de palabras aisladas, las mejoras implementadas durante los años 70´s permitieron reconocer vocabularios medianos (del orden de 100 - 1000 palabras), la introducción de métodos estadísticos aplicados al reconocimiento de voz en los años 80´s que posibilitaron el manejo de vocabularios grandes (del orden de más de 1000 palabras), hasta los desarrollos actuales, sin embargo no se ha desarrollado un sistema que efectúe el reconocimiento de voz continua para el idioma español de la región central de México.

Aunque ya se ha trabajado en México con el sistema Sphinx aun no se ha desarrollado un sistema de reconocimiento de voz continua en el idioma español por gente de nuestro país, ya que el entrenamiento y la codificación solamente se han realizado mediante las aplicaciones que incluye el sistema Sphinx. En el área de procesamiento de voz del posgrado de ingeniería eléctrica existen desarrollos previos de reconocimiento de voz, por lo que es de nuestro interés emplearlos para desarrollar un sistema completo de reconocimiento de voz continua.

Debemos destacar que los sistemas de reconocimiento no son perfectos, de tal manera que tienen mucho que mejorar. Uno de los más usados es el Sphinx, mismo que al ser usado con sus propias bases de datos para el inglés resulta bueno ya que las señales que utiliza para reconocer son escogidas y muy probablemente fueron adquiridas en una sala de grabación libre de ruido. Sin embargo, el usar este sistema en la práctica nos otorga resultados pobres, haciendo poco viable su utilización.

Aunando al desarrollo propio del sistema, me motivan las aplicaciones que le podríamos dar, por ejemplo:

- En automatización de sistemas que requieran del reconocimiento del español, tales como software de dictado, sistemas de reservación de una agencia de viajes, software para telefonía celular, control de iluminación, etc.

- En sistemas multimodales, para recibir instrucciones e interactuar con el usuario, como sería el caso de un edificio inteligente o un robot.

- En minería de datos, para la extracción de información relevante sobre una conversación o un discurso.

Objetivos

- Conocer el sistema Sphinx y con base en el diseñar un sistema propio, basándose en técnicas existentes para realizar el reconocimiento de voz.

- Realizar un diccionario de lenguaje para el español de México.

- Revisar el funcionamiento de las redes neuronales para realizar el reconocimiento de las palabras fonéticamente confusas y continuas.

- Determinar, con base en los resultados, si es conveniente usar las redes neuronales para el reconocimiento de las palabras confusas.

Antecedentes y estado del arte

El reconocimiento automático de voz

Es fácil notar que, en general para las personas, la voz es la manera más natural para comunicarse. El reto que tiene la tecnología es desarrollar sistemas que sean capaces de recibir información de manera oral y que acto seguido puedan reaccionar de manera adecuada y coherente a partir de ésta información. Inclusive, los sistemas podrían solicitar información adicional haciendo uso de la síntesis de voz. De acuerdo con este enfoque, el reconocimiento de voz se puede considerar como un tópico de la inteligencia artificial, es decir, el diseño e implementación de máquinas que tengan la capacidad de escuchar, comprender y reaccionar con base en la información oral que se les de cómo entrada y que además sean capaces de hablar, con lo que ya se podría establecer un dialogo [1].

Es muy probable que hayamos leído o visto, libros, relatos o películas de ciencia ficción, con lo que nos resultará más fácil el imaginar las aplicaciones que tendrían este tipo de tecnologías en la vida cotidiana; sin embargo, no debemos olvidar que en algunos casos aún no contamos con las capacidades técnicas para llevarlas a cabo de manera satisfactoria.

Es conveniente marcar la diferencia entre dos tipos procesos principales: por un lado tenemos lo que se conoce como reconocimiento automático de voz (“Automatic Speech Recognition, ASR”), que hace referencia al proceso mediante el cual una computadora establece una relación entre una señal acústica de voz con su correspondiente representación en texto; mientras que, por otro lado, está la comprensión automática de voz (“Automatic Speech Understanding, ASU”), que se refiere al proceso mediante el cual una computadora establece una relación entre una señal acústica de voz con alguna idea o significado [2].

Es entonces que tenemos varios tipos de sistemas de reconocimiento automático de voz. Por un lado tenemos a los sistemas de reconocimiento de voz dependientes del locutor, cuyo propósito es trabajar con un conjunto ya entrenado de hablantes o usuarios. Por lo general estos sistemas son más simples y por ende más fáciles de desarrollar, resultan más baratos y más precisos, sin embargo no son tan flexibles como los sistemas de

reconocimiento adaptables al locutor, o los independientes del locutor. Por otro lado, están los sistemas de reconocimiento de voz independientes del locutor, cuyo propósito es funcionar con cualquier hablante de un determinado tipo dentro de ciertas especificaciones, que no haya sido entrenado por el sistema, un ejemplo sería un hablante del idioma español de la ciudad de México. Este tipo de sistemas resultan más complejos y por ende son más difíciles en su desarrollo, además de que, resultan más costosos y su precisión, la mayoría de las veces, es más baja que en sistemas dependientes del locutor; aunque tienen la ventaja de que permiten una mayor flexibilidad que los demás. También existen los sistemas de reconocimiento de voz adaptables al locutor, que están habilitados para adecuar su funcionamiento de acuerdo con las características que presenten los nuevos locutores. Estos últimos se encuentran en un nivel intermedio de complejidad en relación con los dos anteriores.

Considerando la complejidad, tenemos a los sistemas de reconocimiento de palabras aisladas, que trabajan con una palabra a la vez; dicho de otra forma, necesitan que exista una pausa entre cada palabra. Estos sistemas implementan los métodos más simples de reconocimiento de voz ya que las palabras son más fáciles de delimitar y además la pronunciación de una palabra no genera afectación a las otras.

La mayor complejidad la encontramos en los sistemas de reconocimiento de voz continua (“Continuous Speech Recognition, CSR”), en donde el usuario tiene permitido pronunciar una frase o mensaje casi sin restricciones. Primeramente, el reconocedor debe ser capaz de manejar los límites temporales de una señal acústica, mismos que son desconocidos. En segundo lugar, deberá tener un buen desempeño en presencia de los efectos coarticulatorios, así como, con las variaciones de articulación; incluyendo inserciones y omisiones de palabras, que se presentan en el lenguaje natural.

Al contrario de los sistemas de reconocimiento de palabras aisladas, los sistemas de reconocimiento de voz continua no requieren de la cooperación del locutor; por lo que deben compensar las fuentes de error, haciendo uso de algoritmos más robustos, que sean capaces de trabajar con una gran cantidad de pequeñas variaciones en el lenguaje continuo.

Los sistemas “CSR” representan una forma más natural de hablar desde el punto de vista del usuario y resultan esenciales en una gran cantidad de las aplicaciones en las que varias personas interactúan con el sistema de reconocimiento. Aquellos que cuentan con un vocabulario extenso deben ser entrenados con unidades de subpalabras, como son, los fonemas, las semisílabas o las sílabas, y las relaciones entre palabras se deben aprovechar para maximizar el desempeño. Resulta entonces comprensible que se tengan mayores requerimientos en el caso de voz continua, ya que para modelar las palabras es necesario considerar las variaciones fonológicas dentro y entre las palabras, sin olvidar el explotar las relaciones probabilísticas entre las unidades de subpalabras.

Reconocimiento de voz continua de vocabulario grande

A lo largo de las últimas décadas, se han llevado a cabo varios avances en el diseño de sistemas modernos de reconocimiento de voz continua de vocabulario grande (“Large Vocabulary Continuous Speech Recognition, LVCSR”), a tal punto que su aplicación se encuentra desde los primeros sistemas de dictado a los sistemas automáticos independientes del hablante, para la transcripción y el indexado de noticias, lecturas, reuniones, reconocimiento de voz legal y médico, y aplicaciones en un “call center”, por mencionar algunas. El uso comercial de esos sistemas es un testimonio impresionante de que tan lejos ha llegado la investigación en "LVCSR”.

Debemos decir, de cualquier manera, que a pesar del uso comercial, el problema del reconocimiento de voz para un gran vocabulario está lejos de ser solucionado, debido a: ruido ambiente, distorsión en los canales de transmisión, acentos extranjeros, voz casual, o cambios inesperados en el tema de conversación, lo que aún causa errores garrafales. Esto se debe a que los sistemas actuales de “LVCSR” no son robustos a fallas durante el alineamiento de las frases o a las condiciones en que se llevan a cabo las pruebas, y no pueden manejar el contexto tan bien como lo puede hacer un escucha humano a pesar de haber sido entrenados con miles de horas de voz y millones de palabras en texto [3].

Arquitectura general para un sistema de reconocimiento de voz

Comúnmente los sistemas de reconocimiento de voz actuales están basados en una arquitectura como la que se observa en la figura 1. Adicionalmente, las aplicaciones se comunican con el decodificador para así apreciar los resultados del reconocimiento, mismos que pueden ser utilizados para adaptar algún otro componente del sistema.

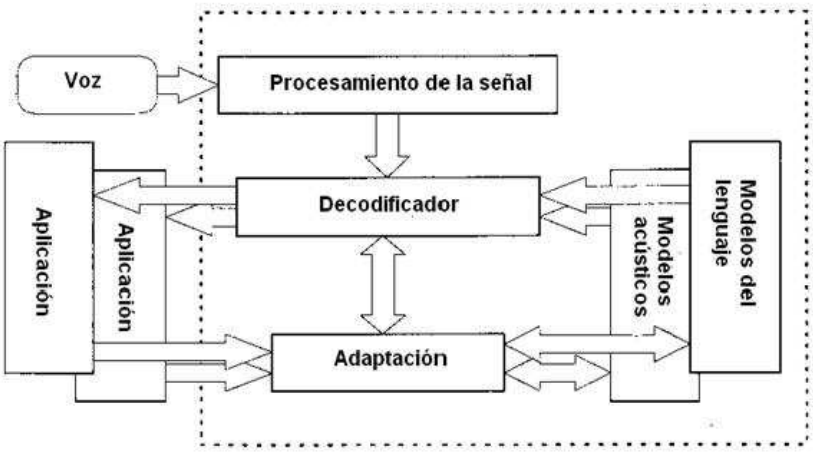


Figura 1. Arquitectura de un sistema de reconocimiento de voz [4].

Los llamados modelos acústicos pueden considerar varios aspectos, tal como acústica, fonética, variaciones del ambiente y del micrófono, el género, los posibles acentos del los hablantes, entre otros. Los modelos del lenguaje consideran lo que constituye una posible palabra, si son parecidas o con qué frecuencia se presentan. Para que se considere como confiable y eficiente a un sistema de reconocimiento deberá ser capaz de lidiar con todas estas cuestiones.

Arquitectura de un sistema de voz continua

Se han logrado mejoras tecnológicas en todas las aéreas del "LVCSR": procesamiento del "front-end", modelado acústico, modelado de lenguaje, búsqueda de hipótesis, y combinación de sistemas. En la figura 2 podemos observar los componentes de un sistema "LVCSR".

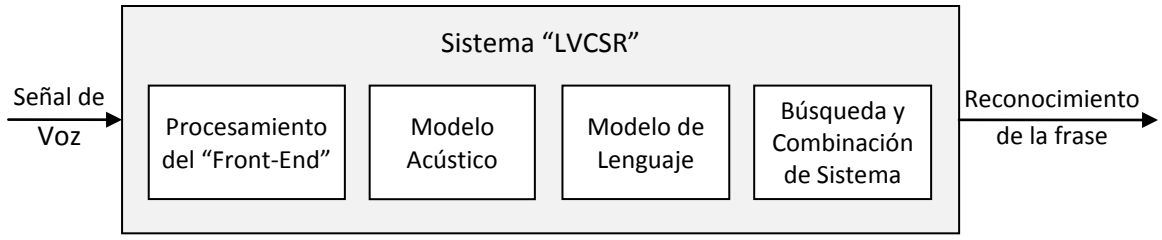


Figura 2. Los cuatro componentes de un sistema "LVCSR" [3].

Dificultades que generalmente se presentan en un sistema de reconocimiento de voz

Detección de inicio y fin.
Las señales de voz tienden a concatenarse, dicho de otro modo, al comunicarnos resulta difícil determinar el inicio y el fin de una palara o frase, así como también, delimitar los sonidos que conforman dicha palabra.

Variabilidad en la voz.
Las señales de voz son variables, dicho de otro modo, dos señales de voz que contengan la misma palabra o la misma frase son distintas incluso habiendo sido pronunciadas por el mismo hablante.

Ambigüedad en el habla. Es decir, hay palabras cuya pronunciación es similar o idéntica, sin embargo su significado es diferente (palabras homófonas), y por otro lado una misma palabra puede tener dos significados dependiendo del contexto en que se encuentre.

Presencia de ruido en la señal de voz. Dado que en el ambiente en que se trabaja existen ciertas perturbaciones que modifican las señales y que interfieren de tal modo que no es sencillo eliminarlas, resultando más difícil la tarea del reconocimiento.

Complejidad del habla. Esta cambia según el grupo de individuos, la región en que se habite, con quien se esté hablando, e incluso el estado de ánimo del hablante.

El tamaño del vocabulario. También llamado corpus del sistema de reconocimiento de voz. Este modifica su complejidad, los requerimientos de procesamiento y la precisión del sistema.

De manera somera, de acuerdo con la cantidad de palabras que manejan, los sistemas actuales se clasifican en alguna de las siguientes categorías:

- Vocabularios pequeños (10 a 100 palabras)
- Vocabularios en los cuales las palabras son habladas en aislamiento de manera deliberada (pueden exceder las 10000 palabras)
- Vocabularios que aceptan voz continua pero que están limitados en sus áreas de aplicación específica, por ejemplo, mensajes que ocurren en un sistema de reservación (típicamente compuestos de 1000 a 5000 palabras)
- Sistemas de reconocimiento de oraciones basadas en trigramas con ausencia de ruido (compuestos por 20000 palabras o más)

Un sistema de reconocimiento de voz resultará exitoso en la medida que sea capaz de sortear los problemas anteriormente planteados.

El sistema de reconocimiento Sphinx

Sphinx es uno de los sistemas de reconocimiento más versátiles que existen. Fue desarrollado por el grupo Sphinx en la "Carnegie Mellon University, CMU" [5] y está basado en la construcción de modelos ocultos de Markov.

Se trata de un sistema de reconocimiento de voz continua, con un gran vocabulario e independiente del locutor, y que está desarrollado bajo licencia "Berkeley Software Distribution, BSD" [6]. Cuenta con varias herramientas y recursos que están en código abierto, conocidos por el nombre de proyecto CMU Sphinxproject, los cuales permiten a los investigadores y desarrolladores implementar sistemas de reconocimiento de voz.

El proyecto CMU Sphinxproject tiene en desarrollo una serie de reconocedores de voz para construir aplicaciones de reconocimiento de voz de alto desempeño, y en los años recientes también incluye recursos relacionados con el reconocimiento de voz, tales como un entrenador acústico, un entrenador del modelo de lenguaje, así como modelos acústicos previamente entrenados [7].

Existen varios proyectos desarrollados en CMU Sphinx, entre otros, destacan:

1. Sphinx 2. Que es un reconocedor de voz de gran vocabulario y de alta velocidad, comúnmente utilizado en sistemas de diálogo y sistemas de aprendizaje de pronunciación.

2. Sphinx 3. Que es un reconocedor de voz un poco más lento que Sphinx 2, pero más preciso. Generalmente se utiliza como una implementación de servidor de Sphinx o para evaluación.

3. Sphinx 4. Que es una versión completa de Sphinx escrita en Java, la cual proporciona una alta precisión y al mismo tiempo una alta velocidad de desempeño.

4. PocketSphinx. Que es un reconocedor de voz que puede ser utilizado en sistemas embebidos.

Detección de palabras confusas

Las palabras confusas son un problema en el reconocimiento automático de voz, "ASR", dado que son una fuente de errores. El vocabulario y la gramática de muchas aplicaciones pueden diseñarse para evitar las palabras confusas y de este modo reducir los errores por confusión. Usualmente esto se realiza probando la aplicación con un vocabulario seleccionado y cambiando las palabras que causan demasiados errores. Dado que este proceso puede ser tedioso, se han realizado varias propuestas para calcular la confusión o perplejidad entre dos palabras sin tener que probar el sistema. Se usa comúnmente "Dynamic Time Warping, DTW" para calcular una distancia entre dos transcripciones

fonéticas. Se utilizan dos diferentes tipos de alineamiento en los trabajos: con o sin inserciones y supresiones. Se ha buscado una medida de perplejidad entre "Hidden Markov Models, HMM" que minimice el error esperado durante el entrenamiento, considerando la información tanto del modelo de lenguaje como del modelo acústico [8].

También se ha propuesto una medida de disimilitud entre transcripciones fonéticas, llamada "Phonetic Acoustic Dissimilarity, PAD" [9], la cual representa una alternativa al empleo de "DTW". En este caso las transcripciones fonéticas se alinean con programación dinámica, haciendo uso de la información fonética de manera independiente de los datos acústicos. Posteriormente, la distancia acumulada se calcula usando la alineación resultante. Todas estas medidas pueden ayudar a conformar una lista de palabras confusas. En varios casos lo que se propone es tomar una decisión empleando una medida de disimilitud entre palabras y clasificarlas en dos clases: palabras confusas y no confusas. Esta estrategia proporciona una herramienta poderosa que puede ser usada para evitar considerar a palabras como confusas cuando no lo son, lo que nos puede representar una mejora en el índice de reconocimiento.

Extracción de patrones característicos y reconocimiento de voz empleando redes neuronales.

Las técnicas de reconocimiento actuales hacen posible que la voz pueda ser empleada para verificar su identidad y controlar el acceso a servicios tales como marcación por voz, banca telefónica, correo de voz, control de seguridad para áreas de información confidencial, acceso remoto a computadoras, entre otras [10]. Uno de los pasos más importantes en los sistemas de reconocimiento de voz radica en extraer cierta información importante de las señales de voz. Los parámetros acústicos de las señales de voz usadas en las tareas de reconocimiento han sido estudiados e investigados en gran medida, lo que ha permitido categorizarlos en dos tipos de procesos: un grupo de parámetros basado en las propiedades espectrales y otro basado en las series de tiempo dinámicas.

Existen varias técnicas de extracción de características que se han desarrollado y que pueden usarse. Esto incluye la transformada rápida de Fourier, "FFT", la predicción lineal

perceptual, “PLP”, los coeficientes cepstral de predicción lineal, “LPCC”, y los coeficientes cepstral en escala mel, “MFCC”. Los “mel-frequency cepstral coefficients, MFCC”, introducidos por Davis y Mermelstein, son quizá los parámetros característicos más empleados en los sistemas de reconocimiento de voz. Esto se puede atribuir probablemente a que los “MFCC” consideran para su adquisición el modelado del sistema de audición humano con respecto a varias frecuencias.

Se puede destacar que, el reconocimiento del alfabeto hablado se utiliza como estándar de comparación para muchos de los sistemas de reconocimiento. Incluso algunos de los sistemas de reconocimiento de palabras aisladas emplean dígitos hablados para probar su desempeño. Del mismo modo, el reconocimiento del alfabeto hablado puede tener varias aplicaciones, entre ellas destacan: asistencia en directorios automatizados, números de teléfono y códigos postales. Si bien este tipo de reconocimiento puede verse como una tarea fácil para los humanos, desafortunadamente, para las máquinas esto puede representar todo un reto debido a las grandes similitudes acústicas entre ciertos grupos de letras. Las grandes similitudes acústicas pueden causar dificultades en la clasificación, mientras que, las similitudes acústicas bajas permiten a los sistemas de reconocimiento de voz una mayor facilidad para discriminar entre las clases.

Otra de las áreas que se ha desarrollado en fechas recientes, con el objetivo de que resulte más fácil el reconocimiento de las letras confusas del alfabeto, contempla el reconocimiento de voz por medio de redes neuronales y algún tipo de parámetros característicos. Recordando que la parte final en un sistema de reconocimiento automático de voz es la etapa de clasificación y que esta etapa conlleva clasificar las señales de entrada para determinar si estas corresponden a la señal objetivo. Uno de los modelos de redes neuronales que comúnmente se usan para las tareas de clasificación es la red neuronal “back-propagation” con una regla de aprendizaje adaptivo [11].

Alcances

En el desarrollo del presente trabajo se lograron las siguientes metas:

- Usar el corpus del laboratorio de procesamiento de voz de la FI UNAM y emplearlo para reconocer en el sistema Sphinx. De tal manera que, tomemos estos primeros resultados como "baseline".

- Conformar otro corpus realizando nuevas grabaciones de frases, considerando el diccionario de palabras existente y buscando tener al menos 3 repeticiones de cada una de las palabras que conforman el corpus. Esto con el fin de mejorar significativamente el porcentaje de reconocimiento.

- Disminuir las tasas de error de frase y error de palabra empleando Sphinx 3.

- Emplear las redes neuronales como método adicional para realizar el reconocimiento de las palabras fonéticamente confusas y continuas, y compararlo con algún otro método. Realizando pruebas con diferentes configuraciones de redes neuronales.

- Establecer las bases de posibles desarrollos a futuro para el reconocimiento de las palabras continuas y confusas.

Estructura del trabajo

A lo largo del presente trabajo se mostrará la puesta en marcha de un sistema de reconocimiento de voz continua, haciendo énfasis en las palabras fonéticamente confusas.

En el primer capítulo comenzamos por ver un panorama general del papel del lenguaje en la comunicación humana, después, se presentan los fundamentos de producción de la voz, se establece una clasificación de los sonidos de la voz y se muestra el estándar del alfabeto

fonético internacional. Es pertinente destacar que, a lo largo del desarrollo del mismo, se hace un análisis de los mecanismos anatómicos y fisiológicos de producción de la voz. Además, algo muy importante es que se menciona como va cambiando la voz con respecto a la edad de las personas.

En el segundo capítulo se hace énfasis en el funcionamiento del sistema Sphinx. Primeramente, se detalla cómo ponerlo en funcionamiento presentando una guía, para después emplear el corpus del laboratorio de procesamiento de voz de la FI UNAM. Debemos destacar que el desarrollo planteado en la tesis se realiza para el español de México, lo cual es de gran importancia ya que los desarrollos existentes son contados. Entre estos desarrollos se encuentra el efectuado por el grupo del Dr. Pineda, IIMAS UNAM. Sin embargo, de acuerdo con lo encontrado, consideramos que la base de datos con la que cuentan es pobre y adicionalmente es "privada". En otros casos se hace uso de HTK, que es otro software de reconocimiento de voz; pero, dado que en el laboratorio de procesamiento de voz de la FI UNAM ya se contaba con otros desarrollos previos, nos inclinamos por trabajar con Sphinx.

Hasta ahora los resultados obtenidos en el laboratorio de procesamiento de voz de la FI UNAM habían sido muy pobres, se tenía una tasa de error de frase de 100% y una tasa de error de palabra de 90.3%, lo cual fue una razón más para continuar trabajando en el tema. Esto nos llevó a proponer conformar un nuevo corpus cuya naturaleza nos permitiera mejorar el desempeño de reconocedor Sphinx, por lo que se grabaron nuevas frases y se etiquetaron para su utilización. Se indica cuál fue la metodología para conformar el nuevo corpus en la búsqueda de mejorar el porcentaje de reconocimiento de voz, se muestran los resultados que se obtuvieron empleando el nuevo corpus para el reconocedor de voz y se comparan los resultados con el corpus existente.

En el tercer capítulo se hace una revisión de los tópicos de redes neuronales haciendo énfasis en aquellos relacionados con el reconocimiento de voz. Primeramente, se presentan algunos conceptos generales, algunas de las diferentes funciones, las arquitecturas y los modelos comúnmente empleados en las redes neuronales. Posteriormente, se mencionan los

inicios de las redes neuronales en el reconocimiento de voz. Finalmente, se hace énfasis en las técnicas de predicción lineal y "Backpropagation", mismas que en este caso se emplean en el reconocimiento de las palabras confusas.

A lo largo del cuarto capítulo se observan las pruebas realizadas y los resultados obtenidos haciendo uso de diferentes configuraciones y funciones de redes neuronales para el reconocimiento de voz. Al mismo tiempo se muestran los resultados de reconocer empleando "Vector Quantization, VQ", técnica que para este caso tomamos como "baseline". Dada la extensión de las pruebas realizadas con redes neuronales se decidió reportarlas en todo un capítulo. Posteriormente, se hace un análisis de los resultados obtenidos.

Finalmente en el capítulo cinco se presentan las conclusiones obtenidas con base en la realización del presente trabajo y en capítulo seis se enlistan las referencias del mismo.

I. Fundamentos de la producción de la voz

1.1. Conceptos generales

1.1.1. Comunicación y lenguaje

Los sistemas de comunicación son capaces de transportar información. El procesamiento de voz se dedica al estudio de un sistema de comunicación específico, es decir, a través de la voz, o dicho de otro modo, señales acústicas tradicionalmente emitidas y recibidas por seres humanos de manera oral. Algunos de los objetivos de este procesamiento son: la representación, análisis, modificación, mejoramiento de la relación señal/ruido de las señales acústicas; la generación de voz de manera artificial o sintética y el que propiamente atañe a este trabajo que es el reconocimiento automático de voz.

A lo largo de la historia y ya desde la antigua Grecia se han realizado intentos por producir voz de manera sintética. Si bien, en algunos casos, se trataba solamente de arreglos de tuberías conectadas entre sí a un locutor humano; en otros casos se trataba de mecanismos acústicos realmente ingeniosos que lograban reproducir sintéticamente algunos sonidos vocálicos [12]. No fue sino hasta el desarrollo de la telefonía a principios del siglo XX que se motivaron diversas investigaciones sobre las propiedades de la voz y la audición, con el fin de mejorar la calidad de la comunicación telefónica. El proceso continuó y ahora las tecnologías existentes permiten, por ejemplo, disponer de sistemas de comunicación oral hombre máquina [12].

En todo sistema de comunicación están presentes diversos componentes: emisor, receptor, mensaje, código, canal y contexto [12]. Es necesario conocer algunos aspectos de cada uno de ellos para poder integrar sistemas que funcionen de manera eficaz y eficiente. En nuestro caso, el emisor es el conjunto integrado por el cerebro que “piensa” el mensaje y el aparato fonatorio que lo “traduce” a una señal acústica. El receptor es el aparato auditivo que recibe la onda sonora y la transforma en impulsos nerviosos que luego son interpretados por el cerebro. El mensaje es la idea a comunicar. El código es el lenguaje hablado. La combinación del mensaje y el código constituyen la señal. El canal puede ser el medio en el cual se propaga la onda sonora (en general el aire).

1.1.2. Algunos conceptos sobre lenguaje

La ***lengua*** es un sistema de signos lingüísticos que permiten la comunicación en una comunidad. Es un sistema pues cada uno de sus elementos tiene entidad propia y entidad relativa a su posición o relación con los otros elementos. Es un código de signos. Tiene carácter social, ya que es común a una sociedad.

El ***habla*** es el acto de seleccionar los signos de entre los disponibles y organizarlos a través de ciertas reglas. Materializa el código. Es individual, por lo que cambia de un individuo a otro. Los signos pueden corresponder al lenguaje escrito o al oral.

El ***lenguaje*** es un sistema articulado ya que los sonidos y otros componentes se integran entre sí. Está formado por signos lingüísticos, nombre que recibe la señal en el lenguaje.

El lenguaje tiene modalidades regionales llamadas ***dialectos***.

Un ***signo*** es algo que reemplaza a otra cosa para comunicarla en un mensaje.

Los signos lingüísticos se clasifican en dos tipos: ***significado*** y ***significante***. El significado es el concepto mental, idea o contenido a comunicar. El significante es la imagen, ya sea gráfica o acústica que se le asigna.

La relación entre significado y significante es arbitraria o convencional, aunque no necesariamente discrecional: involucra acuerdos tácitos, explícitos o normativos en una comunidad lingüística.

En el lenguaje escrito, el significante es la grafía escrita, formada por combinaciones de letras, en tanto que el lenguaje hablado es su realización acústica mediante la palabra hablada.

Las ***palabras*** son los elementos libres mínimos del lenguaje. La sintaxis es el conjunto de reglas para la coordinación de las palabras en frases u oraciones. En su versión escrita las

palabras están formadas por letras o grafemas, es decir unidades graficas mínimas, y, en el caso oral, por fonemas.

Los ***fonemas*** son las unidades teóricas básicas del lenguaje. Se materializan a través de los sonidos, pero de una manera no unívoca. Las variantes de los fonemas se denominan alófonos.

Los ***monemas*** son unidades mínimas con significado, que puede ser gramatical, dando origen a los morfemas, o léxico, representado por los lexemas. Los morfemas tienen relación con la gramática, o la forma de organizar o dar estructura a las categorías básicas del lenguaje (género, número, tiempo o persona de los verbos, etc.), mientras que los lexemas se refieren a significados externos al lenguaje mismo.

Las palabras constan de al menos un monema, siendo las más comunes bimonemáticas, que incluyen un lexema y un morfema. En la tabla I-1 se observan dos ejemplos en los que se identifican los componentes de la palabra.

Tabla I-1. Ejemplos de monemas, grafemas y fonemas [12].

Palabra	Monemas		Grafema	Fonemas
	Lexema	Morfema		
Gato	Gat	o	G, a, t, o	/g/, /a/, /t/, /o/
Amaban	Ama	ban	A, m, a, b, a, n	/a/, /m/, /a/, /b/, /a/, /n/

1.1.3. Fonología y fonética

La fonología estudia los fonemas, es decir, el modelo fonológico convencional e ideal del lenguaje. La fonética, por otro lado, se refiere a los sonidos en el habla, incluyendo su producción acústica, así como los procesos físicos y fisiológicos de emisión y articulación involucrados.

Así, la fonología es el estudio de los sonidos de la lengua en cuanto a su carácter simbólico o de representación mental. Procede detectando regularidades o recurrencias en los sonidos

del lenguaje hablado y sus combinaciones, y haciendo abstracción de las pequeñas diferencias debidas a la individualidad de cada hablante y de características suprasegmentales como la entonación, el acento (tónico, es decir por aumento de la intensidad y agógico por aumento de la duración), etc. Cada uno de los sonidos abstractos así identificados es un fonema. Uno de los objetivos de la fonología es acotar al máximo la cantidad de fonemas requeridos para representar cada idioma de una manera suficientemente precisa.

La fonética estudia experimentalmente los mecanismos de producción y percepción de los sonidos utilizados en el habla a través del análisis acústico, articulatorio y perceptivo. Se ocupa, por consiguiente, de las realizaciones de los fonemas.

1.2. La voz humana

La voz es el sonido producido voluntariamente por el aparato fonatorio humano. Es el instrumento más antiguo y más natural con el que se puede comunicar y hacer música.
Según las leyes de la acústica, hay tres elementos indispensables para la producción de un sonido: un cuerpo vibrante, un medio elástico que propague las vibraciones y una caja de resonancia que las amplifique, a fin de que puedan ser percibidas por el oído.

La voz se produce gracias a la acción coordinada de casi todo nuestro cuerpo. El aparato fonador o vocal está integrado por estructuras musculares de diferentes regiones y por elementos del aparato respiratorio y del aparato digestivo.

Es importante hacer notar que ninguna estructura de nuestro cuerpo tiene como función única ni primaria la producción de la voz. La voz fue una adaptación evolutiva muy posterior a otras acciones imprescindibles para la vida. Así, la laringe, a la que relacionamos de forma automática con la voz, tiene como función principal la de protección de las vías respiratorias. Muchos animales poseen pliegues vocales y no emiten sonidos. En las aves, la laringe no interviene en la producción de sonidos, ya que éstos se originan en la siringe, que se localiza en el extremo inferior de la tráquea.

Cuando hablamos de fonación, hacemos referencia a la voz hablada y cantada, ya que ambas utilizan los mismos mecanismos para su producción, aunque, debido a sus características especiales la voz cantada usará los elementos del aparato fonador de modo más controlado.

1.3. Producción de la voz

La mucosa laríngea forma dos pares de pliegues, como se observa en la figura I-1. El par superior denominado cuerdas vocales falsas y el par inferior, cuerdas vocales verdaderas, al que se alude simplemente como cuerdas vocales. El espacio que hay entre los pliegues ventriculares recibe el nombre de rima vestibuli. Los dos senos (ventrículos) laríngeos son expansiones laterales de la porción media de la laringe, entre las cuerdas vocales falsas y las verdaderas, como se puede observar en la figura I-2.

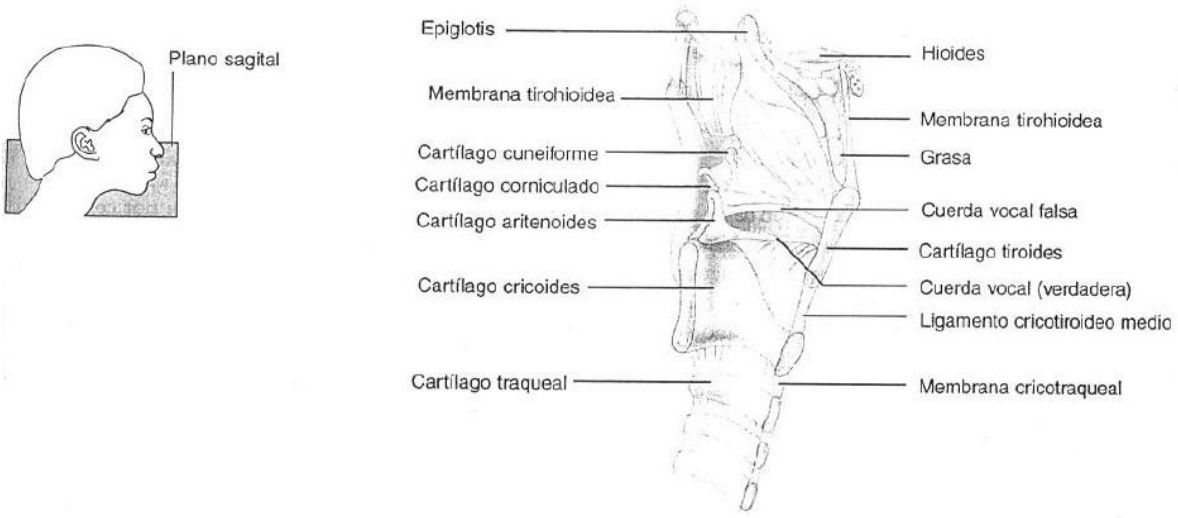


Figura I-1. Corte sagital de la Laringe [13].

Cuando las cuerdas vocales falsas se aproximan entre sí, participan en la contención del aliento que opone resistencia a la presión que hay en la cavidad torácica, como ocurre cuando una persona se esfuerza para levantar un objeto pesado. En plano profundo a la mucosa de las cuerdas vocales verdaderas, formadas por epitelio escamoso estratificado no queratinizado, se encuentran las bandas de ligamentos elásticos estiradas entre los cartílagos rígidos como las cuerdas de una guitarra. Los músculos esqueléticos de la laringe, es decir, sus músculos intrínsecos, se insertan en los cartílagos y las cuerdas

vocales verdaderas, y cuando se contraen, tensan los ligamentos elásticos y estiran estas cuerdas de tal modo que sobresalen las vías respiratorias; esto hace que se angoste la rima glottidis. Así, el aire dirigido contra las cuerdas vocales las hace vibrar con lo que generan ondas sonoras en la columna de aire que fluye en la faringe, nariz y boca. Cuanto mayor sea la presión del aire, tanto más fuerte el sonido.

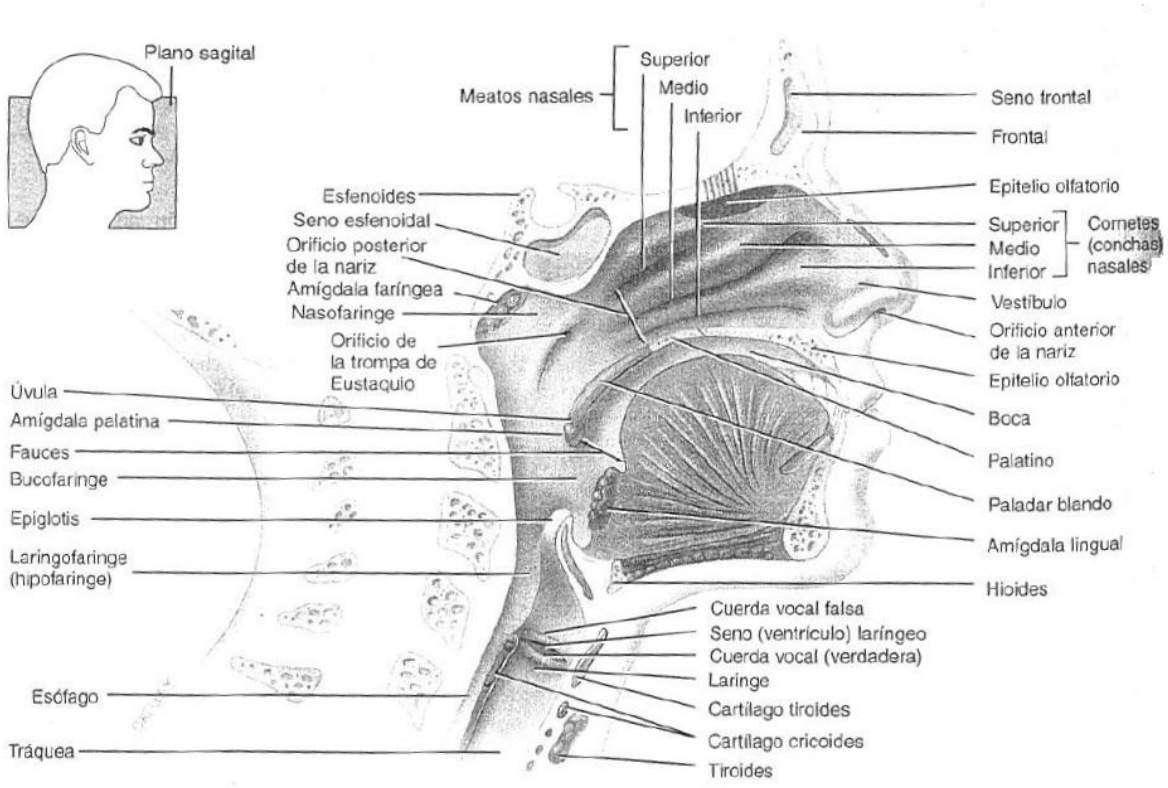


Figura I-2. Corte Sagital del lado izquierdo de la cabeza y cuello que muestra la localización de las estructuras respiratorias [13].

Al contraerse los músculos intrínsecos de la laringe ejercen tracción en los cartílagos aritenoides y los hacen girar. Por ejemplo, la contracción de los músculos cricoaritenoides posteriores separa las cuerdas vocales (abducción) y, con ello, abre la rima glottidis, ver figura I-3. En contraste, la de los músculos cricoaritenoides laterales acerca las cuerdas vocales (aducción) y cierra la rima glottidis, ver figura I-4. Otros músculos intrínsecos pueden alargar (poner en tensión) o acortar (relajar) las cuerdas vocales.

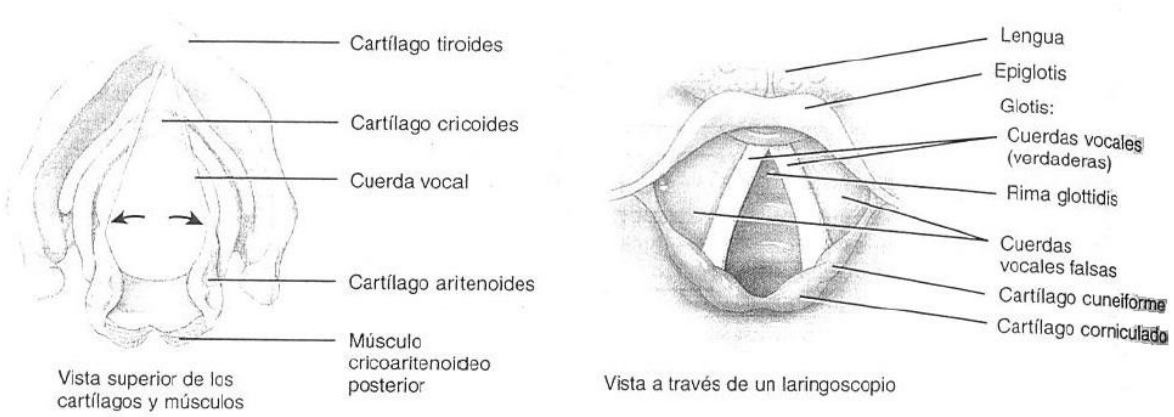


Figura I-3. Movimiento de separación de las cuerdas vocales (abducción) [13].

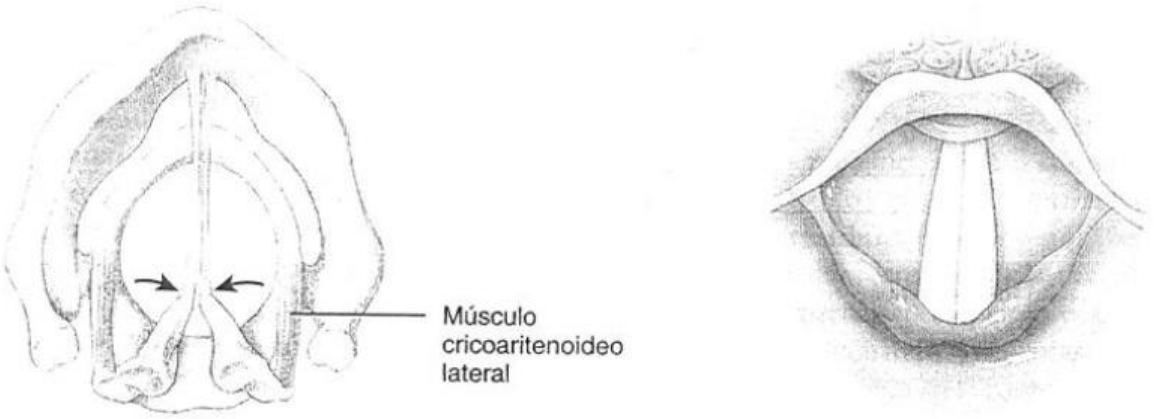


Figura I-4. Movimiento de acercamiento de las cuerdas vocales (aducción) [13].

El tono de la voz se controla con la tensión de las cuerdas vocales. Si los músculos las someten a tensión considerable, vibran más rápidamente y la voz se torna más aguda. Los sonidos graves se producen al disminuir la tensión muscular en dichas cuerdas. En virtud de efectos de los andrógenos (hormonas sexuales masculinas), las cuerdas vocales suelen ser más gruesas y largas en los varones que en las mujeres, por lo que vibran más lentamente. En consecuencia, éstos generalmente tienen voz más grave que las mujeres.

Los sonidos se originan por la vibración de las cuerdas vocales; pero se requieren otras estructuras para convertirlos en habla comprensible. La faringe, boca, nariz y senos paranasales actúan como cámaras de resonancia que confieren a la voz su carácter humano

e individual. Los sonidos vocales se producen al contraer y relajar los músculos de la pared faríngea. Por otra parte, los músculos de la cara, lengua y labios ayudan a la articulación de las palabras.

La voz susurrada se logra al cerrar casi por completo la rima glottidis, salvo su sección posterior, las cuerdas vocales no vibran durante el habla en voz baja, por lo que esta no posee tonalidad. Empero, es factible producir habla inteligible al susurrar mediante cambios en la forma de la boca al articular las palabras. Dichos cambios modifican las cualidades de resonancia de la boca, lo cual confiere, al aire que fluye hacia los labios, tonalidad [13].

1.4. El aparato fonador

A continuación se mencionará de un modo más detallado lo relacionado con el aparato fonador. El aparato fonador se divide para su estudio en tres porciones:

- La ***mancha*** o ***fuelle***. Formada por las estructuras infraglóticas que determinan la mayor o menor presión del aire espirado.
- El ***vibrador***. Constituido por los pliegues vocales (cuerdas vocales) de la laringe.
- Los ***resonadores***. Integrados por las cavidades supraglóticas donde el sonido producido en los pliegues vocales es ampliado y modificado.

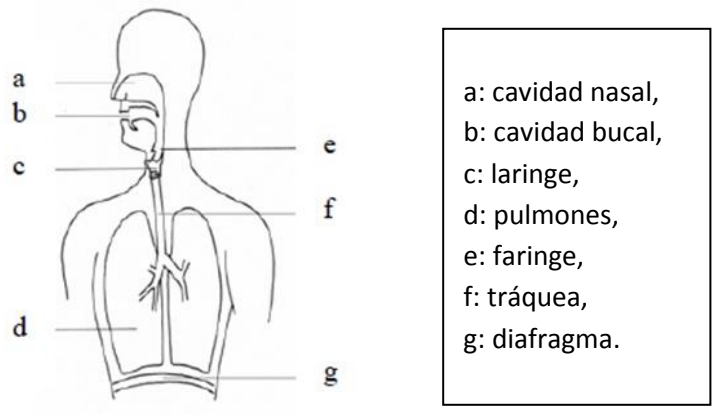


Figura I-5. Aparato Fonador [14].

A pesar de esta división, el aparato fonador es un todo homogéneo e inseparable, por lo cual cualquier alteración o modificación en alguna de sus partes determinará una modificación o alteración en las demás. Cualquier tensión muscular excesiva en cualquiera de ellas provocará problemas en la emisión de la voz y alteraciones a largo o corto plazo en la laringe [14]. En la figura I-5 se indican las partes que integran el aparato fonador.

1.4.1. El fuelle del aparato fonador

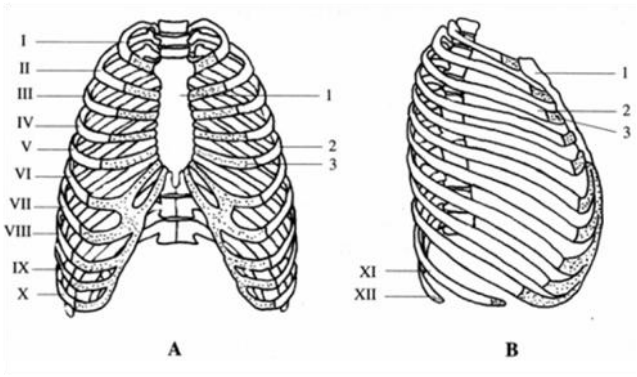
El conjunto formado por los pulmones y la musculatura que suministra la energía necesaria al aire espirado se denomina ***fuelle del aparato fonador*** *o* ***vocal***. Así, como componentes de la mancha o fuelle encontramos a la caja torácica y a los pulmones, el diafragma (musculo inspirador) y los músculos del abdomen (espiradores), así como músculos accesorios de la respiración que actuaran únicamente en casos muy concretos. Clásicamente se describen 3 tipos básicos de respiración: diafragmática, clavicular e intercostal.

La respiración diafragmática es la que se produce en la parte más baja del tórax y en la más alta del abdomen, que es la zona donde radica el mayor control voluntario de la respiración. En esta, el diafragma realiza un movimiento amplio de descenso. Es la óptima para la fonación, principalmente en el canto, ya que no provoca tensiones musculares y deja las estructuras en la posición más adecuada para poder ejercer un control voluntario sobre ellas. Un buen control de la respiración es más importante que un aumento de la capacidad inspiratoria. Una respiración demasiado amplia dificultará la fonación.

La respiración clavicular y la intercostal utilizan músculos del cuello y del tórax que al contraerse crean tensiones que dificultan la fonación y son por ello óptimas en el canto.

Caja torácica

La caja torácica está integrada por la unión de las costillas, el esternón y la porción torácica (dorsal) de la columna vertebral. Esta unión se realiza mediante diversas articulaciones que dotan de movilidad y elasticidad a todo el conjunto, lo que permitirá que, durante la respiración, los diámetros de la caja torácica varíen y los pulmones se llenen y vacíen de aire. En la figura I-6 se indican las diferentes costillas que forman parte de la caja torácica.



A: visión anterior, B: visión lateral,
1: esternón,
2: porción ósea de la costilla,
3: porción cartilaginosa de la costilla,
I-VII: costillas verdaderas,
VIII-X: costillas falsas,
XI-XII: costillas flotantes.

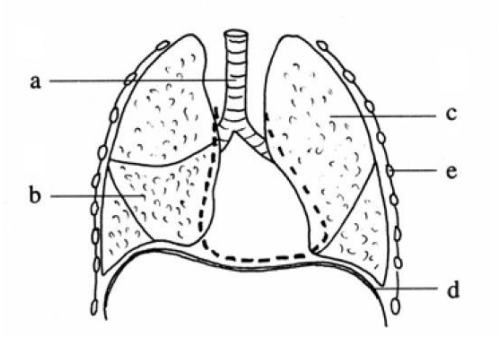
Figura I-6. Caja Torácica [14].

Tráquea y pulmones

La tráquea se sitúa anterior al esófago. Se extiende entre la laringe y los bronquios principales, derecho e izquierdo, donde se bifurca. Su función es la de conducir el aire hacia los pulmones o fuera de ellos, como se observa en la figura I-7.

Los pulmones son los órganos de la respiración, su función básica es la de oxigenar la sangre. Son elásticos, suaves, esponjosos y flotan en el agua. Cada pulmón está envuelto en su pleura. El pulmón derecho está formado por tres lóbulos y el izquierdo por dos, tal como se indica en la figura I-7. La pleura es un saco de doble pared, una interna íntimamente unida al pulmón y una externa adherida a la pared torácica y a la cara craneal del diafragma. Gracias a esta unión de la capa interna y la externa de la pleura, los pulmones seguirán al diafragma y a las costillas en sus movimientos respiratorios.

En la inspiración, la capacidad de la cavidad torácica aumenta en las tres direcciones del espacio. Al ensancharse el pulmón, se produce una reducción de la presión intraalveolar y el aire es inspirado hacia el interior. El aire entra en el pulmón como lo hace un líquido al interior de una jeringa al estirar el émbolo.



a: tráquea, b-c: pulmones,
d: diafragma, e: costillas.

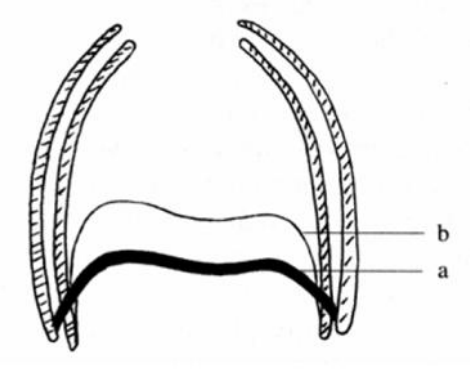
En trazo discontinuo se marca la silueta del corazón que se sitúa anterior a la tráquea.

Figura I-7. Diafragma y estructuras de la cavidad torácica [14].

La espiración normal o tranquila es un proceso pasivo. En la espiración activa, como durante la fonación, intervienen diversos músculos.

Diafragma. Inspiración

El diafragma es el musculo principal de la inspiración. Se sitúa como una lámina que separa la cavidad torácica de la abdominal. Tiene forma de doble cúpula y constituye el suelo de la cavidad torácica y el techo de la abdominal.



a: inspiración,
b: espiración.

Figura I-8. Movimientos del diafragma en la respiración [14].

Cuando el diafragma se contrae, sus cúpulas derecha e izquierda se desplazan hacia abajo aplanándose, como se indica en la figura I-8. Este descenso aumenta el diámetro vertical de la cavidad torácica; por su origen en la abertura inferior de la caja torácica actuará sobre sus

articulaciones determinando el movimiento de las costillas, lo que provoca el aumento de los diámetros anteroposterior y transversal. Con este aumento de los tres ejes del tórax los pulmones se llenan de aire.

La espiración tranquila es un proceso pasivo en el que el ascenso del diafragma se produce por su elasticidad y la de los elementos de la cavidad torácica. Este retorno del diafragma determina la salida de aire intrapulmonar que pasara por la laringe en la cual las cuerdas vocales están abducidas. En el habla o en el canto la respiración es activa y está controlada por la musculatura abdominal. En la fonación las cuerdas vocales están abducidas y el aire espirado sale de los pulmones con una cierta presión para poder abrir la hendidura glótica (glotis) y producir la vibración de los pliegues vocales (cuerdas vocales).

1.4.2. El vibrador del aparato fonador. La laringe.

La laringe tiene la función de proteger las vías respiratorias y de producir los sonidos bajo la acción del aire espiratorio. Se sitúa en la parte medial y anterior del cuello, por delante de la faringe. Cranealmente comunica, a través de la faringe, con la cavidad bucal y las fosas nasales, y caudalmente se continúa con la tráquea. Interviene en la respiración, la deglución y la fonación.

Como en los primates no humanos, la laringe del niño se halla en una posición elevada en el cuello, a la altura de la base occipital o de las primeras vertebras cervicales. Hasta la edad de un año y medio o dos, la posición de la laringe del niño sigue elevada, similar a la de cualquier otro mamífero. Luego, alrededor de los dos años, empieza a descender, lo cual modifica la manera de respirar, de deglutir y emitir sonidos. Debido a este descenso de la laringe se produce una cavidad muy desarrollada por encima de las cuerdas vocales, gracias a la cual los sonidos emitidos por la laringe pueden ser modificados, hacerse audibles y el niño puede empezar a producir los sonidos del habla. A partir de esa edad, el niño no podrá deglutir y respirar a la vez. Posiblemente, la posición baja de la laringe en el ser humano, con la consiguiente expansión del tracto vocal, sean la clave de nuestra capacidad para producir toda la riqueza del lenguaje articulado.

Otro cambio importante sucede durante la pubertad. Por un efecto hormonal la laringe crece en longitud y diámetro, con lo cual las cuerdas vocales crecen en longitud. Durante este proceso se produce la muda vocal. En el niño los pliegues vocales crecen entre 4 y 11 mm. y en la niña entre 1.5 y 4 mm., por ello, el cambio de la voz en el sexo masculino será más evidente. Antes de la pubertad, los pliegues vocales del niño y de la niña tienen una longitud similar mientras que, después de la pubertad, las cuerdas vocales del hombre tienen casi en promedio el doble de longitud que los de la mujer. El cambio que se produce en la voz durante la pubertad no solo se relaciona con el aumento de longitud de las cuerdas vocales sino también con los cambios estructurales que se producen en la estructura histológica de la propia cuerda vocal. Durante la pubertad, la voz femenina desciende alrededor de 2.5 semitonos y la voz masculina alrededor de una octava. La laringe contiene los pliegues vocales denominados comúnmente cuerdas vocales.

Cartílagos y articulaciones de la laringe

La laringe está formada por un esqueleto de piezas cartilaginosas que se articulan entre sí. Los cartílagos de la laringe son nueve: tres de impares (tiroides, cricoides y epiglotis) y tres de pares (aritenoides, corniculados o de Santorini y cuneiformes o de Wrisberg o de Morgagni). Existen, además, pequeños cartílagos inconstantes. En ellos se insertan pequeños músculos (musculatura intrínseca), que actuando sobre sus articulaciones, determinaran los movimientos de los pliegues vocales (cuerdas vocales). En la figura I-9 se indican los cartílagos de la laringe.

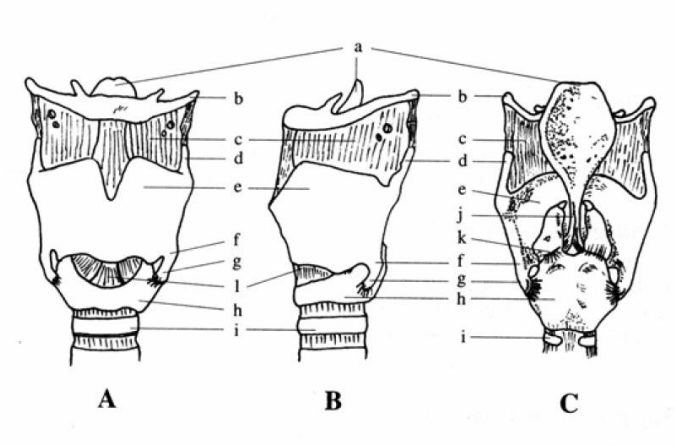
Revisando los cartílagos principales, tenemos:

Tiroides. Constituye la mayor parte de la pared anterior y lateral de la laringe y envuelve parcialmente los demás cartílagos. Está formado por dos láminas que se unen por delante formando la denominada nuez del cuello.

Cricoides. Es el más inferior de la laringe. Tiene forma de anillo de sello con una lámina cuadrilátera posterior dispuesta entre las láminas del tiroides y un arco estrecho en posición anterior que se palpa fácilmente en el cuello.

Aritenoides. Son pares simétricos respecto a la línea media. Tienen forma de pirámide triangular y en ellos se insertan los pliegues vocales (cuerdas vocales).

Epiglotis. Tiene forma de hoja con su peciolo en posición inferior.



A: visión anterior, B: visión lateral, C: visión posterior. a: epiglotis, b: hioides, c:membrana tiroidea, d: asta superior del tiroides, e: tiroides, f: asta inferior del tiroides, g: articulación cricotiroidea, h: cricoides, i: primer cartílago de la tráquea, j: ligamento tiroepiglótico. K: articulación cricoaritenoidea, ligamento cricotoroideo.

Figura I-9. Cartílagos de la laringe [14].

Musculatura intrínseca de la laringe y su inervación

En la laringe podemos distinguir una musculatura intrínseca, que determina los movimientos de las articulaciones laríngeas:

Cricotiroideo. Alarga, tensa y aduce los pliegues vocales.

Cricoaritenoideo posterior. Es el único músculo abductor de las cuerdas vocales.

Cricoaritenoideo lateral. Aductor de los pliegues vocales.

Vocal. Constituye la mayor parte del pliegue vocal. Es el responsable de sus variaciones locales de tensión durante la fonación.

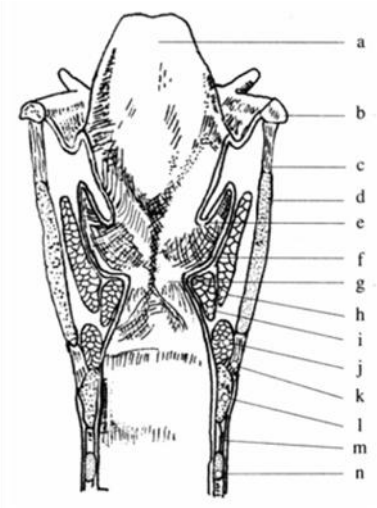
Tiroaritenoideo. Algunas de sus fibras se extienden hasta la epiglotis formando el músculo tiroepiglótico. Es aductor de los pliegues vocales.

Aritenoideo transverso. Aductor de los pliegues vocales.

Aritenoideo oblicuo. Algunas de sus fibras se reflejan hacia la epiglotis constituyendo el músculo aritenoepiglótico. Es aductor de los pliegues vocales.

Constitución de los pliegues vocales y vestibulares

A cada lado de la superficie interna de la laringe encontramos dos pliegues de mucosa superpuestos: los pliegues vestibulares (pliegues ventriculares, cuerdas vocales falsas, cuerdas vocales superiores o bandas ventriculares) situados cranealmente, como se indica en la figura I-10 y los pliegues vocales (cuerdas vocales verdaderas o cuerdas vocales inferiores) en posición caudal, como también se indica en la figura I-10. El pliegue vestibular recubre el ligamento vestibular y se forma a causa de su presencia. El pliegue vocal recubre el ligamento vocal y el músculo vocal, y viene determinado por la existencia de estas estructuras que forman su esqueleto.



a: cartílago epiglotis, b: hueso hioides, c: membrana tirohioidea, d: cartílago tiroides, e: pliegues vestibulares (cuerdas vocales falsas; no tienen ninguna acción fonatoria), f: ventrículo laríngeo, g: pliegue vocal (cuerda vocal), h: musculo tiroaritenoideo, i: musculo vocal, j: musculo cricoaritenoideo lateral, k: ligamento cricoaritenoideo, l: cartílago cricoides, m: ligamento cricotraqueal, n: primer anillo traqueal

Figura I-10. Corte de la laringe que muestra los pliegues vocales derecho e izquierdo [14].

Los pliegues vocales son altamente elásticos y tienen una estructura histológica que le permite a la voz su gran versatilidad. Están conformados por cinco capas de estructura histológica distinta en el pliegue vocal. La más superficial de ellas es el epitelio delgado y lubrificado que cubre la lámina propia, en la que se distinguen tres capas; bajo esta, en la posición más profunda, se encuentra el músculo vocal.

En los pliegues vocales, al paso del aire espirado, se produce un tono complejo que es modificado y amplificado en las cavidades de resonancia supraglóticas. Sin ellas, el sonido producido no sería audible. Las diferentes propiedades mecánicas de las cinco capas son esenciales para los suaves movimientos de los pliegues vocales y su vibración normal.

Durante toda la vida, la laringe se verá afectada por los distintos cambios producidos en los niveles de hormonas sexuales. Durante la pubertad, como ya se ha mencionado, se producen cambios importantes en la laringe. Los pliegues vocales crecen en longitud y se producen también cambios en la masa del pliegue vocal determinando ambos procesos los cambios de la voz. Pero los pliegues vocales también se verán afectados en su constitución bioquímica durante otros procesos que impliquen cambios hormonales importantes, como son la menopausia, el periodo premenstrual o la vejez.

Se ha observado que la frecuencia fundamental de la voz (en estudios hechos con voz cantada) decrece sucesivamente aproximadamente cada diez años en ambos sexos. Para las mujeres, la frecuencia fundamental continúa decreciendo en edades avanzadas y parece crecer más rápidamente después de la menopausia. En los hombres, por el contrario, la frecuencia fundamental de la voz aumenta repentinamente hacia los 60 años y continúa creciendo un poco durante el resto de la vida. De manera que, mientras a los 20 años hay grandes diferencias entre las frecuencias fundamentales en ambos sexos, a los 90 hay muy poca diferencia entre el tono fundamental de la voz masculina y femenina.

Fonación. Tono, timbre e intensidad de la voz.

La fonación exige un cierre y una abertura continuas de los pliegues vocales (cuerdas vocales) con cambios en la longitud y la tensión. Estas variaciones requerirán fluctuaciones continuas de la salida de aire. En el habla normal, la regulación de la salida de este aire es básicamente voluntaria y automática. En los conferenciantes, actores y cantantes se observa, sin embargo, un control en la espiración del aire a través de la hendidura glótica (glotis) cerrada; los pliegues vocales son obligados a separarse y a ponerse en vibración, debido a la presión ejercida por el aire espirado (presión subglótica). El sonido producido en los pliegues vocales sería prácticamente inaudible si este no se modificara y amplificara en las cavidades supraglóticas o resonadores de la voz.

El sonido producido en los pliegues es un tono complejo, que consta de una frecuencia fundamental y de sus armónicos superiores. El tono aumenta cuando los ciclos de cierre y abertura de los pliegues vocales se acortan y se repiten con más frecuencia. La onda

compuesta formada en la laringe pasa a través de las cavidades supraglóticas que actúan como filtros, dejando pasar solo aquellas frecuencias que coinciden con la de las propias cavidades de resonancia. El conjunto formado por el tono fundamental más los armónicos modificados constituyen el timbre de la voz.

El tono de la voz está directamente relacionado con la longitud y el grosor de los pliegues vocales de cada persona. Las diferencias relativas entre hombres y mujeres en cuanto a longitud (aproximadamente 18 mm en los hombres y 10 mm en las mujeres) y el grosor de los pliegues vocales son los determinantes primarios de la diferencia de tono entre personas de ambos sexos (la frecuencia fundamental en el hombre es de unos 125 Hz y en la mujer de unos 200 Hz).

La intensidad o volumen de la voz dependerá principalmente de la presión del aire espirado. La energía con la que el aire es impulsado desde los pulmones determinará una mayor o menor amplitud vibratoria de los pliegues vocales, que provocará un aumento o disminución de la intensidad del sonido producido. Al aumentar la presión del aire espirado crece la amplitud de las vibraciones, ya que los pliegues vocales se distancian y acercan con mayor agilidad.

1.4.3. Los resonadores del aparato fonador

Todas las cavidades situadas por encima de los pliegues vocales (cuerdas vocales) actúan, o pueden actuar, como cajas de resonancia de la voz. Se habla de resonadores o cavidades supraglóticas. Distinguimos la boca, la faringe y las fosas nasales.

Hay resonadores móviles, como la boca, que pueden modificar su forma y volumen adaptándose al sonido producido, y otros fijos, como las fosas nasales, que no podrán cambiar su forma ni su volumen. La boca se modificará en función de la abertura mandibular y de la posición de la lengua y labios. La faringe cambia su morfología principalmente en función de los desplazamientos de la laringe, la lengua y el velo paladar o paladar blando.

De manera general, especialmente en la voz cantada, se ha dado gran importancia a los senos paranasales como resonadores de la voz, pero estas cavidades actúan como zonas en las cuales el aire vibra dando lugar a sensaciones propioceptivas para el cantante, y no como cavidades de resonancia para amplificar el sonido y hacerlo más audible.

1.5. Clasificación de los sonidos de la voz

Los sonidos emitidos por el aparato fonatorio pueden clasificarse de acuerdo con diversos criterios que tienen en cuenta los diferentes aspectos del fenómeno de emisión. Estos criterios son:

- Según su carácter vocálico o consonántico.
- Según su oralidad o nasalidad.
- Según su carácter tonal (sonoro) o no tonal (sordo).
- Según el lugar de articulación.
- Según el modo de articulación.
- Según la posición de los órganos articulatorios.
- Según la duración.

1.5.1. Vocales y consonantes

Desde un punto de vista mecánico-acústico, las ***vocales*** son los sonidos emitidos por la sola vibración de las cuerdas vocales sin ningún obstáculo o constricción entre la laringe y las aberturas oral y nasal. Dicha vibración se genera por el principio del oscilador de relajación, donde interviene una fuente de energía constante en la forma de un flujo de aire proveniente de los pulmones. Son siempre sonidos de carácter tonal (cuasi-periódicos), y por consiguiente de espectro discreto. Las ***consonantes***, por el contrario, se emiten interponiendo algún obstáculo formado por los elementos articulatorios. Los sonidos correspondientes a las consonantes pueden ser tonales o no, dependiendo de si las cuerdas vocales están vibrando o no. Funcionalmente, en el español las vocales pueden constituir palabras completas, no así las consonantes.

1.5.2. Oralidad y nasalidad

Los fonemas en los que el aire pasa por la cavidad nasal se denominan ***nasales***, en tanto que aquellos en los que sale por la boca se denominan ***orales***. La diferencia principal está en el tipo de resonador principal por encima de la laringe (cavidad nasal y oral, respectivamente). En el español son nasales solo las consonantes "m", "n", "ñ".

1.5.3. Tonalidad

Los fonemas en los que participa la vibración de las cuerdas vocales se denominan ***tonales*** o, también, ***sonoros***. La tonalidad lleva implícito un espectro cuasi-periódico. Como se mencionó anteriormente, todas las vocales son tonales, pero existen varias consonantes que también lo son: "b", "d", "m", etc. Aquellos fonemas producidos sin vibraciones glotales se denominan ***sordos***. Varios de ellos son el resultado de la turbulencia causada por el aire pasando a gran velocidad por un espacio reducido, como las consonantes "s", "z", "j", "f".

1.5.4. Lugar y modo de articulación (consonantes)

La ***articulación*** es el proceso mediante el cual alguna parte del aparato fonatorio interpone un obstáculo para la circulación del flujo de aire. Las características de la articulación permitirán clasificar las consonantes. Los órganos articulatorios son los labios, los dientes, las diferentes partes del paladar (alvéolo, paladar duro, paladar blando o velo), la lengua y la glotis. Salvo la glotis, que puede articular por sí misma, el resto de los órganos articula por oposición con otro. Según el lugar o punto de articulación se tienen fonemas:

Bilabiales: oposición de ambos labios.

Labiodentales: oposición de los dientes superiores con el labio inferior.

Linguodentales: oposición de la punta de la lengua con los dientes superiores.

Alveolares: oposición de la punta de la lengua con la región alveolar.

Palatales: oposición de la lengua con el paladar duro.

Velares: oposición de la parte posterior de la lengua con el paladar blando.

Glotales: articulación en la propia glotis.

A su vez, para cada punto de articulación, esta puede efectuarse de diferentes modos; dando lugar a fonemas:

Oclusivos: la salida de aire se cierra momentáneamente por completo.

Fricativos: el aire sale atravesando un espacio estrecho.

Africados: oclusión salida por fricación.

Laterales: la lengua obstruye el centro de la boca y el aire sale por los lados.

Vibrantes: la lengua vibra cerrando el paso del aire intermitentemente.

Aproximantes: la obstrucción muy estrecha que no llega a producir turbulencia.

Los fonemas oclusivos (correspondientes a las consonantes "b" inicial o postnasal, "c", "k", "d", "g" inicial, postnasal o postlateral, "p", "t") también se denominan a veces explosivos, debido a la liberación repentina de la presión presente inmediatamente antes de su emisión. Pueden ser sordos o sonoros, al igual que los fricativos ("b" postvocálica, postlateral y postvibrante, "g" postvocálica y post vibrante, "f", "j", "h" aspirada, "s", "y", "z"). Sólo existe un fonema africado en el español, correspondiente a la "ch". Los laterales ("l", "ll") a veces se denominan líquidos, y son siempre sonoros. Los dos fonemas vibrantes en el español ("r", "rr") difieren en que en uno de ellos ("r") se ejecuta una sola vibración y es intervocálico, mientras que en el otro ("rr") es una sucesión de dos o tres vibraciones de la lengua. Finalmente, los fonemas aproximantes (la "i", y la "u" cerradas que aparecen en algunos diptongos) son a veces denominadas semivocales, pues en realidad suenan como vocales. Pero exhiben una diferencia muy importante: son de corta duración y no son prolongadas.

En la tabla I-2 se indican las consonantes clasificadas según el lugar y el modo de articulación, la sonoridad y la oro-nasalidad. En algunos casos una misma consonante aparece en dos categorías diferentes, de acuerdo con las diferencias antes notadas.

Tabla I-2. Clasificación de las consonantes de la lengua castellana según el lugar y el modo de articulación y la sonoridad [12].

Lugar de articulación	Modo de articulación								
	Oral								Nasal
	Oclusiva		Fricativa		Africada	Lateral	Vibrante	Aproximante	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora	Sonora	Sonora	Sonora
Bilabial	p	b, v		b, v				w	m
Labiodental			f						
Linguodental			z	d					
Alveolar	t	d	s	y	ch	l	r, rr		n
Palatal				(y)	(ch)	ll		i	ñ
Velar	k	g	j	g					
Glotal			h						

1.5.5. Posición de los órganos articulatorios (vocales)

En el caso de las vocales, la articulación consiste en la modificación de la acción filtrante de los diversos resonadores, lo cual depende de las posiciones de la lengua (tanto en elevación como en profundidad o avance), de la mandíbula inferior, de los labios [12] y del paladar blando. Estos órganos influyen sobre los formantes, permitiendo su control. Podemos clasificar las vocales, según la posición de la lengua, como se muestra en la tabla I-3.

Tabla I-3. Clasificación de las vocales castellanas según la posición de la lengua [12].

Posición vertical	Tipo de vocal	Posición horizontal (avance)		
		Anterior	Central	Posterior
Alta	Cerrada	i		u
Media	Media	e		o
Baja	Abierta		a	

Otra cualidad controlable es la ***labialización***, es decir el hecho de que se haga participar activamente los labios. Las vocales labializadas, también definidas como ***redondeadas***, son las que redondean los labios hacia adelante, incrementando la longitud efectiva del tracto vocal. La única vocal labializada en el español es la "u".

En otros idiomas, como el francés, el portugués, el catalán y el polaco, así como en lenguas no europeas como el guaraní o el hindi, existe también el matiz de oralidad o nasalidad [12]. En las vocales orales el velo (paladar blando) sube, obturando la nasofaringe, lo cual impide que el aire fluya parcialmente por la cavidad nasal. En las vocales nasalizadas (u oronasales) el velo baja, liberando el paso del aire a través de la nasofaringe. Se incorpora así la resonancia nasal.

1.5.6. Duración

La duración de los sonidos, especialmente de las vocales, no tiene importancia a nivel semántico en el español, pero sí en el plano expresivo, a través de la agogia, es decir el énfasis o acentuación a través de la duración. En inglés, en cambio, la duración de una vocal puede cambiar completamente el significado de la palabra que contiene.

1.6. El alfabeto fonético internacional

El español es un idioma cuya escritura es eminentemente fonética, ya que salvo pocos casos, hay correspondencia entre grafema y fonema. No todos los idiomas tienen esta característica. El inglés es un caso quizás extremo a tal grado que se han creado posibles ortografías alternativas para algunas palabras basándose en la forma en que sus fonemas aparecen escritos en otras palabras [12]. Estas extrañas ortografías y el análisis correspondiente se muestran en la tabla I-4.

Tabla I-4. Ortografías alternativas de George Bernard Shaw para dos palabras inglesas [12].

Palabra	Ortografía alternativa (según Shaw)	Fonema	Palabra en la que se usa la ortografía alternativa	Escritura en el Alfabeto Fonético Internacional
fish	ghoti	GH	enou**gh**	[ɪˈn∧f]
		O	w**o**men	[ˈwɪmɪn]
		TI	na**ti**on	[ˈneɪʃ ən]
potato	ghoughpteighbteau	GH	hiccou**gh**	[ˈhɪc∧p]
		OUGH	th**ough**	[ðəʊ]
		PT	**pt**eranodon	[traˈnədon]
		EIGH	n**eigh**bour	[ˈneɪbə]
		BT	de**bt**	[det]
		EAU	bur**eau**	[bjʊəˈr əʊ]

Se ha compilado un extenso conjunto de símbolos fonéticos conocido como el *Alfabeto Fonético Internacional* ("International Phonetic Alphabet, IPA") que contiene una gran cantidad de fonemas de los diversos idiomas, y que permite representar de una manera inequívoca los fonemas independientemente del idioma. El subconjunto correspondiente al español se observa en la tabla I-5.

Tabla I-5. Los fonemas del alfabeto fonético internacional utilizados en la lengua castellana [12] [15].

Fonemas castellanos					
Sonido	Ejemplo	Sonido	Ejemplo	Sonido	Ejemplo
[p]	**p**aso	[θ]	**z**orz**z**al, lápi**z**	[ɲ]	ma**ñ**ana, **ñ**o**ñ**o
[b]	**b**ase, **v**ena	[s]	**s**olo, co**s**a	[dʒ]	**y**o, **Y**apeyú
[ß]	la**b**or, la**v**ar	[x]	**g**iro, **j**arabe	[j]	**bi**en, **bi**ólogo
[t]	**t**res, can**t**o	[tʃ]	he**ch**o, **Ch**ubut	[w]	h**u**eso, b**u**itre
[d]	**d**ama, an**d**ar	[r]	ar**r**e**r**, ja**r**abe		
[ð]	ce**d**ro, verda**d**	[rr]	pe**rr**o, **r**ojo	[a]	**ca**ma
[k]	**c**aso, dis**c**o	[l]	**l**oa**bl**e, fie**l**	[e]	**e**spera, v**e**r
[g]	**g**ula, **g**oma	[λ]	**ll**anto, ca**ll**e	[i]	v**i**ne, **i**r**i**s
[γ]	a**g**ua, ne**g**ro	[m]	**m**a**m**á, á**m**bar	[o]	**lo**ro, p**o**s
[f]	**f**ino, ti**f**ón	[n]	**n**e**n**e, jove**n**	[u]	b**u**rla, h**u**racán

II. El sistema de reconocimiento de voz Sphinx

2.1. Consideraciones generales del sistema Sphinx

En las últimas décadas se ha presentado un progreso considerable en el reconocimiento de voz. Han surgido varios sistemas, cada uno de ellos con cierta eficacia, ya que su correcto funcionamiento depende de ciertas consideraciones que los mismos fijan. El sistema Sphinx es un sistema que trata de sobreponerse a tres de dichas limitantes: 1) dependencia del hablante, 2) palabras aisladas y 3) vocabularios pequeños.

La independencia del hablante ha sido vista como la más difícil de las tareas a sobreponerse. Esto porque la mayoría de las representaciones paramétricas de la voz son altamente dependientes del hablante y un conjunto de patrones de referencia adecuados para un hablante puede presentar un pobre desempeño para otro hablante.

El reconocimiento de voz continua es significativamente más complejo que el reconocimiento de palabras aisladas. Su complejidad se da como resultado de tres características propias de la voz continua. Primero, los límites de las palabras son difíciles de determinar. Segundo, los efectos coarticulatorios son más fuertes en la voz continua, ocasionando que se pueda presentar el mismo sonido en varios contextos. Tercero, las palabras de contenido (nombres, verbos, adjetivos, etc.) a menudo se enfatizan, mientras que las palabras función (artículos, preposiciones, pronombres, verbos cortos, etc.) se articulan poco. Las tasas de error se incrementan drásticamente de las palabras aisladas a la voz continua.

Un vocabulario grande típicamente implica un vocabulario de alrededor de 1000 palabras o más. A medida que el tamaño del vocabulario se incrementa, sucede lo mismo con el número de palabras confusas. También, vocabularios más grandes requieren de utilizar modelos de subpalabras. Desafortunadamente, las unidades de subpalabra conllevan a degradar el desempeño debido a que no son capaces de capturar los efectos coarticulatorios tan bien como lo pueden hacer los modelos de palabra. A pesar de estos problemas, los

sistemas de vocabularios grandes se requieren para muchas aplicaciones versátiles, tal como, dictado, sistemas de dialogo y sistemas de traducción de voz [16].

El sistema Sphinx es un sistema de reconocimiento de voz continua, independiente del hablante, para vocabularios grandes. Emplea Modelos Ocultos Discretos y Semidiscretos de Markov ("Hidden Markov Models, HMM's") con parámetros derivados de los coeficientes de predicción lineal ("Linear Predictive Coding, LPC").

2.1.1. Representación de la voz

La voz es muestreada a 16 kHz y pre-enfatizada con un filtro de $1 - 0.97z^{-1}$. Después, se le aplica cada 10 mseg una ventana de hamming con un ancho de 20 mseg. Se aplica un análisis de autocorrelación de orden 14 el cual es seguido por un análisis "LPC" del mismo orden. Finalmente, se calculan 12 coeficientes "LPC" cepstral a partir de los coeficientes "LPC", y esos coeficientes "LPC" cepstral se transforman a una escala mel empleando una transformada bilineal. Estos 12 coeficientes son cuantizados vectorialmente en un "codebook" de 256 vectores prototipo.

2.1.2. Entrenamiento de los HMM independientes del contexto

Sphinx se basa en modelos fonéticos ocultos de Markov. Se identifica un conjunto de 48 fonos, y se entrena un modelo oculto de Markov por cada fono. Cada "HMM" fonético contiene tres distribuciones de salida discretas de símbolos de "VQ". Cada distribución es la densidad conjunta de tres "codebooks" de funciones de densidad de probabilidad (pdf's), las cuales se asumen como independientes.

Se inicializa el sistema con la base de datos a emplear (corpus). En esta inicialización se emplea el algoritmo "forward-backward" para entrenar los parámetros de los 48 "HMM's" fonéticos. Por cada frase, se construyen "HMM's" por palabra concatenando "HMM's" de fonos. Estos "HMM's" de palabras son concatenados en una frase larga de "HMM" y entrenados en la correspondiente voz. Debido a que la estimación inicial es buena, solo se ejecutan dos iteraciones del algoritmo "forward-backward". Esta fase de entrenamiento produce 48 modelos acústicos independientes del contexto.

En la figura II-1 podemos observar un diagrama de bloques que muestra el procedimiento que sigue el sistema Sphinx durante el entrenamiento [17].

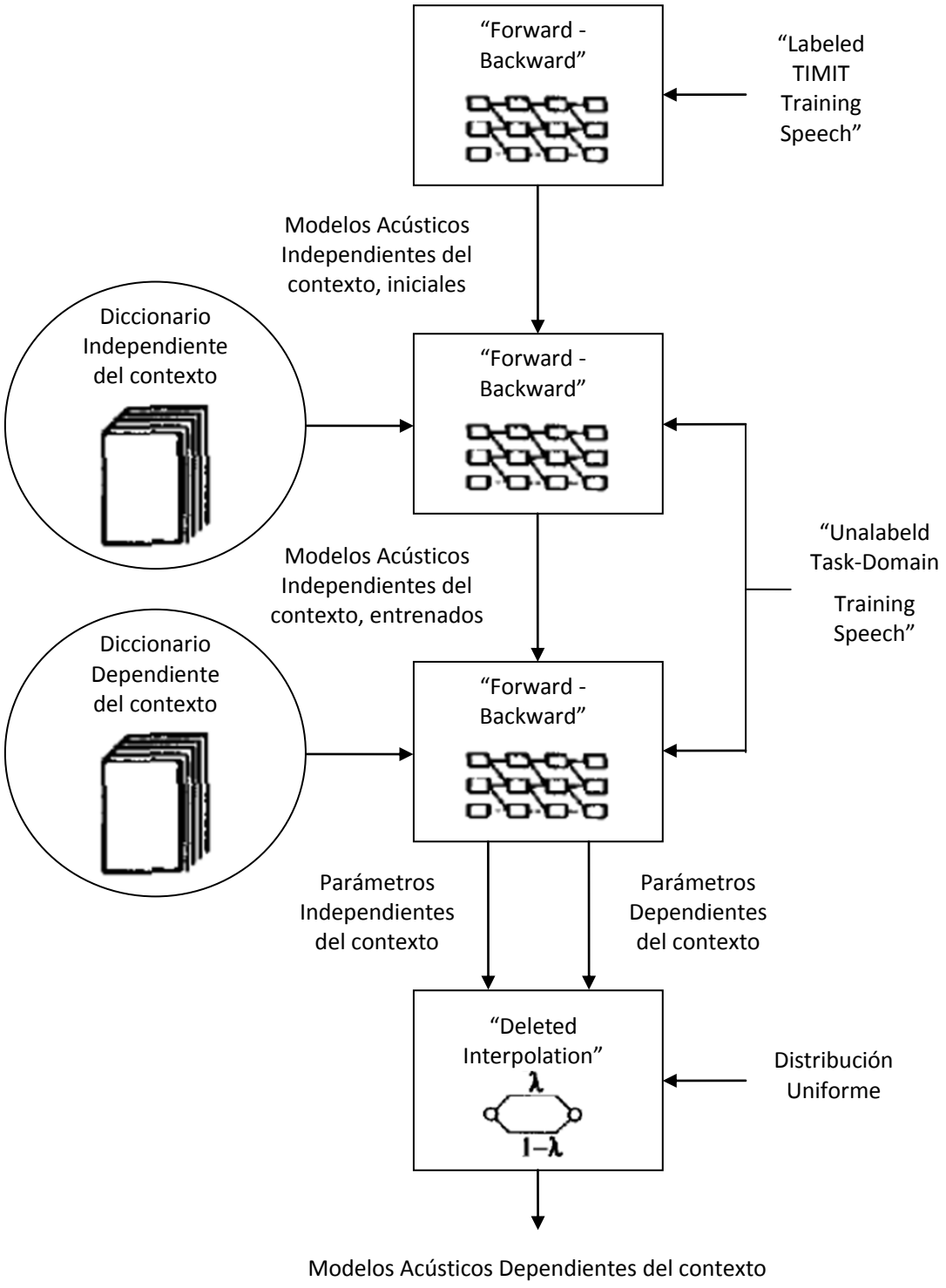


Figura II-1. Procedimiento de entrenamiento en Sphinx [17].

2.1.3. Modelos dependientes palabra función y frase función

Un problema con el reconocimiento de voz continua es la no tan clara articulación de palabras función. En el idioma inglés, dado que el conjunto de palabras función es limitado y las palabras función ocurren de manera frecuente, se modela cada fono en cada palabra función por separado.

Se ha notado que las palabras función son difíciles de reconocer cuando ocurren en grupo. Las palabras están menos claramente articuladas y tienen fuertes efectos coarticulatorios entre sí. Debido a esto se crea un conjunto de modelos acústicos específicos a frases función, las cuales son frases que están compuestas únicamente por palabras función.

2.1.4. Modelos tri-fono generalizados.

Los modelos acústicos dependientes, palabra función y frase función, permiten una mejor representación de las palabras función. De cualquier forma, los modelos acústicos simples para las palabras no función son inadecuados, debido a que la realización de un fono depende de manera crucial del contexto. Para modelar de una mejor manera se propuso el modelo de trifonemas. Se emplea un modelo de trifonema para cada uno de los contextos, izquierdo y derecho. La desventaja radica en que mientras que los modelos de trifonema son sensibles a los contextos fonéticos vecinos y han llevado a buenos resultados, existe un número muy grande de ellos, cada uno de los cuales deber ser entrenado de manera separada.

El algoritmo de generalización de contextos permite un justo medio para encontrar el equilibrio entre la capacidad de entrenamiento y la sensibilidad. Dada una cantidad fija de datos de entrenamiento, es posible encontrar el número más grande de modelos detallados entrenables.

2.1.5. Modelado de coarticulación entre palabras

Trifonemas y modelos trifono generalizados son técnicas de modelado de subpalabra poderosas porque toman en cuenta los contextos fonéticos izquierdo y derecho, los cuales

son las principales causas de variabilidad fonética. Sin embargo, estos modelos consideran solamente el contexto dentro de la palabra. Donde se complica es al considerar los modelos de trifonos para modelar las coarticulaciones entre palabras, debido a que el número de trifonos crece significativamente. Es por ello que, los trifonos generalizados son particularmente recomendables para modelar las coarticulaciones entre palabras.

2.1.6. Reconocimiento de los "HMM" con la duración de la palabra

Para el reconocimiento, se emplea una búsqueda Viterbi que encuentra la secuencia óptima en una red de "HMM" grande. En el nivel más alto, esta "HMM" es una red de palabra "HMM's", ordenada de acuerdo a la gramática. Cada palabra se relaciona con su red de pronunciación fonética, y cada fono se relaciona con su modelo acústico correspondiente. En la figura II-2 se muestra un esquema del decodificador del sistema Sphinx.

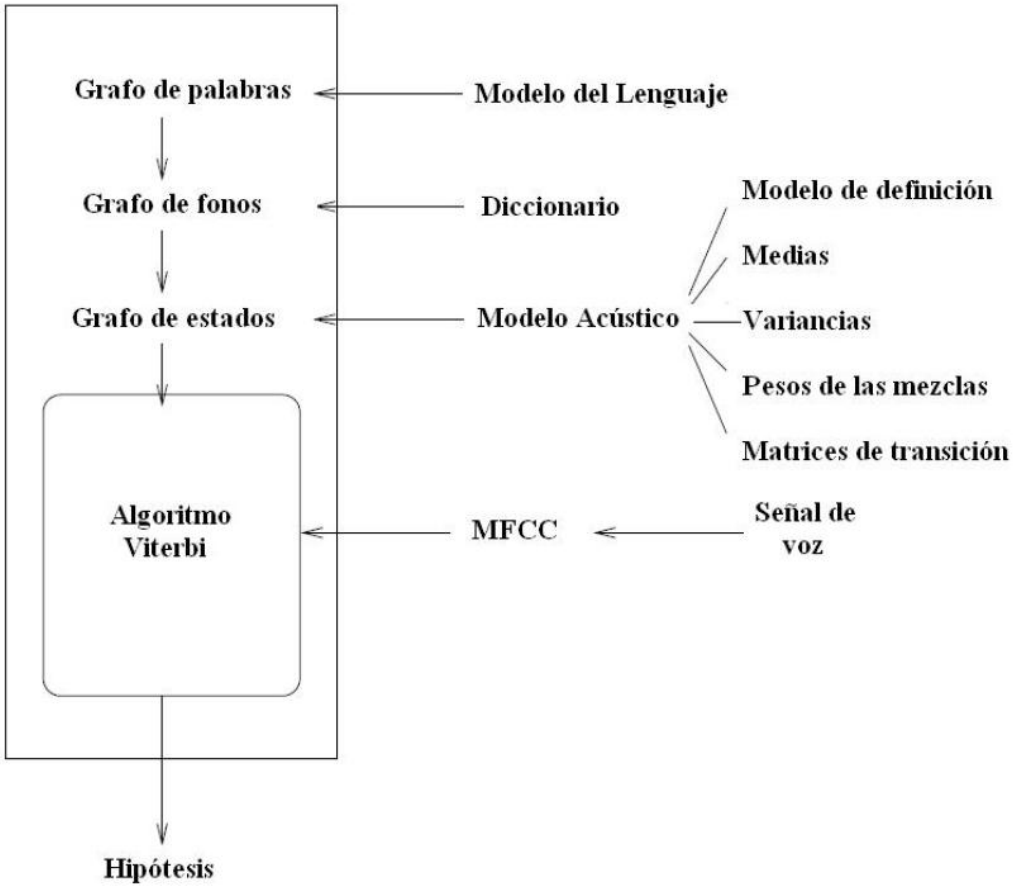


Figura II-2. Esquema del decodificador de Sphinx [2].

2.2. "Front-end" de Sphinx

2.2.1. Descripción general del "front-end"

Los siguientes pasos describen el "front-end" del sistema de reconocimiento Sphinx 3. La tarea del "front-end" consiste en transformar una señal de voz en un conjunto de parámetros que se utilizan en el reconocimiento, específicamente, se trata de coeficientes "Mel-Frecuency Cepstral, MFCC" [18].

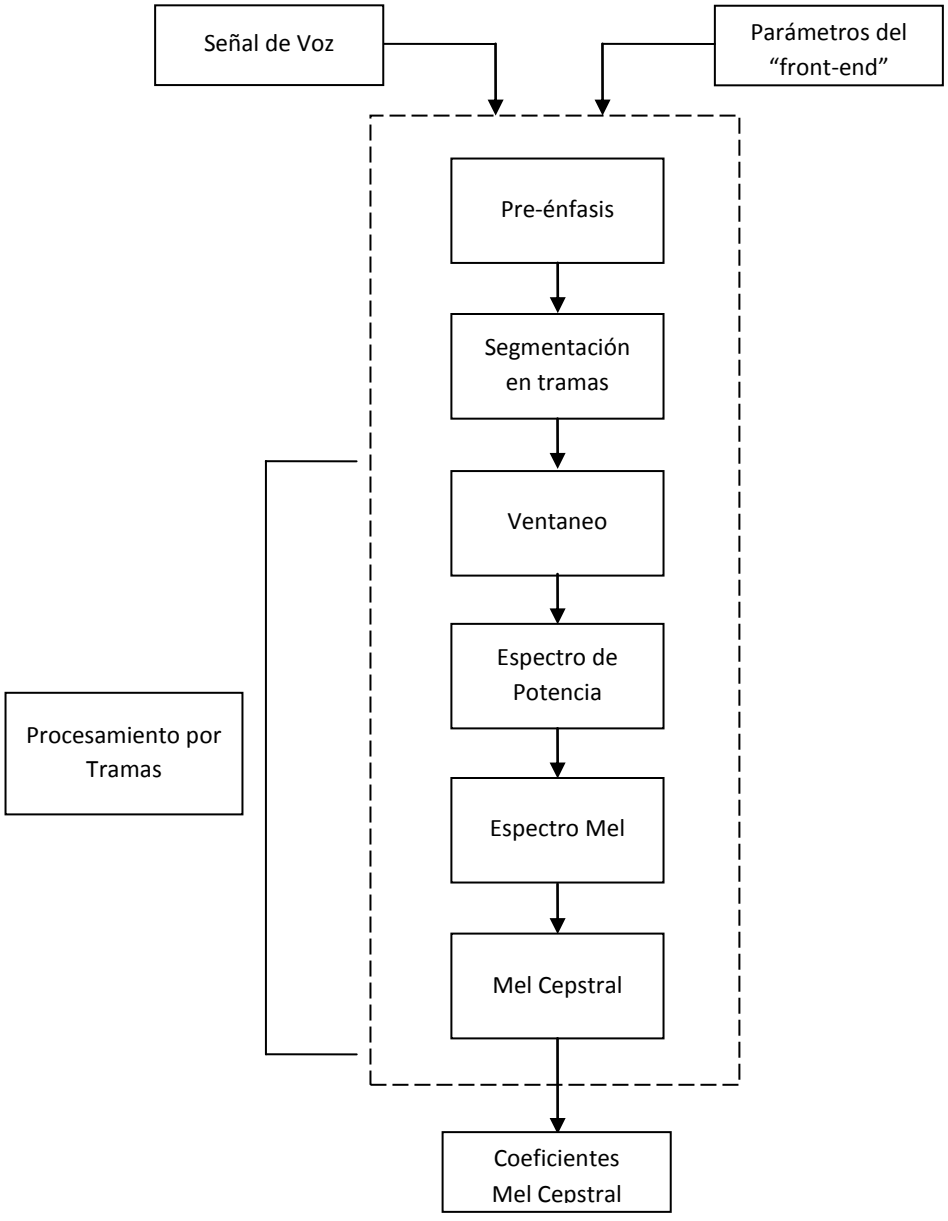


Figura II-3. "Front-end" de Sphinx 3 [18].

En la figura II-3 podemos ver un diagrama de bloques en el que se observan las operaciones que realiza el sistema Sphinx 3 para extraer las características de los archivos de voz.

2.1.2. Procesamiento del "front-end"

Pre-énfasis

Se le aplica el siguiente filtro FIR de pre-énfasis a la señal de entrada:

$$y[n] = x[n] - \alpha x[n-1]$$

α es proporcionada por el usuario o se puede optar por el valor por defecto. Si $\alpha = 0$, entonces se hace caso omiso de este paso y se continúa con el siguiente.

Las siguientes operaciones se realizan por trama.

Ventaneo

La trama es multiplicada por la siguiente ventana de Hamming:

$$w[n] = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right)$$

Donde N es la longitud de la trama

Espectro en potencia

El espectro en potencia de la trama se calcula por medio de una DFT de una cierta longitud, y después calculando su magnitud al cuadrado.

$$s[k] = \big(real(X[k])\big)^2 + \big(imag(X[k])\big)^2$$

Espectro MEL

El espectro MEL de la trama se calcula multiplicando el espectro en potencia por cada uno de los filtros triangulares ponderados Mel e integrando el resultado.

$$\tilde{s}[l] = \sum_{k=0}^{N/2} S[k]M_l[k] \qquad l = 0,1,\dots,L-1$$

Donde N es la longitud de la DFT, y L es el número total de filtros triangulares ponderados.

MEL Cepstral

Se le aplica una DCT al logaritmo natural del espectro Mel para obtener los Mel Cepstral.

$$c[n] = \sum_{i=0}^{L-1} ln(\tilde{s}[i])\, cos\left(\frac{\pi n}{2L}(2i+1)\right) \qquad c = 0,1,\dots,C-1$$

Donde C es el número de coeficientes cepstral.

Valores por defecto del "front-end"

Dentro de la tabla II-1 podemos ver los valores por defecto que utiliza el "front-end" del sistema Sphinx 3 para su funcionamiento.

Tabla II-1. Valores predeterminados del "front-end" de Sphinx 3 [18].

Parámetro	Valor por defecto
Tasa de muestreo	16000 [Hz]
Tasa por Trama	100 tramas/seg
Tamaño de la ventana	0.025625 seg
Tipo de Baco de filtros	Banco de filtros Mel
Numero de Cepstral	13
Numero de filtros Mel	40
Tamaño de la DFT	512
Frecuencia mínima de filtrado	133.33334 [Hz]
Frecuencia máximo de filtrado	6855.4976 [Hz]
Valor de α para el pre-énfasis	0

2.3. Consideraciones y guía para la puesta en marcha de Sphinx

Con la finalidad de que desarrollos posteriores ahorren tiempo en la puesta en marcha del sistema Sphinx 3 y se aboquen a otros tópicos se presentan una serie de consideraciones a manera de guía. Esto resulta particularmente importante ya que no existía. Los autores, en "CMU", no lo tienen y ya no trabajan en el sistema:

- K.F.Lee realizó un manual, pero es para la primera versión de Sphinx, de manera que actualmente es insuficiente, ya que el sistema ha cambiado y muchas de las instrucciones ya no son validas.

- E. Gouvea lo realizó para Sphinx 3, pero centrándose en java, de manera que no nos sirve ya que el desarrollo que hicimos es para windows.

- R. Singh, R.Stern y M.Seltzer. Realizaron para Sphinx 2 varios reportes. Sin embargo no están integrados, o solamente forman parte de algún artículo.

Para hacer funcionar el sistema requeriremos del Sphinx train y del Sphinx decoder, que son aplicaciones que posibilitan dicha tarea.

2.3.1. Componentes para realizar el entrenamiento

El entrenador del Sphinx consta de un conjunto de programas, cada uno para una tarea específica, y un conjunto de scripts que organizan el orden en que los programas se ejecutan. Esto se puede realizar en una gran variedad de compiladores.
El entrenador aprende los parámetros de los modelos de las unidades de sonido empleando un conjunto de muestras de señales de voz, a lo que se le conoce como base de datos para el entrenamiento. El usuario también requiere definir cuales unidades de sonido tomará en cuenta el sistema para realizar el entrenamiento y, por lo menos, la secuencia en la cual ocurren dentro de cada señal de voz de la base de entrenamiento. Esta información se debe proporcionar por medio del archivo de transcripción (transcrip file), en el cual la secuencia de palabras y demás sonidos se escriben tal y como ocurren dentro de la señal de voz, seguidos por una etiqueta la cual puede ser usada para asociar esta secuencia con su correspondiente señal de voz.

Después el entrenador busca dentro de un diccionario, el cual mapea cada palabra a una secuencia de unidades de sonido, para derivar la secuencia de unidades de sonido asociada con cada señal. De este modo, además de las señales de voz, también se darán un conjunto de transcripciones para la base de datos (en un solo archivo) y dos diccionarios, uno dentro del cual se incluirán las palabras válidas, mismas que son secuencias mapeadas de unidades de sonido (o “sub word units”), y otro en el cual los sonidos que no se consideran señales de voz son mapeados a sus correspondientes unidades ya sea que se consideren similares a una unidad de sonido, o que simplemente no se consideren como voz. Nos referiremos al

primero como diccionario de lenguaje (“languaje dictionary”) y al segundo como diccionario de relleno (“filler dictionary”).

En suma, los componentes necesarios para realizar el entrenamiento serán:

1. El código fuente del entrenador.
2. Las señales de voz.
3. Los archivos de transcripción correspondientes.
4. Un diccionario de lenguaje.
5. Un diccionario de relleno.

2.3.2. Componentes para realizar el reconocimiento (decodificación)

De manera similar al sistema de entrenamiento, el decodificador también está compuesto por un conjunto de programas, los cuales han sido compilados para tener un solo ejecutable que sea capaz de realizar el reconocimiento, siempre y cuando se le den como entrada los parámetros correctos. Las entradas que tienen que proporcionarse son: los modelos acústicos del entrenamiento, un archivo de índice de modelo, un modelo de lenguaje, un diccionario de lenguaje, un diccionario de relleno y el conjunto de señales acústicas que se desea reconocer. A esta información, frecuentemente, se le conoce como datos de prueba.

En suma, los componentes necesarios para llevar a cabo el reconocimiento son:

1. El código fuente del decodificador (reconocedor).
2. El diccionario de lenguaje.
3. El diccionario de relleno.
4. El modelo de lenguaje.
5. Los datos de prueba.

Adicionalmente a estos componentes, se requieren también los modelos acústicos resultantes del entrenamiento, así como los archivos de índice de modelo, que son archivos que contienen identificadores para cada estado de cada “HMM”, mismos que son utilizados tanto por el entrenamiento como por el reconocimiento para acceder a los parámetros correctos de esos estados de “HMM”.

2.3.3. Configuración del sistema

Para poder configurar el sistema de reconocimiento es necesario contar con varios componentes. Una vez que se cumple con los requerimientos se procede a descargar los archivos propios del sistema de reconocimiento, es decir, el paquete de datos, el entrenador y uno de los decodificadores de Sphinx. A continuación se hace mención a algunos de los pasos que se deben seguir para la instalación.

Perl

Se requiere de Perl para poder utilizar los scripts que se proporcionan. La mayoría de las versiones de Linux cuentan ya con una versión de Perl; sin embargo, si se busca "correr" en windows, una versión popular es el Active Perl.

Compilador

Sphinxtrain y Sphinx 3 usan GNU autoconf para encontrar la información básica sobre el sistema de equipo, lo que permite que se pueda compilar en la mayoría de las versiones de Linux. Al mismo tiempo, en windows, es posible ejecutar los archivos de compilación usando Microsoft Visual C++, ya que se proporcionan los archivos de solución (.sln) y los proyectos (.vcproj) necesarios para compilar el código en dicho sistema operativo.

Ajuste de palabras ("word alignment")

Para facilitar la medición de la exactitud del decodificador es necesario emplear un programa de ajuste de palabras. Uno comúnmente empleado es el sclite, que está disponible en el "National Institute of Standards and Technology, NIST".

2.3.4. Configurar los datos

A fin de realizar pruebas de configuración, el Sphinx Group tiene disponibles dos bases de datos de audio, mismas a las que haremos referencia a lo largo de la configuración. Cada una tiene sus peculiaridades. Sin embargo, estas bases de datos no son suficientes para construir un sistema de reconocimiento de voz de alto desempeño. Como se mencionó, se emplean únicamente con el fin de verificar la correcta configuración del sistema.

Estas bases de datos son: AN4 o RM1. La base AN4 incluye el audio, pero es una base de datos muy pequeña. Se puede seleccionar si se quiere incluir la creación de los archivos de características en los experimentos. La base RM1 es un poco más grande, lo que nos permite obtener un sistema con un desempeño ligeramente mejor. En esta base no se proporciona el audio.

Pasos a seguir:

1. Crear un directorio para el sistema y moverse a dicho directorio:

   mkdir tutorial

   cd tutorial

2. Descargar los archivos de la base de datos, ya sea AN4 o RM1. Salvarlos en el mismo directorio del tutorial que se creó.

En windows, se descomprimen los archivos en la misma usando el explorador.

Al terminar, se tendrá un directorio llamado tutorial con los siguientes archivos:

Tutorial
- an4
- an4_sphere.tar.gz

O

Tutorial
- rm1
- rm1_cepstra.tar.gz

2.3.5. Configuración del entrenamiento (“trainer”)

Obtención del código

Se deben descargar los archivos para el entrenador, considerando que es importante hacer la descarga en el mismo directorio en el cual creamos (tutorial), en donde ya tenemos la base de datos. En windows, auxiliándonos del explorador, accedemos a la carpeta tutorial y extraemos en ella el Sphinxtrain.

Al terminar se tendrá la carpeta con los archivos siguientes:

Tutorial

- an4
- an4_sphere.tar.gz
- Sphinxtrain
- SphinxTrain.nightly.tar.gz

O

Tutorial

- rm1
- rm1_cepstra.tar.gz
- Sphinxtrain
- SphinxTrain.nightly.tar.gz

Compilación de sphinxtrain

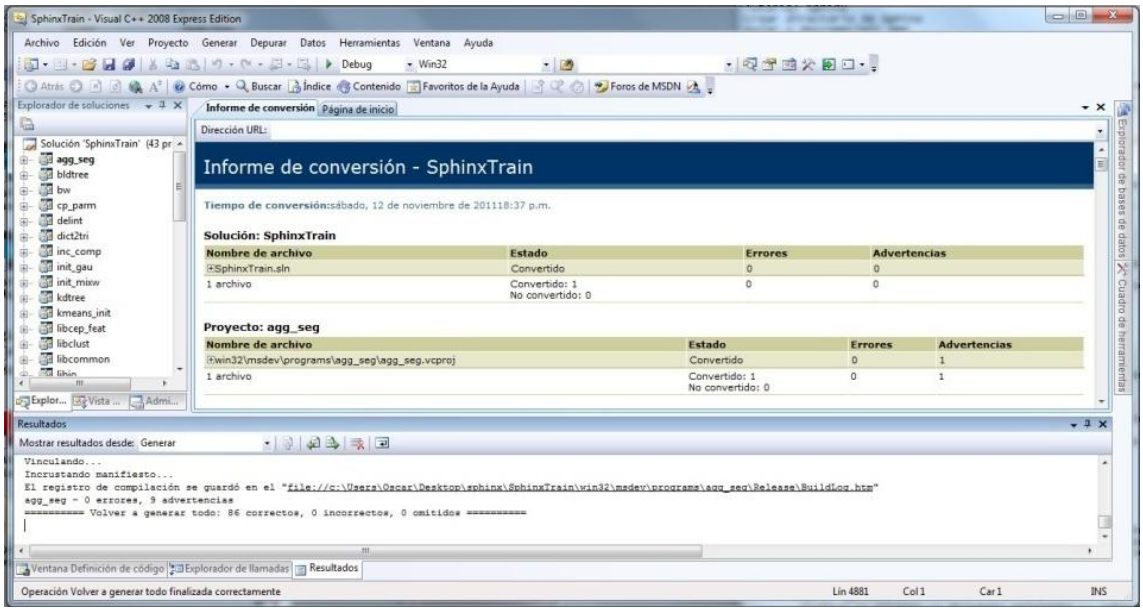


Figura II-4. Compilación del archivo Sphinxtrain.

En la figura II-4 se observa una captura de pantalla del resultado obtenido al compilar el archivo Sphinxtrain con el programa ***Visual C++ 2008 Express Edition***.

En windows:

1. Damos doble clic en el archivo tutorial/Sphinxtrain/SphinxTrain.sln. Esto deberá abrir el MS Visual C++, mismo que debe estar instalado previamente.
2. En el menú Build escogemos Batch Build y seleccionamos todos los objetos. Damos Clic en Rebuild All. Esto generará todos los ejecutables requeridos por el entrenador.

En el caso de la base de datos que estamos empleando como prueba, es necesario que después de compilar el código, hagamos la configuración pertinente copiando todos los ejecutables indispensables y los scripts a la misma área de los datos. Si pensamos que estamos trabajando en el directorio tutorial que creamos, debemos hacer lo siguiente:

```
cd SphinxTrain
# si instalamos AN4
perl scripts_pl/setup_tutorial.pl an4

# si instalamos RM1
perl scripts_pl/stup_tutorial.pl rm1
```

2.3.6. Configuración del decodificador (“decoder”)

El Sphinx Group tiene varios decodificadores cuyas características pueden servirnos como guía para seleccionar cual es el adecuado, estos se describen a continuación.

- PocketSphinx: esta es una versión modernizada del Sphinx 2, especialmente optimizada para sistemas embebidos y “Personal Digital Assistant, PDA”. al mismo tiempo consume en promedio 20% menos memoria y del 5-20% menos tiempo de procesamiento del CPU que Sphinx 2. Sin embargo, se encuentra en constante desarrollo, por lo que la interfaz y sus características pueden ser inestables.

- Sphinx 3: utiliza HMMs continuos. Soporta decodificación tanto batch como live. Actualmente este es el decodificador más desarrollado.

- Sphinx 4: utiliza HMMs continuos. Fue realizado en el lenguaje de programación Java. Permite una alta flexibilidad y una gran precisión y velocidad para pequeñas tareas.

2.3.7. Instalación de Sphinx 3

Obtención del código

Se deben descargar los archivos sphinx3 y sphinxbase. Es importante guardarlos en el directorio tutorial que hemos creado y extraer su contenido en la misma carpeta.

En windows, auxiliándose del explorador de windows, accedemos a la carpeta tutorial y extraemos en ella los archivos tanto del Sphinx3 como del Sphinxbase.

Al terminar tendremos en la carpeta los archivos siguientes:

Tutorial

- an4
- an4_sphere.tar.gz
- Sphinxtrain
- SphinxTrain.nightly.tar.gz
- sphinx3
- sphinx3.nightly.tar.gz
- sphinxbase
- sphinxbase.nightly.tar.gz

O

Tutorial

- rm1
- rm1_cepstra.tar.gz
- Sphinxtrain
- SphinxTrain.nightly.tar.gz
- sphinx3
- sphinx3.nightly.tar.gz
- sphinxbase

- sphinxbase.nightly.tar.gz

Compilación de sphinxbase y sphinx3

En la figura II-5 y II-6 se observa una captura de pantalla del resultado obtenido al compilar el archivo sphinxbase y sphinx3, respectivamente, con el programa ***Visual C++ 2008 Express Edition***.

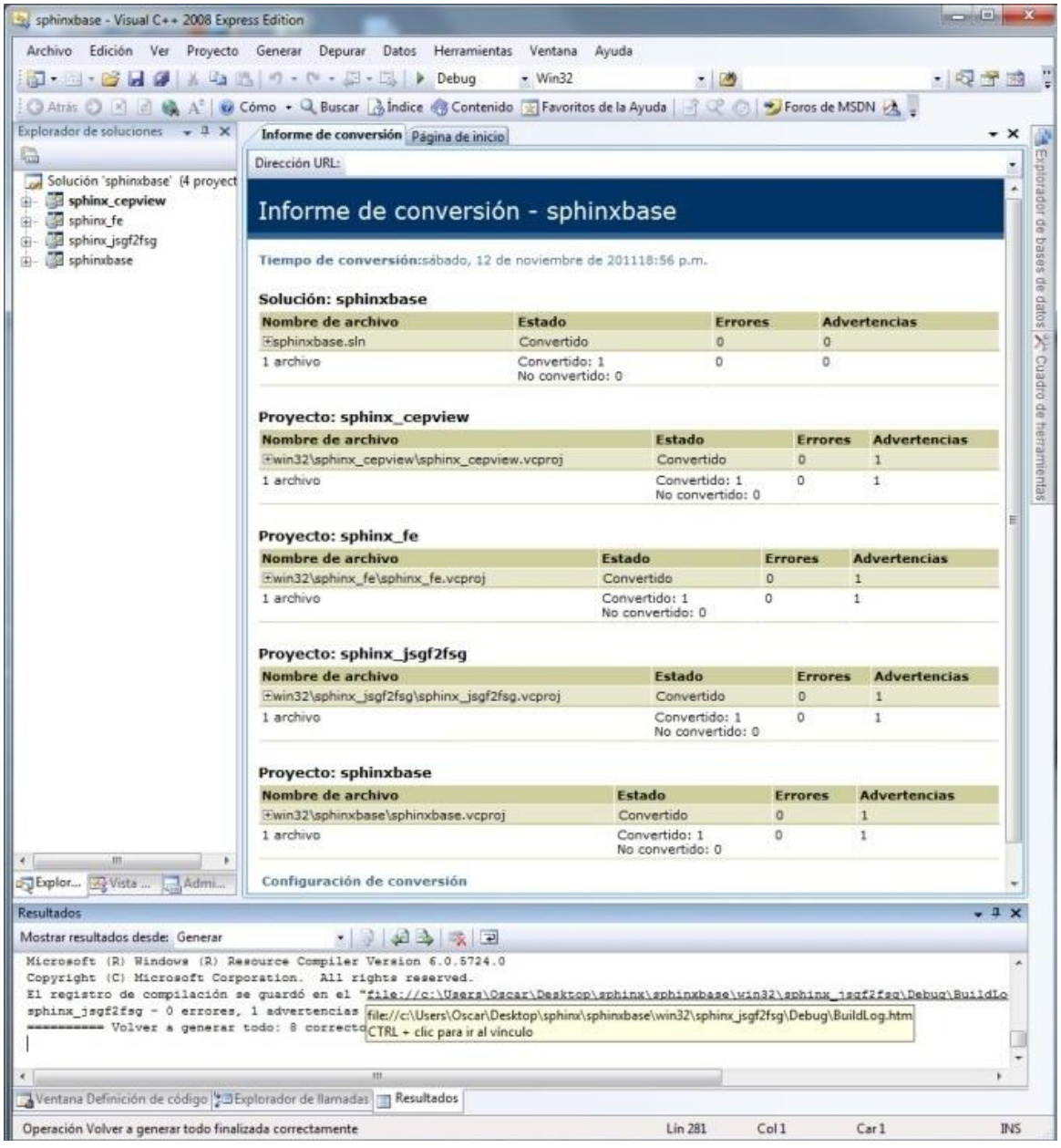


Figura II-5. Compilación del archivo sphinxbase.

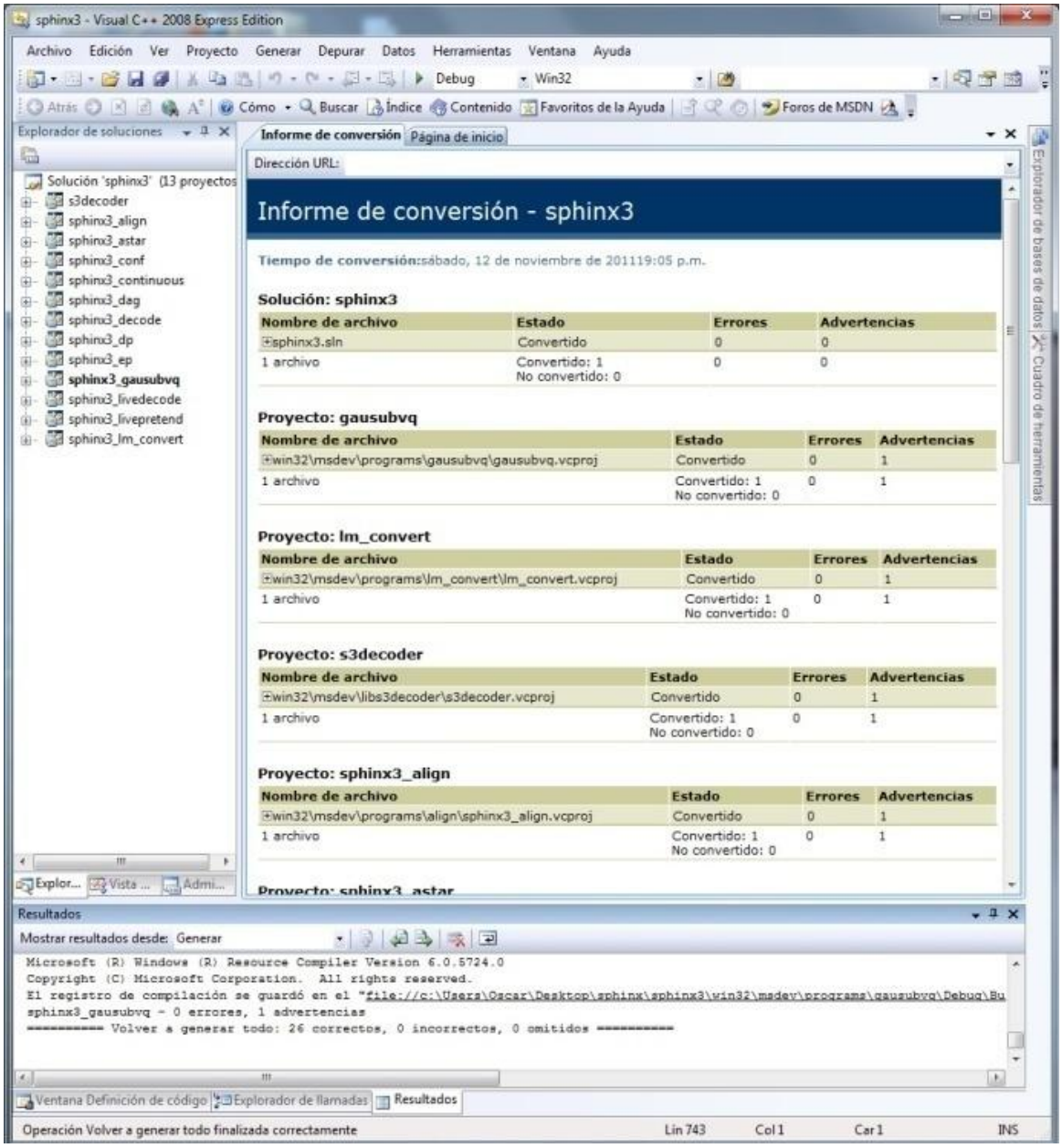


Figura II-6. Compilación del archivo sphinx3.

En windows, primero es necesario renombrar la carpeta sphinxbase-xx por sphinxbase y entonces:

1. Dar doble clic en el archivo tutorial/sphinxbase/sphinxbase.sln. Esto abrirá el MS Visual C++, mismo que debe estar previamente instalado.

2. En el menú build escogemos batch build y seleccionamos todos los objetos. Damos clic en Rebuild All, lo cual generará todas las librerías en el paquete de sphinxBase.

3. Damos doble clic en el archivo tutorial/sphinx3/programs.sln. Esto abrirá el MS Visual C++, mismo que debe estar previamente instalado.
4. En el menú Build, escogemos Batch build y seleccionamos todos los objetos. Damos clic en Build All, lo cual generará todos los ejecutables en el paquete de sphinx3.

2.3.8. Configuración del tutorial

Después de haber compilado el código, debemos configurar el tutorial copiando los ejecutables y los scripts necesarios al mismo lugar donde están los datos.

Si pensamos que estamos trabajando en el directorio tutorial que creamos, debemos hacer lo siguiente:

```
cd sphinx3

# si instalamos AN4
perl scripts/setup_tutorial.pl an4

# si instalamos RM1
perl scripts/stup_tutorial.pl rm1
```

2.3.9. Como llevar a cabo una corrida de entrenamiento preliminar o de prueba

Primero es necesario ir al directorio donde se instalaron los datos. Habiendo hecho esto, en windows, hacemos lo siguiente:

```
# si estamos usando AN4
cd..\an4

# si estamos usando RM1
cd..\rm1
```

El sistema no trabaja directamente con señales acústicas. Las señales son transformadas primero en una secuencia de vectores de características, los cuales son utilizados en lugar de las señales acústicas correspondientes. Para llevar a cabo esta transformación (obtención de los parámetros) desde el directorio an4, se teclea la siguiente instrucción en la línea de comandos. Recordando que si se está trabajando en windows se debe reemplazar / por \.

Por otro lado, si se está trabajando con rm1 no es necesario realizar esta transformación, ya que estos ya son datos en el formato "cepstral". La instrucción es:

Perl scripts_pl\make_feats.pl –ctl etc\an4_train.fileids

Este script calculará, para cada secuencia de entrenamiento, una secuencia de de vectores de dimensión 13 (vectores de características) conformados por coeficientes "Mel-Frequency Cepstral, MFCC's". Hay que notar que la lista de archivos wave contiene una lista con las rutas completas a los archivos de audio. Dado que los datos están localizados en el mismo directorio en que se trabaja, las rutas son relativas, no absolutas. Es posible que esto se quiera cambiar, además del archivo an4_test.fileids, si la ubicación de los datos es diferente. Este paso requiere de unos pocos minutos, si se realiza en una computadora medianamente rápida, aunque el tiempo puede variar.

Los "MFCC's" se guardan automáticamente en el directorio .\feat. Hay que hacer notar que el tipo de vectores de características que se calculan a partir de las señales de voz para el entrenamiento y el reconocimiento no está restringido al uso de los "MFCC's"; es decir, que es posible emplear otro tipo de vectores de características.

En el directorio de los scripts (.\scripts_pl), se generan varios directorios numerados del 00* al 99*. Cada directorio tiene a su vez un directorio nombrado como slave*.pl. Se puede acceder a cada uno de estos directorios y ejecutarlos. O de manera alternativa, se puede ejecutar simplemente el script RunAll.pl.

Perl scripts_pl\RunAll.pl

Los scripts cargaran tareas en el equipo, y las mismas durarán algunos minutos para realizarse. Antes de correr los scripts podemos abrir los directorios para observar que archivos se encuentran en ellos, de manera que si los abrimos nuevamente después de "correr" los scripts notemos que se crean nuevas carpetas y nuevos archivos. Uno de los archivos que se crean en el directorio de trabajo es un .html, llamado en este caso an4.html o rm1.html, dependiendo de la base de datos que se esté utilizando. Este archivo contiene un reporte del estado de las tareas que se hayan ejecutado hasta el momento. Antes de abrirlo se recomienda verificar si las tareas se han terminado de realizar.

Al pasar por el proceso de ejecutar los scripts del 00* al 99*, se habrán generado un conjunto de modelos acústicos, cada uno de los cuales puede ser empleado en el reconocimiento. Podremos notar también que algunos de los pasos se requieren únicamente para la creación de los modelos semicontinuos.

Una vez que se han terminado de ejecutar completamente los procesos cargados hasta el 20.ci_hmm se habrán entrenado entonces los modelos independientes del contexto para las unidades de palabra en el diccionario. Cuando se han ejecutado completamente las tareas cargadas del directorio 30.cd_hmm_untied, se habrán entrenado los modelos para las unidades de palabra dependientes del contexto (trifonemas) con estados no ligados.

Estos son llamados modelos no ligados dependientes del contexto y son necesarios para construir los arboles de decisión de manera que se puedan ligar los estados. Las tareas en 40.buildtrees generan los árboles de decisión para cada estado por unidad de palabra. Las tareas en 45.prunetree llevan a cabo el ligado de los estados considerando los árboles de decisión. Siguiendo con el proceso, las tareas en 50-cd-hmm_tied van a entrenar los modelos finales para los trifonemas en el corpus de entrenamiento, estos se conocen como los modelos ligados dependientes del contexto. Los modelos ligados dependientes del contexto se entrenan en varias etapas, comenzando con una gaussiana por estado y así sucesivamente hasta el número previamente definido de gaussianas.

Una vez terminado lo anterior, se ha completado el entrenamiento. La ubicación de los modelos finales dependerá de la base de datos y el tipo de modelo que se esté usando.
Durante este entrenamiento aparecen algunos errores no críticos como:

“This step had 6 ERROR messages and 2 WARNING messages. Please check the log file for details”.

No hay que preocuparse por ellos, es muy probable que se presenten una serie de errores de alineación en la base de datos.

2.3.10. Como llevar a cabo una decodificación preliminar o de prueba

La decodificación es relativamente fácil de realizar. Primero se calculan los "MFFC's" para todas las iteraciones en el conjunto de datos prueba. Si se trabaja con los archivos rm1, no se requiere hacer nada más, ya que estos datos ya están en el formato "cepstral". En el caso de trabajar con los datos de an4, se requiere ejecutar la siguiente instrucción para obtener los parámetros necesarios:

Perl scripts_pl\make_feats.pl –ctl etc\an4_test.fileids

Esto lleva unos cuantos minutos, al término de los cuales ya se puede realizar la decodificación. Para ello escribimos el siguiente comando:

Perl scripts_pl\decode\slave.pl

Esto utiliza todos los componentes proporcionados para realizar la decodificación, incluyendo los modelos acústicos y el archivo de índices de modelo que previamente se generó durante el entrenamiento. Cuando el proceso de reconocimiento es completado, el script calcula la tasa de error de palabra para el reconocimiento ("Word Error Rate, WER") o la tasa de error de oración ("Sentence Error Rate, SER").

Cuando se ejecuta el script de decodificación, este imprime la información acerca de la precisión hasta arriba de la pagina .html para el experimento en que se esté trabajando. Al mismo tiempo genera dos conjuntos de archivos. Uno de estos conjuntos, con extensión .match, contiene las hipótesis como salidas a través de decodificador. El otro conjunto, con extensión .align, contiene la alineación generada por algún programa de alineación programado, o por el que contiene Sphinx, con el resultado de la comparación entre las hipótesis del decodificador y las transcripciones dadas.

En la figura II-7 podemos ver una captura de pantalla del resultado obtenido al ejecutar el programa Sphinx 3 empleando las bases de prueba an4.

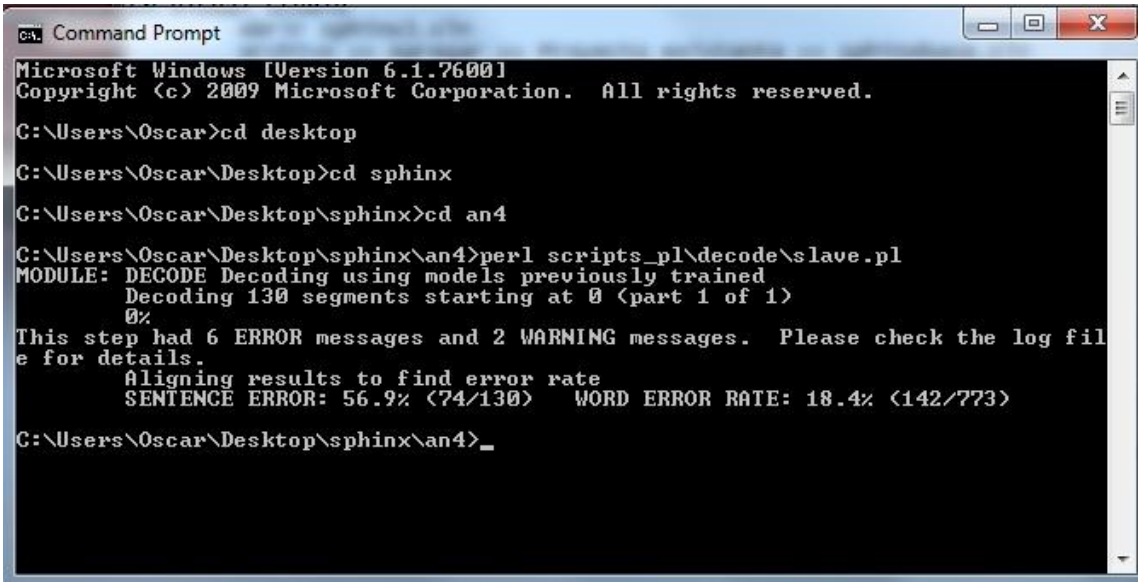


Figura II-7. Corrida del programa Sphinx 3 empleando las bases de prueba, an4.

Si se usa la herramienta del "NIST", el archivo .html contendrá una línea como la siguiente:

Si se utilizó an4:

SENTENCE ERROR: 56.9% (74/130) WORD ERROR RATE: 18.4% (142/773)

O, si se utiliza rm1:

SENTENCE ERROR: 38.833% (233/600) WORD ERROR RATE: 7.640% (434/5681)

2.3.11. Como realizar un entrenamiento y algunos temas claves

Ahora estamos listos para comenzar con los experimentos propios. Para cada entrenamiento y decodificación, es necesario primero fijarle un nombre. Un nuevo proyecto se crea a partir de uno ya existente. Asumiendo que se copia la configuración de una ya existente.

En el siguiente ejemplo, se hace justamente eso: se copia la configuración existente en an4. Es importante recordar que debemos estar en el directorio ya existente para entonces poder "correr" el siguiente script:

```
cd an4
perl script_pl\copy_setup.pl –task  $taskname
```

Esto creará una nueva configuración al volver a "correr" la configuración de Sphinxtrain y después volver a "correr" la configuración de decodificador tal y como se realizó en la configuración original de an4 y finalmente copiando los archivos de configuración, localizados dentro de la carpeta etc, a la nueva configuración, donde se deben renombrar los archivos para que correspondan con el nombre del nuevo proyecto y se pueda realizar la decodificación

Hay que tener cuidado al realizar la copia, ya que el script copy_setup.pl también copia los datos contenidos dentro de feat y de wav, a la nueva ubicación. Por lo que si el conjunto de datos es grande, esta duplicidad nos llevara a un desperdicio del espacio en memoria.

2.4. Puesta en marcha del reconocedor con el corpus del laboratorio de procesamiento de voz de la FI UNAM

Una vez que se generó la carpeta con una copia de la configuración para "correr" Sphinx 3, nos dimos a la tarea de realizar la configuración, así como, las adecuaciones pertinentes para llevar a cabo el reconocimiento de voz, empleando el corpus del laboratorio de procesamiento de voz de la FI UNAM, conformado por el M.I. Jaime Alfonso Reyes Cortés.

Como se menciona previamente, para realizar el reconocimiento sobre otras bases (corpus) diferentes de las de prueba, realizamos lo siguiente:

```
cd an4
perl script_pl\copy_setup.pl –task  $Tesis
```

Como se menciona en la tesis de maestría del M.I. Jaime Reyes [2], antes de realizar el proceso de entrenamiento en el sistema Sphinx 3 se requieren de los siguientes elementos de apoyo:

- Un conjunto de archivos de características calculados previamente con las señales de audio para el entrenamiento, donde cada archivo representa una grabación que se tiene en el "corpus" de entrenamiento. Las grabaciones se pueden convertir en una secuencia

de vectores de características haciendo uso del ejecutable wav2feat proporcionado con el software Sphinxtrain y cuyos parámetros tienen un valor por defecto.

Para el trabajo del M.I. Jaime Reyes se trabaja con los archivos que se encuentran dentro de Tesis_train.fileids y tienen formato raw, los cuales se encuentran en el directorio wav, que tienen extensión pcm, mismos que se guardan en el directorio feat, con la extensión mfc, con una frecuencia de muestreo de 16 kHz, en donde se utiliza un banco de 40 filtros triangulares.

- Un segmento de un archivo de características se muestra a continuación:

$$
\begin{array}{l}
226 frames \\
11.851 \ -0.916 \ -0.660 \ -0.203 \ -0.215 \ 0.454 \ -0.148 \ -0.183 \ 0.179 \ -0.081 \\
12.079 \ -0.707 \ -0.379 \ -0.291 \ -0.300 \ 0.208 \ -0.264 \ -0.301 \ 0.085 \ -0.065 \\
11.728 \ -0.594 \ -0.405 \ -0.211 \ \ 0.004 \ 0.317 \ -0.138 \ -0.034 \ \ 0.204 \ \ 0.046 \\
11.168 \ -0.423 \ -0.432 \ -0.348 \ -0.253 \ 0.287 \ -0.129 \ -0.224 \ \ 0.287 \ \ 0.098 \\
\vdots
\end{array}
$$

Donde “frames” indica el número de tramas que forman a la señal y los valores de cada renglón representan un segmento de los coeficientes “MFCC” calculados.

- Un archivo de control que contiene una lista de nombres de archivos y la ruta completa de los vectores de características. Un ejemplo del contenido de este archivo para el entrenamiento es el siguiente:

BeatrizC/TRAIN/PARA/BC01
BeatrizC/TRAIN/PARA/BC02
BeatrizC/TRAIN/PARA/BC03
⋮

Estos archivos no llevan las extensiones. Estas serán proporcionadas de manera separada durante el entrenamiento, debido a que se desea dar nombres únicos a todos los archivos. De igual forma se incluye la dirección completa para hacer cada entrada única en el archivo de control.

- Un archivo de transcripción en el cual se enlista la transcripción textual correspondiente a los archivos de características, exactamente en el mismo orden como se enlistan los nombres de los archivos de características en el archivo de control. Un segmento del archivo de transcripción utilizado es:

  <S> ++HEM++ PARA OCHENTA Y OCHO PUNTO NUEVE NOTICIAS </S> (BC01_00)
  <S> PARA PANORAMA INFORMATIVO OCHENTA Y OCHO PUNTO NUEVE NOTICIAS </S> (BC01_1)
  <S> PARA EL AUDITORIO QUE NOS ESCUCHA EN LA CIUDAD DE MÉXICO HAY TRÁNSITO PESADO EN </S> (BC02_1)
  ⋮

- Un léxico o diccionario de pronunciación, que es un archivo que especifica las pronunciaciones de las palabras que son de interés para el entrenador y el decodificador. Por ejemplo, para una parte del vocabulario utilizado en el trabajo del M.I. Jaime Reyes el diccionario es:

AUDITORIO	a u d i t o r(i o
CIUDAD	Z i u d a d
DE	d e
EL	e l
EN	e n
ESCUCHA	e s k u TS a
HAY	a i

  ⋮

- Un diccionario de relleno o diccionario de ruido, que usualmente define las palabras legales que no están contenidas en el modelado del lenguaje pero que son parte del lenguaje mismo, como aspiraciones, pausas, etc.

  Este diccionario debe tener al menos los siguientes valores:
  <S> SIL
  </S> SIL

<SIL> SIL

Donde <S> es el silencio del comienzo de la pronunciación de la oración, <SIL>, es el silencio dentro de la pronunciación y </S> el silencio al final de la pronunciación de la oración. Estas palabras son tratadas como palabras especiales y es necesario que sean representadas dentro del diccionario de relleno. Al menos una de ellas debe estar relacionada con un fonema llamado "SIL", que representa el silencio entre palabras.

- Un archivo con la lista de fonemas, que es una lista de todas las unidades acústicas que se han seleccionado para entrenar el modelo acústico. Es importante notar que Sphinx no permite unidades acústicas diferentes a las de sus diccionarios. Todas las unidades en sus dos diccionarios deben ser enlistadas aquí. Dicho de otra forma, su lista de fonemas debe contener exactamente las mismas unidades usadas en sus diccionarios, ni más ni menos. Cada fonema debe ser enlistado en una línea separada dentro del archivo, comenzando a la izquierda, con ningún espacio extra después del fonema. Como ejemplo se tiene:

a
a_7
b
k
TS
d
⋮

Una vez que se tienen todos los archivos recién mencionados se puede continuar con el entrenamiento. En este punto es necesario renombrar los archivos para que coincidan con los generados con la instrucción perl script_pl\copy_setup.pl –task $Tesis, de tal modo, que sean estos y no los generados con copy los que operen.

Se debe modificar en el archivo sphinx_train.cfg, el número de estados ligados (fijándolos a 500) y el número de gaussianas (fijándolo a 1).

Ahora sí se puede ejecutar el comando para el entrenamiento

perl scripts.\pl makefeats

perl scripts.pl\runall

Esto, recordando las direcciones correctas del directorio en donde se está trabajando.

2.4.1. Prueba del funcionamiento del sistema Sphinx 3 con el corpus del laboratorio de procesamiento de voz de la FI UNAM

Una vez terminado el proceso de entrenamiento se puede realizar el reconocimiento, habiendo realizado las configuraciones pertinentes en el archivo sphinx_decode.cfg.

En la figura II-8 se muestra una captura de pantalla de la "corrida" del programa de reconocimiento empleando el corpus del laboratorio de procesamiento de voz de la FI UNAM, conformado por el M.I. Jaime Reyes, para el entrenamiento y reconocimiento.

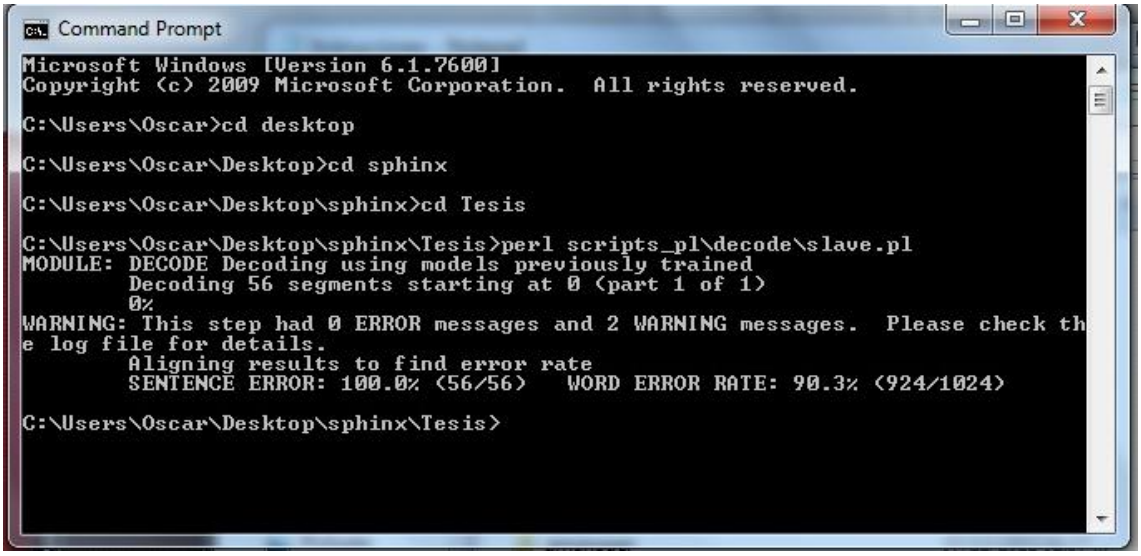


Figura II-8. "Corrida" del programa Sphinx 3 empleando el corpus del laboratorio de procesamiento de voz de la FI UNAM, conformado por el M.I. Jaime Reyes.

SENTENCE ERROR: 100.0% (56/56) WORD ERROR RATE: 90.3% (924/1024)

Finalmente, si queremos conocer los detalles de la ejecución del software Sphinx 3 podemos abrir el archivo .html que se genera.

2.5. Propuesta para mejorar el desempeño del sistema de reconocimiento, conformación de un nuevo corpus

Dado que el desempeño del sistema de reconocimiento de voz Sphinx 3 con el corpus disponible en el laboratorio de procesamiento de voz de la FI UNAM no es lo suficientemente bueno, como podemos observar en la figura II-8 ("SER" del 100% y "WER" del 90.3%), nos propusimos buscar la manera de mejorarlo.

Revisando el corpus notamos que una probable limitante para el correcto funcionamiento del sistema era el reducido número de repeticiones de las palabras que lo conforman, por lo que buscamos una forma de mejorarlo. Nuestra propuesta contempló el aumentar a por lo menos 3 repeticiones de cada una de las palabras.

Para conseguir aumentar el número de repeticiones se grabaron 300 nuevas frases adicionales a las que se tenían para conformar el nuevo corpus. El aumentar las repeticiones de ciertas palabras nos condujo a proponer las frases a grabar, esto se hizo considerando el diccionario del corpus existente. Durante la generación de las frases a grabar se buscó que las palabras presentaran distinta ubicación dentro de las frases, de manera que los parámetros obtenidos resultaran más representativos.

2.6. Metodología para llevar a cabo las grabaciones del corpus

1. Se conformaron las frases para grabar, tomando como base el corpus del laboratorio de procesamiento de voz de la FI UNAM.
2. Se emplearon las mismas palabras, es decir que no se agregaron nuevas palabras en el diccionario, únicamente se agregaron nuevas pronunciaciones.
3. Se conformó el nuevo corpus teniendo al menos 3 pronunciaciones de cada una de las palabras.

4. Para realizar la grabación de las frases se empleó el software Adobe Audition, utilizando la siguiente configuración para realizar las grabaciones.
   - Archivos de audio en formato .wav
   - Frecuencia de muestreo de 44100 [Hz]
   - 16 bit
   - Mono

En la figura II-9 se muestra una señal de voz visualizada en Adobe Audition.

Figura II-9. Visualización de una señal de voz en Adobe Audition (señal en el tiempo arriba, y su espectro abajo).

5. Las grabaciones se realizaron en una habitación común y corriente, cuidando en la medida de lo posible que los niveles de ruido ambiente fueran los mínimos.

2.6.1. Hardware utilizado para las grabaciones de voz

- Micrófono con conexión USB marca Logitech.
- Pedestal para micrófono (con el fin de reducir las vibraciones).
- Computadora portátil, con tarjeta de sonido integrada a la tarjeta madre.

2.6.2. Procesamiento posterior de las grabaciones de voz

1. Una vez terminado el proceso de grabación en la computadora, se procedió a delimitar las frases dejándolas con una pequeña porción de silencio al inicio y al final.
2. Se sub-muestrearon las señales de audio para adecuarlas a las especificaciones del sistema Sphinx 3.
3. Posteriormente se procedió a convertir el formato de los archivos de audio de formato .wav a formato .raw con codificación pcm, para homologarlos con los existentes en el corpus anterior.
4. A continuación se realizó el etiquetado de las frases auxiliándonos del software Praat.

En la figura II-10 se observa la visualización de una señal de voz para su etiquetado en Praat.

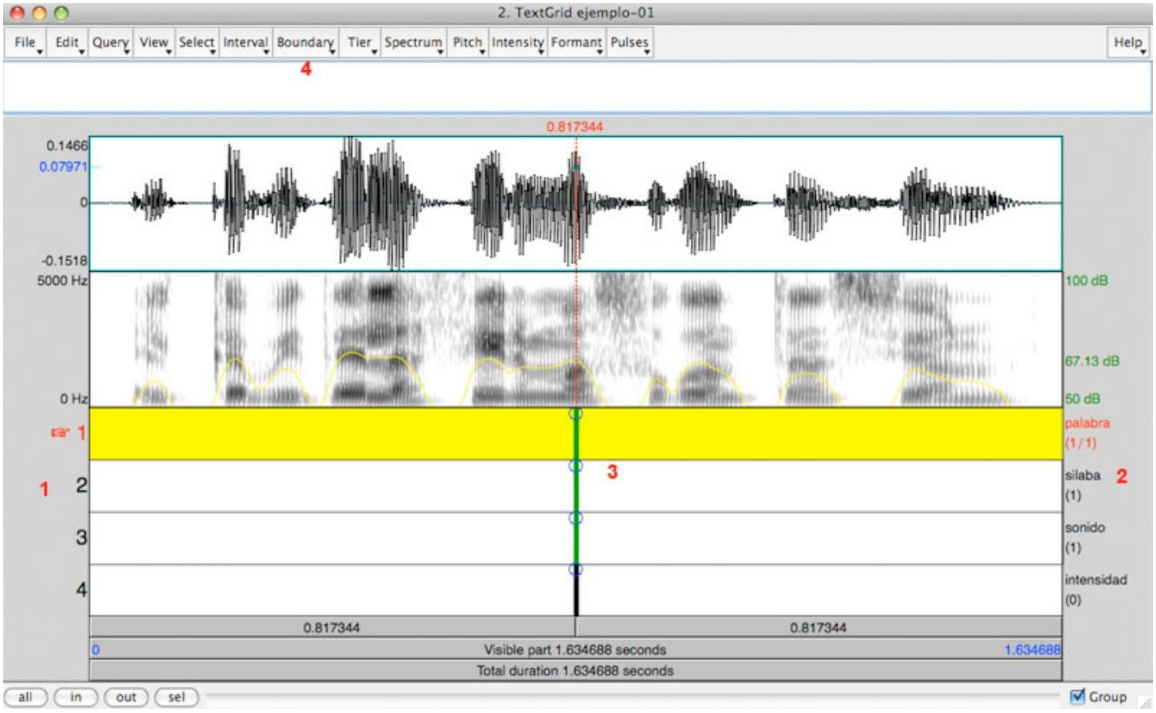


Figura II-10. Visualización de una señal de voz en Praat al realizar el etiquetado (señal en el tiempo, espectro, y los diversos niveles de etiquetado).

2.6.3. Ajustes a los archivos de configuración para agregar las nuevas frases al corpus.

Para poder emplear los nuevos archivos, junto con los existentes en el corpus del laboratorio de procesamiento de voz de la FI UNAM, es necesario crear o modificar algunos de los utilizados por el sistema Sphinx 3. Estos son:

- El archivo de transcripción fonética para el entrenamiento.
- El archivo de identificación para el entrenamiento.
- El archivo de transcripción fonética para la prueba.
- El archivo de identificación para la prueba.

Se colocaron los archivos de audio en las mismas carpetas donde se encuentran los archivos del corpus del laboratorio de procesamiento de voz de la FI UNAM.

2.7. Pruebas con el nuevo corpus

2.7.1. Resultados obtenidos con el nuevo corpus

Ahora sí se puede ejecutar el comando para el nuevo entrenamiento

perl scripts.\pl makefeats

perl scripts.pl\runall

Esto, recordando las direcciones correctas del directorio en donde se está trabajando.

Una vez terminado el proceso de entrenamiento y habiendo realizado las configuraciones pertinentes en el archivo sphinx_decode.cfg, se realizó el reconocimiento. En la figura II-11 se muestra una captura de pantalla de la “corrida” del programa de reconocimiento, empleando el nuevo corpus para el entrenamiento y reconociendo frases del corpus original.

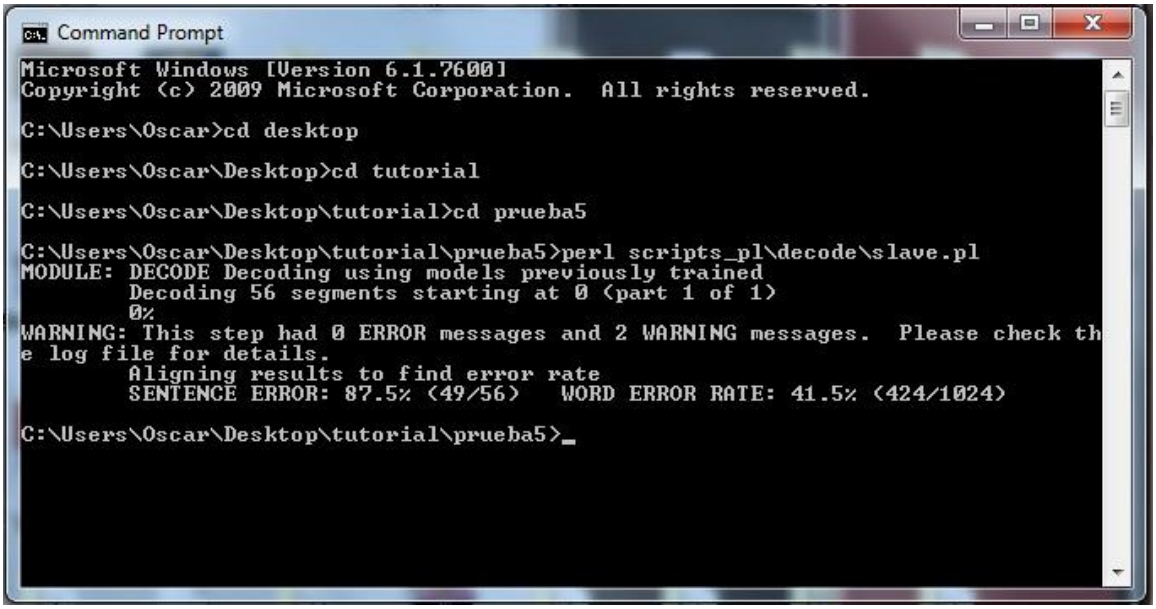


Figura II-11. Captura de pantalla al ejecutar el programa de reconocimiento Sphinx 3, empleando el nuevo corpus para el entrenamiento y reconociendo frases del corpus original.

SENTENCE ERROR: 87.5% (49/56) WORD ERROR RATE: 41.5% (424/1024)

En la figura II-12 se muestra una captura de pantalla de la corrida del programa de reconocimiento empleando el nuevo corpus para el entrenamiento y reconociendo frases del nuevo corpus.

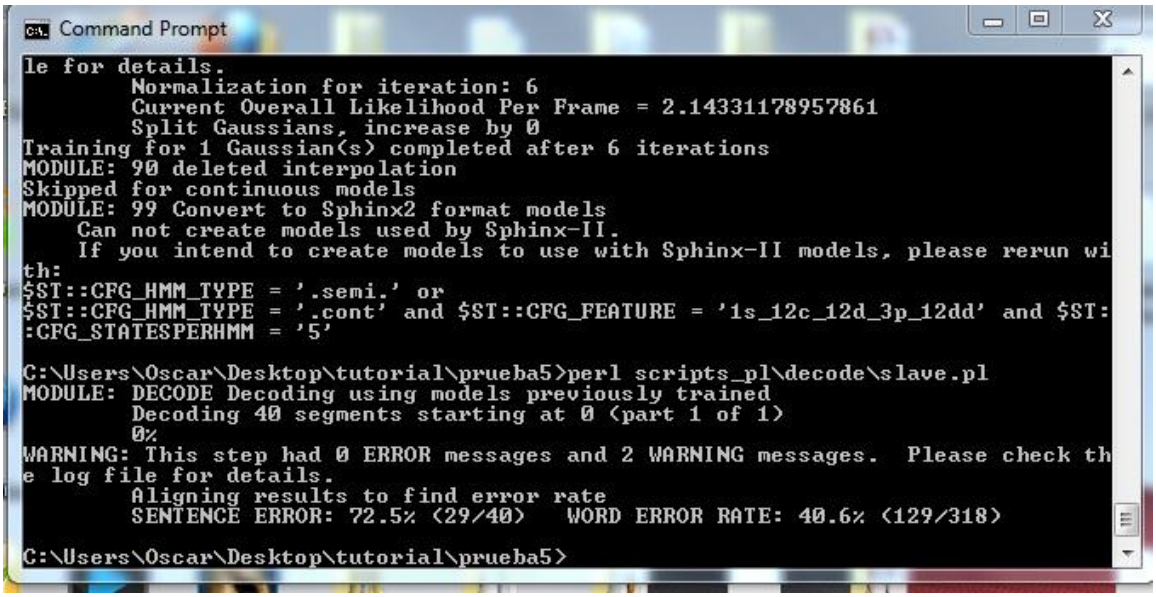


Figura II-12. Captura de pantalla al ejecutar el programa de reconocimiento Sphinx 3, empleando el nuevo corpus para el entrenamiento y reconociendo frases del nuevo corpus.

SENTENCE ERROR: 72.5% (29/40) WORD ERROR RATE: 40.6% (129/318)

Para validar el correcto funcionamiento de las nuevas frases, se realizaron pruebas por separado, logrando resultados mejores. Para realizar el entrenamiento y reconocimiento se consideraron únicamente las frases nuevas. El resultado de dichas pruebas se observa en la figura II-13.

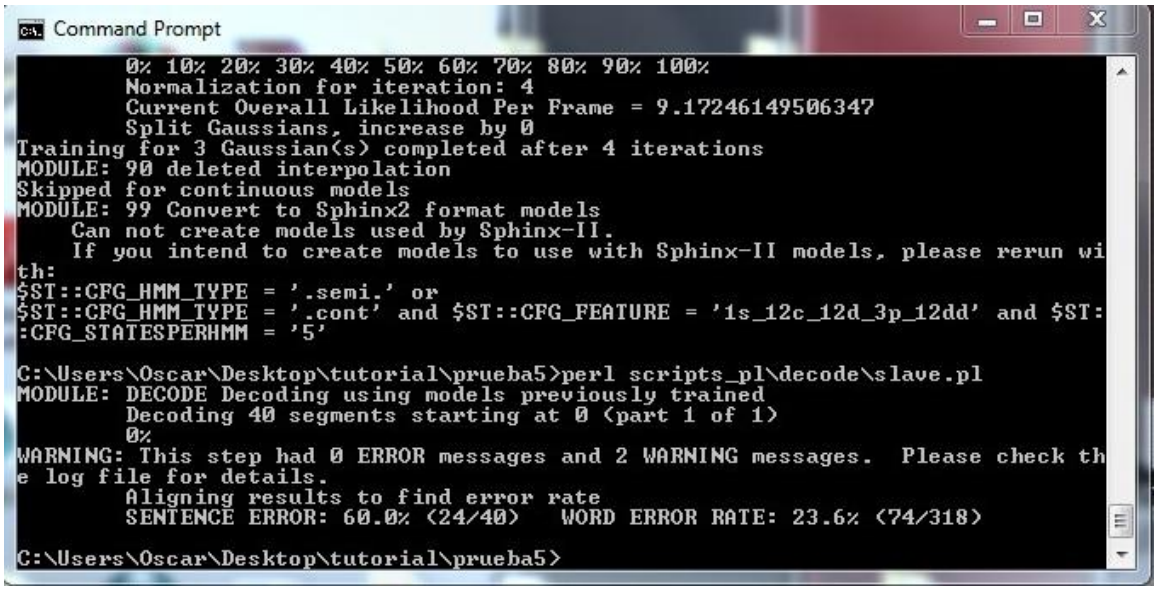


Figura II-13. Captura de pantalla al ejecutar el programa de reconocimiento Sphinx 3, haciendo uso únicamente de las frases recién grabadas.

SENTENCE ERROR: 60.0% (24/40) WORD ERROR RATE: 23.6% (74/318)

2.7.2. Comparación de resultados

La evaluación fuera de línea es quizás una manera más objetiva de realizar la evaluación de los sistemas de voz. Usualmente se graba primero un conjunto de audios de prueba y después se puede calcular el "Sentence Error Rate, SER" o el "Word Error Rate, WER". Recordando que entre menor sea el error es mejor el desempeño [7].

El "Sentence Error Rate" se define como el porcentaje del número de frases no reconocidas durante la prueba y de manera similar el "Word Error Rate" como el porcentaje del número de palabras no reconocidas durante la prueba.

Por otro lado el "Word Accuracy" (A) usualmente se define como:

$$A = \frac{H - S - I - D}{H} x\ 100\%$$

Esto de manera que se tengan en cuenta los errores por inserción (I) o por eliminación (D) al momento de calcular la exactitud "accuracy" (A).

Donde:

H es el número total de palabras en referencia

S es el número de errores de sustitución

I es el número de errores de inserción

D es el número de errores por eliminación

En la tabla II-2 se comparan los resultados obtenidos empleando los diferentes corpus y considerando diferentes frases de prueba.

Tabla II-2. Comparación de resultados usando Sphinx 3.

	"Sentence Error Rate"	**"Word Error Rate"**
Corpus Original - Frases Originales	100.00%	90.3%
Corpus Nuevo - Frases Originales	87.50%	41.50%
Corpus Nuevo - Frases Nuevas	72.50%	40.60%
Corpus Adicional - Frases Adicionales	60.00%	23.60%

III. Redes neuronales y procesamiento de voz

3.1. Conceptos generales

Una red neuronal es un modelo artificial inspirado en el sistema nervioso del cerebro humano [19]. Por artificial debemos entender algo que es producido por el humano más que de forma natural. Entonces, una red natural artificial, RNA, es una representación matemática del sistema nervioso del cerebro humano. Las neuronas reales en el cerebro humano integran un amplio rango de señales temporales por medio de las dendritas. En el cerebro humano la transformación de la información es llevada a cabo por las neuronas.

En la figura III-1 se observa una porción de una red compuesta por dos células. La célula contiene un **núcleo** (la porción de procesamiento central de la neurona). A la izquierda de la célula 1, las **dendritas** proporcionan las señales de entrada a las células. A la derecha, el **axón** envía señales de salida a la célula 2 por medio de las terminales del axón. Las terminales del axón se unen con las dendritas de la célula 2. Las señales pueden ser transmitidas sin cambio alguno, o pueden ser modificadas por medio de sinapsis. Una **sinapsis** es capaz de incrementar o disminuir la fuerza de la conexión de neurona a neurona y causar excitación o inhibición de una neurona subsecuente. Es aquí donde se almacena la información.

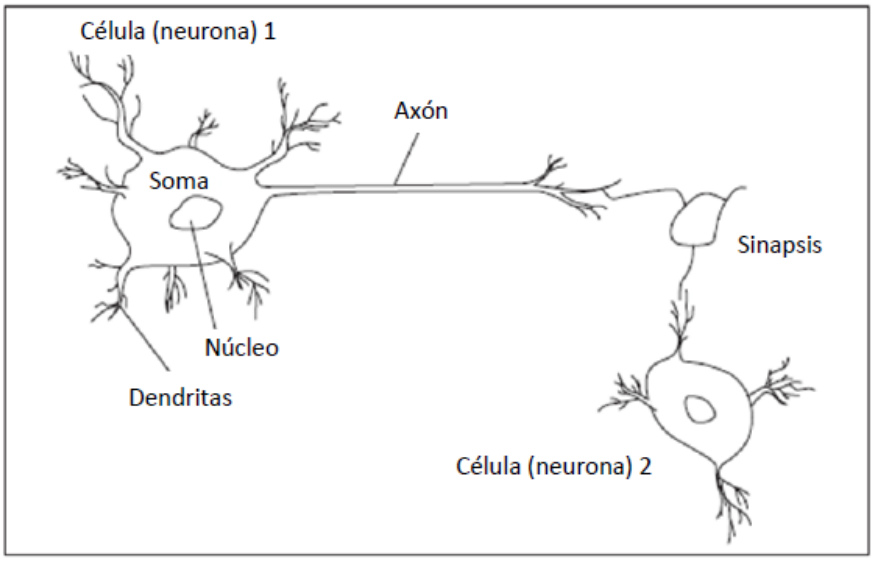


Figura III-1. Porción de una red: dos células biológicas conectadas [20].

El modelo de una RNA, o "ANN" por sus siglas en inglés, emula una red neuronal biológica. Actualmente, la computación neuronal utiliza un conjunto muy limitado de conceptos de los sistemas neuronales biológicos. Es más una analogía al cerebro humano que un modelo exacto de él. Los conceptos neuronales son usualmente implementados como simulaciones de software de procesos masivos paralelos que involucran elementos de procesamiento (también llamados neuronas artificiales, o "neurodes") interconectadas en una arquitectura de red.

La neurona artificial recibe entradas análogas a los impulsos electroquímicos que las dendritas de las neuronas biológicas reciben de otras neuronas. La salida de las neuronas artificiales corresponde a las señales enviadas hacia afuera por una neurona biológica a través de su axón. Estas señales artificiales pueden ser cambiadas por medio de pesos de una manera similar al caso de los cambios físicos que ocurren en la sinapsis. Esto se puede observar en la figura III-2.

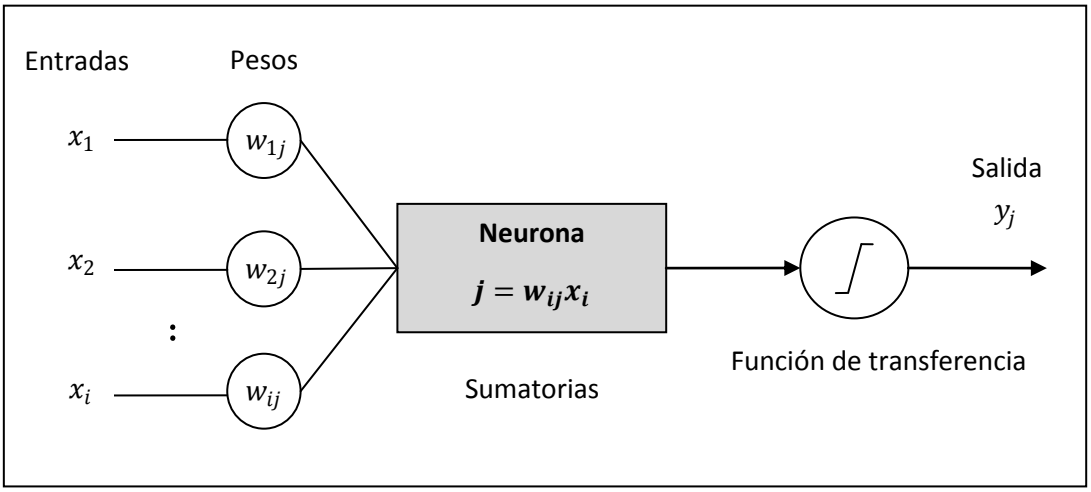


Figura III-2. Procesando información en una neurona artificial [20].

3.2. Historia

La historia de las redes neuronales artificiales está llena de esfuerzos, de varios investigadores de diferentes campos del conocimiento, varios de los cuales se llevaron a cabo durante décadas para desarrollar conceptos que ahora podemos dar por sentados.

Los conceptos y sus desarrollos matemáticos no son suficientes, a menos, que exista la manera de implementar los sistemas. Por ejemplo, las matemáticas necesarias para la reconstrucción de imágenes escaneadas por un tomógrafo asistido por computadora ("CAT") se conocieron varios años antes de que se contara con computadoras de gran capacidad y alta velocidad, que permitieran implementar de manera eficiente los algoritmos que permitirán realizar una implementación práctica.

El enfoque moderno de las redes neuronales comenzó en la década de los 1940's con el trabajo de Warren McCullcoh y Walter Pitts, que mostraron que las redes neuronales artificiales podrían calcular, en principio, cualquier función aritmética o lógica. Su trabajo es a menudo conocido como el origen del campo de conocimiento de las redes neuronales. McCulloch y Pitts fueron seguidos por Donald Hebb, quien propuso que el condicionamiento clásico se presenta debido a las propiedades de neuronas individuales. El propuso un mecanismo de aprendizaje en las neuronas biológicas.

La primer aplicación práctica de las redes neuronales artificiales se dió un poco más tarde en 1950's, con la invención por Frank Rosenblatt de la red perceptron y la regla de aprendizaje asociada. Rosenblatt y sus colegas construyeron una red perceptron y demostraron su habilidad para realizar reconocimiento de patrones. Este primer acontecimiento generó un gran interés en las investigaciones sobre redes neuronales. Desafortunadamente, más tarde se demostró que la red perceptron básica puede resolver solamente problemas con un número limitado de clases.

Al mismo tiempo, Bernard Widrow y Ted Hoff introdujeron un nuevo algoritmo de aprendizaje y lo usaron para entrenar una red neuronal lineal adaptiva, el cual era similar en estructura y capacidad al perceptron de Rosenblatt. La regla de aprendizaje de Widrow-Hoff aún se emplea.

Desafortunadamente, tanto la red de Rosenblatt como la de Widrow sufrieron a causa de las mismas limitaciones, es decir, a causa de no poder realizar una gran cantidad de cálculos rápidamente. Ellos estaban conscientes de esas limitaciones y propusieron nuevas redes que

pudieran sobreponerse a las mismas. De cualquier manera, no fueron capaces de modificar satisfactoriamente sus algoritmos de aprendizaje para entrenar redes más complejas.

Varios investigadores creían que las investigaciones en redes neuronales habían llegado a un punto muerto. Esto, aunado al hecho de que no existían computadoras digitales lo suficientemente poderosas en las cuales experimentar, causó que muchos de ellos abandonaran el campo. A lo largo de una década, las investigaciones sobre redes neuronales permanecieron estancadas.

Algunos trabajos importantes, continuaron durante los 70's. En 1972 Teuvo Kohonen y James Anderson, de manera independiente y separada, desarrollaron nuevas redes neuronales que podían actuar como memorias. Stephen Grossberg también estuvo muy activo durante este periodo en la investigación de las redes auto-organizadas ("self-organizing").

El interés en las redes neuronales había caído a finales de los 1960's debido a la falta de nuevas ideas y computadoras poderosas en las cuales experimentar. Durante los 1980's, ambas limitaciones fueron superadas, y la investigación en las redes neuronales se incrementó dramáticamente. Las nuevas computadoras personales y las estaciones de trabajo, las cuales crecían rápidamente en capacidad, se volvieron ampliamente disponibles. Además, se introdujeron importantes conceptos nuevos.

Dos conceptos nuevos fueron los responsables del resurgimiento de las redes neuronales. El primero fue el uso de métodos estadísticos para explicar el funcionamiento de ciertas clases de redes neuronales recurrentes, las cuales podían ser usadas como memoria asociativa. Esto fue descrito por John Hopfield. El segundo desarrollo clave de los 1980's fue el algoritmo de "backpropagation" (retropropagación) para el entrenamiento de redes perceptron multicapa, el cual fue descubierto de manera independiente por varios investigadores diferentes. La publicación más influyente del algoritmo de "backpropagation" fue realizada por David Rumelhart y James McClellnad.

Estos nuevos desarrollos revigorizaron el campo de las redes neuronales. En los últimos años, se han escrito miles de artículos, y las redes neuronales han encontrado muchas aplicaciones [21].

3.3. Funciones de señales neuronales

La activación interna en las neuronas es transformada en señales con ayuda de una función señal. La función de activación es empleada en neuronas para calcular la respuesta de salida de la neurona con la entrada pesada. La respuesta de salida es producida por medio de la suma de las señales de entrada pesadas aplicadas a la activación. Algunas de las funciones de señales se mencionan a continuación:

3.3.1. Función lineal

La función está dada por

$$f(t) = t\ para\ toda\ t$$

Esta es la función más simple. Esta función no está limitada ($t \in \{-\infty, \infty\}$). Esta función es no acotada.

3.3.2. Función binaria

La función de umbralización binaria está dada por

$$f(t) = \begin{Bmatrix} 1 & t \geq 0 \\ 0 & t < 0 \end{Bmatrix}$$

Estas funciones no son diferenciables y la función señal $f(t) \in \{0,1\}$.

3.3.3. Función umbral bipolar

La función está dada por

$$f(t) = \begin{Bmatrix} +1 & t \geq 0 \\ -1 & t < 0 \end{Bmatrix}$$

Para sobreponerse al problema de discontinuidad se introduce la umbralización bipolar. Esta define que $f(t) = 0$ es una activación ambigua y no se puede traducir

satisfactoriamente. Ahora la señal es $f(t) \in \{-1,1\}$ en lugar de $\{0,1\}$ y el comportamiento lógico de una neurona se extiende a bipolar. Esta función también se conoce como función signum, se denota sign(x).

3.3.4. Función de umbralización lineal.

La función está dada por

$$f(t) = \begin{cases} 0 & t \leq 0 \\ \alpha t & 0 \leq t \leq t_m \\ 1 & t \geq t_m \end{cases}$$

Donde $\alpha = 1/t_m$ es un parámetro pendiente de función y la función es diferenciable gracias al parámetro. La función lineal no está acotada. Esta es la versión acotada de la función lineal.

3.3.5. Función señal sigmoide

La función está dada por

$$f(t) = 1/\left(1 + e^{\lambda t}\right)$$

Donde λ es un factor escalar de ganancia. Esta función es diferenciable y $f(t) \in \{0,1\}$. esta función es monotónica, lo que significa que los valores se incrementan o decrecen en función del valor de entrada. La función sigmoide se grafica usualmente como una curva y se usa comúnmente en redes multicapa.

3.3.6. Función tangente hiperbólica

La función está dada por

$$f(t) = k tanh(\lambda t)$$

Donde k es un valor constante mayor que 0 $(k > 0)$, y λ es un factor pendiente. Se trata también de una función diferenciable. Esta función es un ejemplo de una función bipolar donde $f(t) \in \{-1,1\}$.

3.3.7. Función gaussiana

La función está dada por

$$f(t) = exp(-(t-c)^2/2\sigma^2)$$

Donde σ es el factor de distribución gaussiana y c es el centro. En la función básica radial ("Radial Basis") la función gaussiana es empleada porque se trata de una red neuronal base central. Es un ejemplo de una función no-monotónica, la cual se incrementa primero a partir de 0 y después decrece hacia 0, por ejemplo $f(t) \in \{0,1\}$.

3.3.8. Función estocástica

La función está dada

$$f(t) = \begin{Bmatrix} +1 & con\ probabilidad & P(t) \\ -1 & con\ probabilidad & 1-p(t) \end{Bmatrix}$$

Esta función P(x) es la función de activación probabilística. Las neuronas estocásticas se usan en las redes de aprendizaje reforzado y los métodos Boltzmann. Para ambos casos de las neuronas estocásticas binario o bipolar, la función señal es $f(t) \in \{0,1\}$ o $\{-1,1\}$. Esta función neuronal es no-diferenciable.

3.4. Arquitectura de las redes neuronales

Las redes neuronales conectan a las neuronas artificiales básicamente con dos diferentes tipos de arquitecturas: "feed-forward" (alimentación hacia adelante o positiva) y "feed-back" (alimentación hacia atrás o negativa). En la figura III-3 se observan algunas de las familias de redes más comunes [22].

3.4.1. Arquitectura "feed-forward"

La arquitectura "feed-forward" de las redes es estática por naturaleza donde la entrada esta directamente conectada a la salida o a las neuronas de la capa oculta. No hay un ciclo en la red porque la arquitectura es hacia adelante por naturaleza.

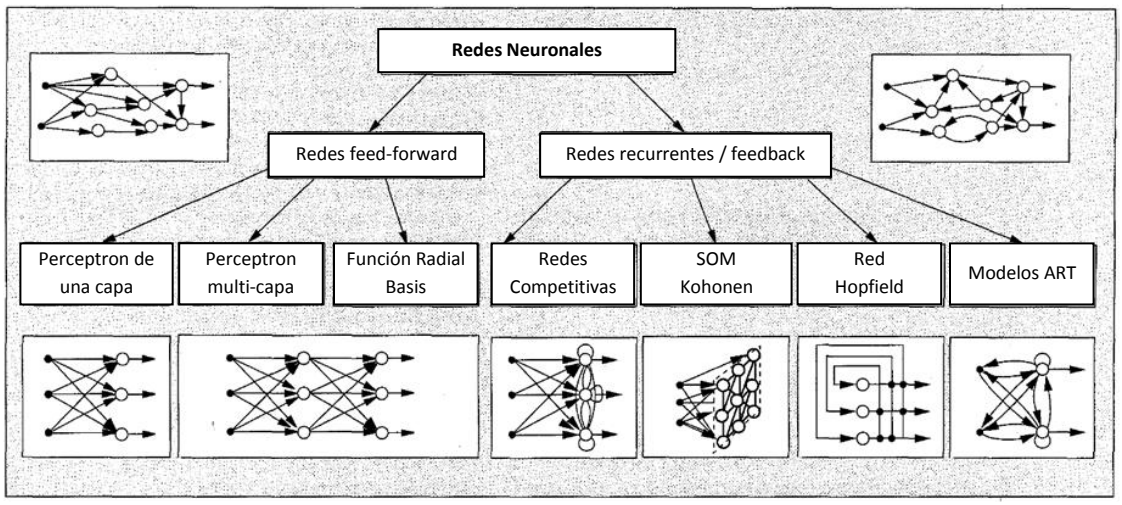


Figura III-3. Taxonomía de arquitecturas de redes "feed-forward" y recurrentes/"feed-back" [22].

3.4.2. Arquitectura "feed-back"

Las redes "feed-back" son lo contrario a las "feed-forward". Estas redes tienen ciclos en la red debido a que la neurona transporta una señal hacia atrás a través de la red. La red recurrente es también un tipo de arquitectura "feed-back".

3.5. Modelos de redes neuronales

De acuerdo con la arquitectura de las redes neuronales existen diferentes tipos de modelos de redes neuronales, los cuales han sido propuestos a lo largo de los años y sus áreas de aplicación se han ido incrementando. Algunos de estos modelos son:

3.5.1. Perceptron

El perceptron fue desarrollado por Frank Rosenblatt en 1958. El perceptron es un modelo de red neuronal "feed-forward" de un solo nodo el cual emplea un umbral binario como función de activación. La raíz del modelo es el modelo de McCulloch Pitts. Este fue el primer modelo de red neuronal. El perceptron trabaja en una aproximación de aprendizaje supervisado por corrección de error. En el perceptron la red de entrenamiento continúa hasta que el error sea cero. El modelo de red neuronal perceptron resulta bueno para clasificación, con la restricción de que las clases deben ser linealmente separables. Algunas

de sus aplicaciones se encuentran en el campo del reconocimiento, como el reconocimiento de espectros con ruido, el reconocimiento de caracteres escritos y muchas más.

3.5.2. Adaline

La red neuronal Adaline fue dearrollada por Widrow y Hoff en 1960's. Adaline es la abreviatura de "Adaptive Linear Neuron". El comportamiento adaptivo es un tipo de comportamiento que se ajusta a otro tipo de situación. La neurona es adaptiva cuando sus pesos pueden cambiar de acuerdo con una regla de entrenamiento bien definida [19]. Widrow llamó a esta neurona un elemento lineal adaptable. El modelo de red neuronal Adaline tiene una arquitectura de un solo nodo "feed-forward", que trabaja con una función de activación lineal. Adaline trabaja con el algoritmo de aprendizaje supervisado, con la regla del mínimo medio cuadrado o la regla delta. La aplicación del modelo Adaline se da en el dominio de la regresión, debido a que las salidas pueden tomar valores reales sin restricción, a diferencia del perceptron. De manera similar resulta benéfico en el campo de las señales adaptivas con el algoritmo LMS.

3.5.3. Perceptron multicapa

Esta red neuronal fue desarrollada por Rumelhart, Hinton y Williams en 1986. El perceptron multicapa es un modelo de red feed-forward con una arquitectura de capa. Esta red utiliza la función sigmoide no lineal como función de activación en la capa oculta. La no-linealidad es diferenciable en cualquier lugar, por lo que está suavizada y no está restringida como el perceptron de un solo nodo. El modelo perceptron multicapa es una aproximación de aprendizaje supervisado el cual emplea la regla delta. Es entrenado empleando el algoritmo de "back-propagation" el cual es muy rápido y eficiente. Los dominios de aplicación para el perceptron multicapa son, entre otras: reconocimiento de emociones, reconocimiento de imágenes y texto, reconocimiento de caracteres.

3.5.4. Aprendizaje por refuerzo

Este aprendizaje fue desarrollado por Sutton y Barto en 1998. Los algoritmos de aprendizaje por refuerzo se basan en el principio de reforzamiento. Si una acción de un

sistema es seguida por una respuesta satisfactoria entonces se fortalece la tendencia de producir dicha acción. La red de aprendizaje por refuerzo se basa en esta aproximación y evalúa la señal de éxito o fracaso la cual es proporcionada por el verificador. El propósito general de este aprendizaje es maximizar la tasa de éxito de la red neuronal. Este aprendizaje es un tipo de aprendizaje supervisado el cual emplea una función de activación de umbral binario. El aprendizaje por refuerzo ha probado ser una herramienta computacional práctica y se emplea en el área de control. Las redes de aprendizaje por refuerzo pueden resolver el problema de reconocimiento de objetos, reconocimiento de patrones en 3D, etc.

3.5.5. "Support vector machine"

"SVM" fue desarrollada por Vapnik y Corinna Cortes en 1990's. Esta red neuronal está basada en una arquitectura de red "feed-forward" multicapa. Emplea el umbral binario como función de activación. Es una aproximación de aprendizaje supervisado y una arquitectura basada en un "kernel". A pesar de que en un principio se empleó para trabajar en el reconocimiento de caracteres ópticos y en clasificadores, dado el buen desempeño en el campo del reconocimiento se ha aplicado en las tareas de reconocimiento de objetos. Así mismo, se emplea satisfactoriamente el campo de la meteorología.

3.5.6. Función "radial basis"

En 1998, las redes función "radial basis" ("RBFN's") fueron desarrolladas por Broom Head y Lowe. Se trata de una red neuronal "feed-forward" la cual esta encapsulada dentro de una red neuronal de dos capas que calcula la activación en la capa oculta. Se calcula empleando un exponencial de la medida de la distancia entre el vector de entrada y el vector prototipo en las neuronas de la capa oculta que caracterizan a la función señal. En la interpolación de los puntos dato en un entrenamiento finito se introduce un conjunto de "RBFN's". Puede ser usada tanto para interpolación como para función de aproximación. En los problemas de optimización donde hay un gran número de posibles soluciones para un pequeño problema, la red Hopfield ha encontrado aplicación. La función Radial Basis es una red neuronal "feed-forward" pero es ligeramente distinta a las redes neuronales estándar. De manera general se añade un peso de "bias" en la salida de la neurona lineal. El aprendizaje en

"RBFN" puede ser implementado de diversas formas como: aprendizaje supervisado, selección de centros aleatorios, selección de centros basado en agrupamiento. Se usa en problemas de clasificación, regresión, interpolación, reconocimiento de caracteres y reconocimiento de patrones.

3.5.7. Red Hopfield

En el año 1982, John Hopfield propuso un modelo de red neuronal el cual se conoce como red Hopfield. Se trata una red recurrente de una sola capa. Donde, todas las neuronas en la red son realimentadas a partir de las otras neuronas presentes. La red Hopfield emplea un umbral binario/sigmoide como función de activación y tiene aplicaciones interesantes en el reconocimiento de objetos en 3D.

3.5.8. Maquina Boltzmann

En 1985, David H. Ackey, Hinton y T.J. Sejnowski propusieron un modelo al cual llamaron máquina de Boltzmann. La máquina de Boltzmann es un modelo de red neuronal que emplea métodos estocásticos en su funcionamiento. El aprendizaje es implementado con la combinación compleja de un método de búsqueda estocástico llamado "simulated annealing" y "gradient descent". De manera que, el aprendizaje Boltzmann es un aprendizaje supervisado estocástico. Este modelo de red neuronal tiene una arquitectura de dos capas con una unidad de umbralización binaria. Este tipo de arquitectura es empleada en técnicas de optimización, en la cuales las neuronas estocásticas cambian sus estados dependiendo de una probabilidad, que es calculada a partir del cambio de energía, mismo que resulta al conmutar la neurona su señal. La distribución de probabilidad en la red se produce al utilizar un procedimiento basado en la correlación. La máquina de Boltzmann se emplea para el reconocimiento de patrones.

3.5.9. Memoria asociativa bidireccional

La memoria asociativa bidireccional ("BAM") fue desarrollada por Bart Kosko en 1988. Es una extensión de las redes Hopfield a un resultado de arquitectura de dos campos. En la memoria asociativa bidireccional los campos de los algoritmos se actualizan uno a la vez.

La capacidad de corrección de error de la "BAM" es mucho mayor en comparación con la red Hopfield debido a su arquitectura de dos capas. Emplea un umbral de activación binario como unidad de activación en la red. Se aplica en el campo de la autentificación, así como, en el reconocimiento de caracteres.

3.5.10. Teoría de resonancia adaptiva

"Adaptive Resonance Theory, ART" fue desarrollada por Grosberg y Carpenter de 1987 a 1990. Se han propuesto diferentes tipos de ART's. La arquitectura de las ART's es un sistema de red neuronal de dos capas que se logra por medio del uso de instars y outstars en una red bidireccional. Unas outstars que disminuyen la información espacial de los patrones en vectores pesados, cuando las neuronas son activadas en la red. Por otro lado las instars varían su vector de pesos para ajustar el vector de entrada durante el entrenamiento. Las ART usan la activación "faster than linear" resultado del comportamiento ganador toma todo ("winner takes all"). En "Adaptive Resonance Theory" esta función desempeña una gran competencia entre las neuronas de la red, en los campos de aplicación como son clasificación y agrupamiento. La aplicación de ART1 es la clasificación de parámetros binarios. Este modelo se extiende a ART2, el cual trabaja con parámetros de entrada analógicos. Y está también ART3 para sofisticados neurotransmisores, basados en la búsqueda paralela de códigos de reconocimiento distribuidos en redes multicapa. Los módulos ART tienen sus aplicaciones en varios campos, tales como robots móviles, reconocimiento de objetivos, reconocimiento de rostros, reconocimiento visual de objetos 3D, reconocimiento de caracteres y símbolos, etc.

3.5.11. Cuantización vectorial, "VQ"

Teuvo Kohonen desarrolló un modelo de red neuronal el cual se conoce como cuantización vectorial. La cual es una red neuronal "feed-back" de una sola capa. Básicamente se introdujo para comprimir información, almacenar y transmitir voz o datos. Utiliza una aproximación por aprendizaje competitivo. El campo de aplicación de la cuantización vectorial está en el agrupamiento ("clustering"). La cuantización vectorial trabaja tanto en la aproximación por aprendizaje competitivo supervisado como no supervisado. Kohonen introdujo la versión supervisada de la cuantización vectorial, conocida como "Learning

Vector Quantization, LVQ". Utiliza la función de activación "faster tan linear". No hay ninguna red que trabaje con competición suave, esta competición fuerte, la cual emplea "LVQ", reduce la posibilidad de errores en la clasificación comparada con la cuantización vectorial no supervisada. El agrupamiento y la cuantización tienen un área de aplicación para la cuantización vectorial, misma, que también se emplea en el campo del reconocimiento, por ejemplo, en el reconocimiento de caracteres.

3.5.12. Red mexican hat

La red neuronal mexican hat se basa en la competencia. Esta red se sobrepone la aproximación fuerte con la aproximación suave; donde en la red un grupo de neuronas, las cuales se encuentran alrededor de la neurona ganadora, se convierten en la ganadora. En esta red no hay algoritmos de aprendizaje y los pesos empleados son fijos. Esta red usa un umbral lineal como unidad de activación. Cuenta con una arquitectura "feed-back" de una sola capa y emplea una red compleja de conectividad que involucra excitaciones "feed-back" de rango corto y de excitación de bajo nivel de rango largo. Esta red básicamente se utiliza en actividades de agrupamiento, promediando el nivel de un grupo donde las neuronas vecinas de la entrada máxima, encienden y todas las demás neuronas se desactivan, y en lugar de que gane una neurona es un grupo ("cluster") el que lo hace.

3.5.13. Mapas de características de organización automática de Kohonen, "SOFM"

Los "Self Organizing Features Maps, SOFM" fueron desarrollados por Teuvo Kohonen en 1972. Es una red neuronal de una sola capa la cual es capaz de reproducir aspectos importantes de la arquitectura del sistema de neuronas humano. Se utilizan los mapas de topología para representar los datos y los "SOFM" conservan información topológica importante, la cual se puede obtener por un proceso de aprendizaje no supervisado. Los algoritmos "SOFM" son una cuantización vectorial competitiva, en la cual los grupos de neuronas ganan, en lugar de una sola neurona. Esta propiedad de estructura topológica solo se encuentra en los "Self Organizing Maps, SOM". El propósito original para desarrollar esta red neuronal fue el campo del análisis de señales, clasificación de patrones y agrupación de datos.

3.6. Primeras etapas de las aplicaciones de las RNA al reconocimiento de voz

Los reconocedores de voz convencionales basados en "HMM's" han mostrado ser de utilidad bajo ciertas condiciones limitadas, por ejemplo, el caso de la dependencia del locutor, reconocimiento de grandes vocabularios de pequeñas secuencias de palabras conectadas y el caso de la independencia del locutor, reconocimiento de pequeños vocabularios de palabras aisladas.

Actualmente, estos sistemas son parte de una nueva oportunidad de realizar desarrollos haciendo uso de las tecnologías de procesamiento de voz. Sin embargo, también se ha resaltado, que estos sistemas son incapaces de llevar a cabo de una manera satisfactoria algunas de las tareas que permitan a las personas hablarle a los sistemas, de la misma manera en que se realiza la comunicación humano-humano por medio de la voz.

La razón más seria, para explicar estas fallas, ha sido la limitada precisión en la clasificación de los sistemas convencionales. Para lidiar con este problema, algunas investigaciones se han centrado en desarrollar metodologías alternativas, por ejemplo, las redes neuronales, "ANN", con un énfasis en su alta capacidad discriminante. La mayoría de los reconocedores basados en "ANN", tal como, las "TDNN" y las hibridas "ANN/HMM". En la tabla III-1 se mencionan algunos de los reconocedores basados en "ANN".

Tabla III-1. Ejemplos típicos de reconocedores de voz basados en "ANN" [23].

"ANN" empleada para el reconocimiento
"Time delay neural network (TDNN)"
"Multi-state TDNN"
"Learning vector quantization (LVQ) network"
"Shift-tolerant LVQ network"
"HMM-based time warping/ANN hybrid"
"ANN-based probability estimator/HMM hybrid"

En las primeras aplicaciones, las "ANN" se utilizaron para incrementar la precisión en el reconocimiento de pequeños segmentos, tal como, fonemas. La "TDNN" y la "STLVQ" son ejemplos típicos de este tipo de aplicaciones de "ANN". Se incorporó en la arquitectura de la red una estructura de retraso en el tiempo, buscando normalizar las variaciones temporales de pequeños segmentos. El efecto de estos reconocedores basados en "ANN" en el reconocimiento de fonemas fue bastante bueno. Sin embargo, no fueron suficientes para reconocer segmentos más largos y comunes de voz, como palabras o frases, porque la estructura de retraso en el tiempo era un simple mecanismo de normalización para segmentos de voz comparativamente pequeños (y estacionarios), los cuales usualmente corresponden a fonemas. Los reconocedores basados en "ANN" no pueden manejar las dinámicas de segmentos más largos como son las palabras. En la siguiente etapa, los reconocedores tal como la "TDNN" fueron usadas de manera hibrida con "DTW" y "HMM's", los cuales tienen una arquitectura ajustada para modelar las características no estacionarias de los segmentos largos de voz.

Las "ANN" han contribuido de manera satisfactoria mejorando los índices de reconocimiento de algunos sistemas de reconocimiento en varias pruebas experimentales. Sin embargo, el uso de este tipo de aplicaciones de "ANN" se ha limitado al reconocimiento de señales de voz que han sido extraídas a priori. Obviamente, aún existen varios detalles por resolver en dichas técnicas, y varios retos en los desarrollos basados en "ANN" [23].

3.7. Técnicas adicionales de redes neuronales

3.7.1. Código de predicción lineal

Una gran parte de las aplicaciones relacionadas con el procesamiento de la voz están basadas en el análisis "LPC", dado que es capaz de extraer la información característica de la porción de voz analizada. La predicción lineal modela el rango de voz humana como una respuesta al impulso infinita, capaz de generar la señal de voz.

El término de predicción lineal se refiere al método para predecir o aproximar una muestra de una señal dada en el dominio del tiempo $s[n]$, basada en varias muestras anteriores $s[n-1], s[n-2], s[n-M]$.

$$s[n] \approx \tilde{s}[n] = -\sum_{i=1}^{M} a_i s[n-i]$$

Donde $s[n]$ es llamada señal muestreada y $a_p, i = 1,2, \dots M$ son los predictores o coeficientes "LPC". Un pequeño número de coeficientes LPC $a_1, a_2, \dots, a_M$ pueden ser usados para representar eficientemente una señal $s[n]$. Los valores $a_1, a_2, \dots, a_M$ son la base para la realización de este trabajo debido a que nos ayudan a modelar los parámetros de la voz de cada uno de los hablantes que se emplean en el sistema propuesto.

3.7.2. La red Backpropagation

En 1986, Rumelhart, Hinton y Williams, basados en otros trabajos, propusieron un método para que una red neuronal aprendiera la asociación que existe entre los patrones de entrada a la misma y las clases correspondientes, utilizando mas niveles de neuronas que los que utilizó Rosenblatt para desarrollar el Perceptron. Este método es conocido como "Backpropagation" (retropropagación) que es un tipo de red de aprendizaje supervisado, el cual emplea un ciclo propagación-adaptación de dos fases.

Una vez aplicado un patrón de entrenamiento a la entrada de la red, este se propaga desde la primera capa a través de las capas subsecuentes de la red, hasta generar una salida, la cual es comparada con la salida deseada y se calcula una señal de error para cada una de las salidas, a su vez esta es propagada hacia atrás, empezando en la capa de salida, hacia todas las capas de la red hasta llegar a la capa de entrada, con la finalidad de actualizar los pesos de conexión de cada neurona, para hacer que la red converja a un estado que le permita clasificar correctamente todos los patrones de entrenamiento. La estructura general se muestra en la figura III-4.

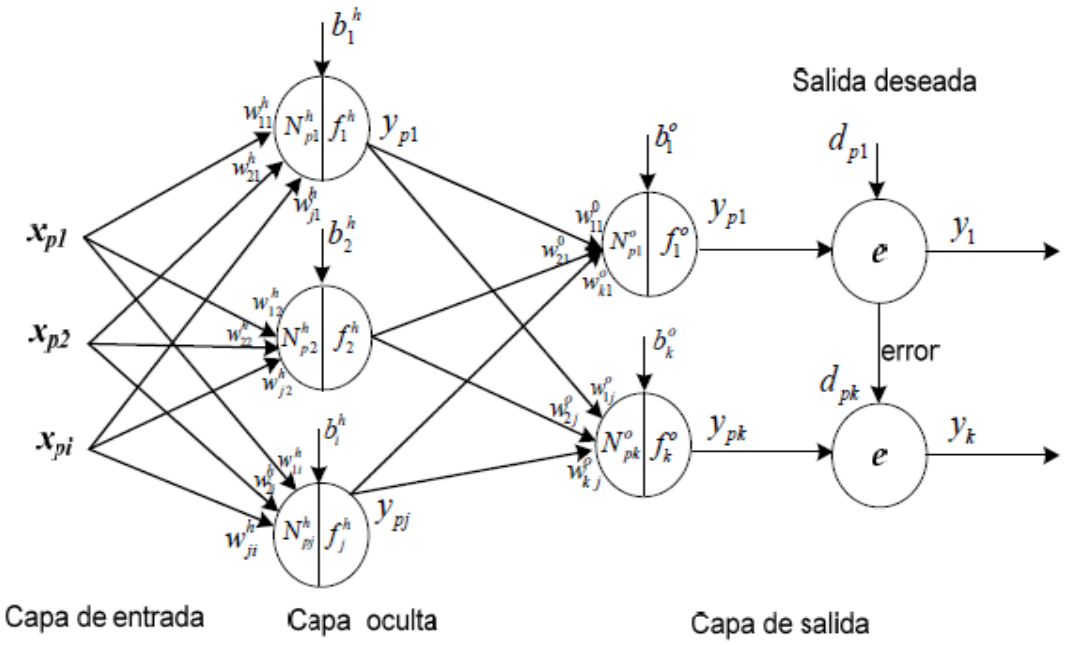


Figura III-4. Modelo de retropropagación "ANN" [24].

3.7.3. Algoritmo de entrenamiento de la red

A continuación se presenta un algoritmo que permite el entrenamiento de retropropagación "ANN" [24].

1. Inicializar los pesos de la red (w) con valores aleatorios pequeños.
2. Mientras la condición de paro sea falsa realizar los pasos (3-6).
3. Se presenta un patrón de entrada, $(x_{p1},\ x_{p2}, \dots,\ x_{pi})$ y se especifica la salida deseada que debe generar la red $(d_{p1},\ d_{p2}, \dots,\ d_{pk})$.
4. Se calcula la salida actual de la red, para ello se presentan las entradas a la red y se va calculando la salida que presenta cada capa hasta llegar a la capa de salida $(y_1, y_2, \dots, y_k)$. Los pasos son los siguientes:
   a) Se determinan las entradas netas para las neuronas ocultas procedentes de las neuronas de entrada.

$$N_{pj}^{h} = \sum_{i=1}^{m} w_{ji}^{h} x_{pi} + b_{i}^{h}$$

b) Se aplica la función de activación a cada una de las entradas de la neurona oculta para obtener su respectiva salida.

$$y_{pj} = f_j^h\left(N_{pj}^h = \sum_{i=1}^{m} w_{ji}^h x_{pi} + b_i^h\right)$$

c) Se realizan los mismos cálculos para obtener las respectivas salidas de las neuronas de la capa de salida.

$$N_{pk}^o = \sum_{j=1}^{m} w_{kj}^o y_{pj} + b_k^o$$

$$y_{pk} = f_k^o\left(N_{pk}^o = \sum_{j=1}^{m} w_{kj}^o y_{pj} + b_k^o\right)$$

5. Se determinan los términos de error para todas las neuronas

a) Cálculo de error (salida deseada-salida obtenida)

$$e = \left(d_{pk} - y_{pk}\right)$$

b) Obtención de la delta (producto del error con la derivada de la función de activación, con respecto a los pesos de la red).

$$\delta_{pk}^o = e^* f_k^{o'}\left(N_{pk}^o\right)$$

6. Actualización de los pesos. Se emplea el algoritmo recursivo del gradiente descendente, comenzando por las neuronas de salida y trabajando hacia atrás hasta llegar a la capa de entrada.

a) Para los pesos de las neuronas de la capa de salida:

$$w_{kj}^o(t+1) = w_{kj}^o(t) + \Delta w_{kj}^o(t+1)$$

$$\Delta w_{kj}^o(t+1) = miu\delta_{pk}^o y_{pj}$$

b) Para los pesos de las neuronas de la capa oculta:

$$w_{ji}^h(t+1) = w_{ji}^h(t) + \Delta w_{ji}^h(t+1)$$

$$\Delta w_{ji}^{h}(t+1) = miu\delta_{pj}^{h}\, x_{pi}$$

7. Se cumple la condición de paro (error mínimo ó número de iteraciones máximo alcanzado)

3.8. Redes neuronales para el reconocimiento de palabras aisladas.

Después de haber implementado el sistema Sphinx para el reconocimiento de palabras continuas y haber mejorado el corpus de entrenamiento, aumentando el número de repeticiones del las palabras contenidas en dicho corpus, los resultados que se obtuvieron, si bien, mostraron una mejoría en el desempeño del sistema con que contábamos, distan aún de ser un resultado digno de un buen reconocedor. Por tal motivo nos dimos a la tarea de proponer un método alternativo y/o adicional al sistema Sphinx.

En principio la hipótesis que se planteó consistió en contar con un corpus de palabras fonéticamente confusas, mismas que, al ser reportadas como palabras reconocidas por el sistema Sphinx nos alertan de un posible error en el reconocimiento. De manera que, al detectarse una de estas palabras se pretende realizar de nueva cuenta el reconocimiento, pero solamente para el segmento de voz que contiene la palabra en cuestión; con lo que se espera que, al realizar una verificación del resultado otorgado por el sistema Sphinx, disminuyamos aún más la tasa de error resultante de los experimentos realizados únicamente con el corpus mejorado.

Se realizaron pruebas con dos grupos de palabras, el primero compuesto por palabras fonéticamente distintas y el segundo por palabras fonéticamente confusas. Para las palabras fonéticamente distintas se seleccionaron palabras de manera aleatoria pero que su uso es común en el español hablado en México, mientras que, para las palabras fonéticamente confusas se tomó como base los trabajos previamente realizados en el laboratorio de procesamiento de voz de la FI UNAM. El resultado de estas pruebas se observa en el capítulo 4.

IV. Pruebas y análisis de resultados

4.1. Experimentos de reconocimiento de palabras aisladas con redes neuronales

Las pruebas de reconocimiento de voz con redes neuronales se basan en el siguiente procedimiento:

Partimos de definir dos grupos de palabras, las que se conocen como corpus, un primer grupo conformado por palabras que se pueden considerar como fonéticamente distintas y otro grupo, el cual basándonos en desarrollos previos, consideramos como fonéticamente confusas.

Posteriormente, se realizaron las grabaciones de las palabras y luego se realizó un acondicionamiento de las mismas para que fuese posible usarlas en el sistema de reconocimiento.

El sistema de reconocimiento hace uso de los corpus conformados por las grabaciones de las palabras, para obtener los parámetros característicos de dichas señales de voz, en este caso "LPC's", para después usar algunos de ellos para realizar el entrenamiento del sistema y reservar los restantes para las pruebas de reconocimiento.

4.2. Palabras que conforman los corpus de pruebas

Para la realización de las pruebas de reconocimiento de voz empleando redes neuronales se seleccionaron las palabras que conformaron los corpus de prueba. Como ya se mencionó, se consideraron dos corpus, uno con palabras fonéticamente distintas y otro con palabras fonéticamente confusas. Las palabras que los conforman se muestran en las tablas IV-1y IV-2, respectivamente.

Es importante recalcar que al seleccionar las palabras que se emplearon para el corpus de palabras fonéticamente confusas se consideraron los desarrollos previos realizados en el laboratorio de procesamiento de voz de la FI UNAM [25].

Tabla IV-1. Corpus con palabras fonéticamente distintas.

Palabras Fonéticamente Distintas	
1. Casa	9. Pera
2. Silla	10. Bote
3. Reja	11. Vaca
4. Oso	12. Mesa
5. Mano	13. Copa
6. Sala	14. Dado
7. Goma	15. Llave
8. Auto	16. Lata

Tabla IV-2. Corpus con palabras fonéticamente confusas.

Palabras Fonéticamente Confusas	
1. Acto	9. Alas
2. Actos	10. Bajas
3. Actor	11. Bala
4. Pacto	12. Blas
5. Tacto	13. Bolas
6. Auto	14. Galas
7. Alto	15. Malas
8. Balas	16. Rama

4.3. Grabación de las señales de voz

Para poder realizar los experimentos de una mejor manera y además siendo congruentes con el procedimiento que se empleó durante la realización de las grabaciones para mejorar el corpus empleado paras pruebas con el sistema Sphinx 3, mismo que se detalló en el capítulo 2, es necesario que las grabaciones cumplan con ciertas especificaciones en cuanto a formato. Para obtener las grabaciones con dichas especificaciones se realizaron las grabaciones de las señales de voz, en este caso palabras, auxiliándonos del software Adobe Audition, mismo que nos permite ajustar varios parámetros para las grabaciones. Este procedimiento se describe a continuación:

1. Para realizar la grabación de las palabras se empleó el software Adobe Audition, utilizando la siguiente configuración para realizar las grabaciones:
   - Archivos de audio en formato .wav
   - Frecuencia de muestreo de 44100 [Hz]

- 16 bits
- Mono

2. Se grabaron las palabras de los corpus de las tablas IV-1 y IV-2.
3. Las grabaciones se realizaron en una habitación común y corriente, cuidando en la medida de lo posible que los niveles de ruido ambiente fueran los mínimos.

4.3.1. Hardware utilizado para las grabaciones

- Micrófono con conexión USB marca Logitech.
- Pedestal para micrófono (con el fin de reducir las vibraciones).
- Computadora portátil, con tarjeta de sonido integrada a la tarjeta madre.

4.3.2. Procesamiento posterior de las grabaciones

Las grabaciones realizadas con estas especificaciones las consideramos como señales base, posteriormente se realizaron adecuaciones a las mismas para utilizarlas en el sistema de reconocimiento. Estas adecuaciones se realizaron en función de los parámetros que tiene como requerimiento el sistema Sphinx 3, cabe mencionar, que no se realizaron las grabaciones de las palabras directamente con las especificaciones marcadas en los requerimientos del sistema Sphinx 3 ya que estas son menores en calidad a las grabaciones obtenidas, de modo que, las grabaciones obtenidas al ser de mayor calidad se pueden utilizar en desarrollos posteriores en el laboratorio de procesamiento de voz de la FI UNAM.

Las adecuaciones que se realizaron en las señales de voz consistieron en modificar la frecuencia de muestreo mediante un sub-muestreo de las señales, delimitar las señales de voz únicamente a la porción que contiene la voz, así como también, realizar una homogeneización de los niveles de la señales de voz, mediante un ajuste de ganancia en aquellas que lo requirieran.

Las especificaciones de las señales reacondicionadas se enlistan a continuación:

- Archivos de audio en formato .wav
- Frecuencia de muestreo de 16000 [Hz]

- 16 bits
- Mono
- Amplitud de alrededor de -3[dB] en el Adobe Audition

4.4. Pruebas de reconocimiento con redes neuronales.

Para observar el desempeño de las redes neuronales en el reconocimiento de voz se plantearon diversas pruebas, mismas que se mencionan a continuación:

Un primer grupo de pruebas de reconocimiento de voz empleando "VQ", las cuales se tomaron como referencia o "baseline".

Un segundo grupo de pruebas, ya con redes neuronales, con la finalidad de determinar que funciones emplearíamos en las pruebas subsiguientes.

Un tercer grupo de pruebas, ya solamente con dos funciones de entrenamiento, traingda y trainscg, y en donde se variaron una serie de parámetros como: el número de capas, el número de segmentos y las repeticiones de prueba.

Un cuarto grupo de pruebas, adicionales, empleando "Adaptive Resonance Theory, ART"

Y, finalmente, una serie de pruebas de reconocimiento de las vocales con redes neuronales.

4.5. Pruebas de reconocimiento con "VQ"

Para poder comparar el desempeño del reconocimiento con redes neuronales, consideramos como punto de partida ("baseline") una serie de pruebas realizadas haciendo uso del reconocimiento con "VQ", técnica que empleé en el desarrollo de mi tesis de licenciatura, Sistema Automatizado de Iluminación de una casa mediante comandos de voz [26]. De manera que las primeras pruebas reportan los resultados obtenidos al utilizar esta técnica y posteriormente se muestran las pruebas realizadas haciendo uso de las redes neuronales.

Se consideraron 4 segmentos en cada palabra para luego obtener los vectores característicos de dichos segmentos, salvo en el caso que se especifique otra cosa. De manera que se obtuvieron 16 centroides de orden 8 por cada uno de los segmentos, es decir 64 en total por cada palabra.

4.5.1. Reconocimiento de palabras con "VQ" (Pruebas con las repeticiones del entrenamiento)

Para la realización de estas pruebas se consideró el corpus de las palabras fonéticamente distintas y el corpus de las palabras fonéticamente confusas, respectivamente, se emplearon 10 repeticiones de cada palabra para el entrenamiento y posteriormente se observó el resultado del sistema usando cada una de ellas como entrada, es decir que, se probaron 160 palabras de cada corpus, de las cuales se obtuvo un reconocimiento del 100% en ambos casos, como se observa en la tabla IV-3 y IV-4, respectivamente.

Tabla IV-3. Reconocimiento de palabras fonéticamente distintas usando "VQ".

Total de Palabras Reconocidas	160
% de Palabras Reconocidas	100
% de Error	0

Tabla IV-4. Reconocimiento de palabras fonéticamente confusas usando "VQ".

Total de Palabras Reconocidas	160
% de Palabras Reconocidas	100
% de Error	0

4.5.2. Reconocimiento de palabras con "VQ" (Pruebas con repeticiones diferentes a las del entrenamiento).

Para la realización de estas pruebas se consideró el corpus de las palabras fonéticamente distintas y el corpus de las palabras fonéticamente confusas, respectivamente, se emplearon 10 repeticiones de cada palabra para realizar el entrenamiento y posteriormente se

realizaron las pruebas con 5 repeticiones, diferentes a las usadas en el entrenamiento, de cada una de ellas. Se obtuvo un resultado de 96.25 % de palabras reconocidas con éxito para el caso del corpus de las palabras fonéticamente distintas y un 82.5% de palabras reconocidas con éxito para el corpus de las palabras fonéticamente confusas, lo cual se puede ver en la tabla IV-5 y IV-6, respectivamente.

Tabla IV-5. Reconocimiento de palabras fonéticamente distintas con "VQ". (Prueba con palabras diferentes a las del entrenamiento).

Total de Palabras Reconocidas	77
% de Palabras Reconocidas	96.25
% de Error	3.75

Tabla IV-6. Reconocimiento de palabras fonéticamente confusas con "VQ". (Prueba con palabras diferentes a las del entrenamiento).

Total de Palabras Reconocidas	66
% de Palabras Reconocidas	82.5
% de Error	17.5

4.6. Pruebas de reconocimiento de palabras con redes neuronales, empleando diferentes funciones para el entrenamiento.

Tomando como base los resultados obtenidos con "VQ" se procedió a realizar las pruebas de reconocimiento haciendo uso de las redes neuronales. Se consideraron diversas funciones de MATLAB para el entrenamiento de la red neuronal: "Gradient descent with adaptive learning rule backpropagation" (traingda), "Levenberg-Marquardt backpropagation" (trainlm), "Gradient descent with momentum and adaptive learning rule backpropagation" (traingdx) y "Scaled conjugate gradient backpropagation" (trainscg).

En el caso del reconocimiento con redes neuronales los vectores característicos, previamente calculados, son reordenados para que sea posible su utilización.

4.6.1. Reconocimiento de palabras fonéticamente distintas con redes neuronales

Para la realización de estas pruebas se consideró el corpus de las palabras fonéticamente distintas y el corpus de las palabras fonéticamente confusas, respectivamente, se emplearon 10 repeticiones de cada palabra para el entrenamiento y se usaron otras 5 repeticiones para realizar las pruebas. Se realizó 3 veces cada una de las pruebas y se consideró una red neuronal con dos capas, 45 neuronas en la primera capa y 16 en la segunda capa, que a su vez resulta ser la capa de salida. El desempeño del reconocimiento con redes neuronales empleando diferentes funciones para el entrenamiento de la red se observa en la tabla IV-7 y IV-8, respectivamente.

Tabla IV-7. Reconocimiento de palabras fonéticamente distintas con redes neuronales.

Prueba	traingda	trainlm	traingdx	trainscg
1	55	30	54	54
2	51	14	51	55
3	57	6	55	51
Total	**163**	**50**	**160**	**160**
%	**67.91**	**20.83**	**66.67**	**66.67**

Tabla IV-8. Reconocimiento de palabras fonéticamente confusas con redes neuronales.

Prueba	traingda	trainlm	traingdx	trainscg
1	31	6	28	34
2	30	19	34	33
3	29	8	31	31
Total	**90**	**33**	**93**	**98**
%	**37.50**	**13.75**	**38.75**	**40.83**

A partir del desempeño de las últimas dos pruebas determinamos que las siguientes pruebas que realizaríamos serían únicamente considerado dos funciones para el entrenamiento, traingda y trainscg. Al mismo tiempo notamos que contrario a lo que reporta la teoría, la función Levenberg-Marquardt no resultó en una mejora del desempeño ya que es demasiado lenta en comparación con las demás y al reducir el número máximo de iteraciones para hacer posible su funcionamiento, el desempeño resultó bastante pobre.

4.7. Reconocimiento de palabras con redes neuronales, empleando las funciones traingda y trainscg para el entrenamiento

A lo largo de las siguientes pruebas buscamos observar como era el reconocimiento si modificábamos el número de capas que conformaban a la red neuronal, por lo que realizamos una serie de pruebas comparativas usando redes de 2, 3 y 4 capas. También se varió el número de segmentos y las repeticiones de prueba.

4.7.1. Reconocimiento de palabras con redes neuronales (Pruebas con diferente número de capas).

Para la realización de estas pruebas se consideró el corpus de las palabras fonéticamente distintas y el corpus de las palabras fonéticamente confusas, respectivamente, se emplearon 10 repeticiones de cada palabra para el entrenamiento y se reservaron 5 repeticiones para el reconocimiento, se utilizaron las funciones de entrenamiento traingda y trainscg, así como también, se usaron redes neuronales de 2 capas (45 neuronas en la primera y 16 en la segunda), 3 capas (25 neuronas en la primera, 45 en la segunda y 16 en la tercera) y 4 capas (25 en la primera, 45 en la segunda, 45 en la tercera y 16 en la cuarta). En la tabla IV-9 y IV-10, respectivamente, se comparan los resultados de estas pruebas.

Tabla IV-9. Reconocimiento de palabras fonéticamente distintas con redes neuronales (prueba con diferente número de capas).

	traingda	trainscg	traingda	trainscg	traingda	trainscg
Prueba	**[45 16]**	**[45 16]**	**[25 45 16]**	**[25 45 16]**	**[25 45 45 16]**	**[25 45 45 16]**
1	55	55	55	43	43	47
2	51	55	44	50	50	47
3	57	54	51	52	53	40
Total	**163**	**164**	**150**	**145**	**146**	**134**
%	**67.92**	**68.33**	**62.50**	**60.42**	**60.83**	**55.83**

Tabla IV-10. Reconocimiento de palabras fonéticamente confusas con redes neuronales (prueba con diferente número de capas).

	traingda	trainscg	traingda	trainscg	traingda	trainscg
Prueba	**[45 16]**	**[45 16]**	**[25 45 16]**	**[25 45 16]**	**[25 45 45 16]**	**[25 45 45 16]**
1	34	27	29	24	24	26
2	29	32	38	30	28	26
3	34	31	31	26	23	26
Total	**97**	**90**	**98**	**80**	**75**	**78**
%	**40.42**	**37.50**	**40.83**	**33.33**	**31.25**	**32.50**

Posteriormente, al observar el pobre desempeño del reconocimiento empleando redes neuronales, se planteó probar con algunas variantes en las especificaciones para mejorar el reconocimiento: usar una red neuronal por cada segmento, considerando 1 segmento, y posteriormente hacer uso del método de K-neighborg para determinar al ganador.

4.7.2. Reconocimiento de palabras con redes neuronales (una red neuronal por cada segmento).

Para la realización de estas pruebas se consideró el corpus de las palabras fonéticamente distintas y el corpus de las palabras fonéticamente confusas, respectivamente, se emplearon 10 repeticiones de cada palabra para el entrenamiento y se reservaron 5 repeticiones para el reconocimiento, se utilizaron las funciones de entrenamiento traingda y trainscg. En este caso se probó con una red de 2 capas (45 neuronas en la primera y 16 en la segunda) y 3 capas (45 neuronas en la primera y 16 en la segunda). En la tabla IV-11 y IV-12, respectivamente, se comparan los resultados de estas pruebas.

Tabla IV-11. Reconocimiento de palabras fonéticamente distintas con redes neuronales (una red neuronal por cada segmento).

	traingda	trainscg	traingda	trainscg
Prueba	**[45 16]**	**[45 16]**	**[25 45 16]**	**[25 45 16]**
% 1	41.5625	52.5	47.1875	44.375
% 2	48.4375	45.9375	41.5625	34.0625
% 3	45.9375	42.5	39.6875	35.9375
% Promedio	**45.31**	**46.98**	**42.81**	**38.13**

Tabla IV-12. Reconocimiento de palabras fonéticamente confusas con redes neuronales (una red neuronal por cada segmento).

	traingda	trainscg	traingda	trainscg
Prueba	**[45 16]**	**[45 16]**	**[25 45 16]**	**[25 45 16]**
% 1	21.25	18.75	19.6875	23.125
% 2	22.5	21.5625	23.125	21.875
% 3	24.375	18.75	25.9375	20.3125
% Promedio	**22.71**	**19.69**	**22.92**	**21.77**

Como otra posibilidad para mejorar el reconocimiento, se propuso obtener los vectores característicos considerando la palabra completa, en lugar de obtenerlos por cada segmento.

4.7.3. Reconocimiento de palabras con redes neuronales (considerando 1 segmento, o vectores característicos con la palabra completa).

Para la realización de estas pruebas se consideró el corpus de las palabras fonéticamente distintas y el corpus de las palabras fonéticamente confusas, respectivamente, se emplearon 10 repeticiones de cada palabra para el entrenamiento y se reservaron 5 repeticiones para el reconocimiento, se utilizaron las funciones de entrenamiento traingda y trainscg. En este caso se probó con una red de 2 capas (45 neuronas en la primera y 16 en la segunda) y 3 capas (45 neuronas en la primera y 16 en la segunda). En la tabla IV-13 y IV-14, respectivamente, se comparan los resultados de estas pruebas.

Tabla IV-13. Reconocimiento de palabras fonéticamente distintas con redes neuronales (considerando 1 segmento, o vectores característicos con la palabra completa).

	traingda	trainscg	traingda	trainscg
Prueba	**[45 16]**	**[45 16]**	**[25 45 16]**	**[25 45 16]**
% 1	66.25	61.25	67.5	61.25
% 2	62.5	61.25	56.25	58.75
% 3	63.75	62.5	65	42.5
% Promedio	**64.17**	**61.67**	**62.92**	**54.17**

Tabla IV-14. Reconocimiento de palabras fonéticamente confusas con redes neuronales (considerando 1 segmento, o vectores característicos con la palabra completa).

	traingda	trainscg	traingda	trainscg
Prueba	**[45 16]**	**[45 16]**	**[25 45 16]**	**[25 45 16]**
% 1	38.75	40	37.5	32.5
% 2	40	40	38.75	23.75
% 3	37.5	36.25	38.75	36.25
% Promedio	**38.75**	**38.75**	**38.33**	**30.83**

Posteriormente, se probó usando en el reconocimiento las mismas repeticiones empleadas en el entrenamiento, para observar como era el comportamiento del sistema en este caso.

4.7.4. Reconocimiento de palabras con redes neuronales (mismas repeticiones que en el entrenamiento).

Para la realización de estas pruebas se consideró el corpus de las palabras fonéticamente distintas y el corpus de las palabras fonéticamente confusas, respectivamente, se emplearon 10 repeticiones de cada palabra para el entrenamiento y se realizaron las pruebas con esas mismas repeticiones, se utilizaron las funciones de entrenamiento traingda y trainscg. En este caso se probó con una red de 2 capas (45 neuronas en la primera y 16 en la segunda) y 3 capas (25 neuronas en la primera, 45 en la segunda y 16 en la tercera). En la tabla IV-15 y IV-16, respectivamente, se comparan los resultados de estas pruebas.

Tabla IV-15. Reconocimiento de palabras fonéticamente distintas con redes neuronales (mismas repeticiones que en el entrenamiento).

	traingda	trainscg	traingda	trainscg
Prueba	**[45 16]**	**[45 16]**	**[25 45 16]**	**[25 45 16]**
% 1	76.25	70.625	73.125	66.875
% 2	74.375	75	68.75	70
% 3	70	73.75	69.375	66.875
% Promedio	**73.54**	**73.13**	**70.42**	**67.92**

Tabla IV-16. Reconocimiento de palabras fonéticamente confusas con redes neuronales (mismas repeticiones que en el entrenamiento).

	traingda	trainscg	traingda	trainscg
Prueba	**[45 16]**	**[45 16]**	**[25 45 16]**	**[25 45 16]**
% 1	51.25	49.375	52.5	53.125
% 2	54.375	26.875	48.75	43.75
% 3	55.625	53.75	51.25	48.125
% Promedio	**53.75**	**43.33**	**50.83**	**48.33**

De manera adicional, se planteó la utilización de redes adaptivas ("Adaptive Resonance Theory, ART"). En este caso usamos las redes ART y ARTMAP.

4.8. Pruebas de reconocimiento de palabras con redes neuronales, empleando "Adaptive Resonance Theory, ART"

4.8.1. Reconocimiento de palabras fonéticamente distintas con red neuronal ART (pruebas con mismos vectores empleados durante el entrenamiento de la red).

Para la realización de esta prueba se consideró el corpus de las palabras fonéticamente distintas, se emplearon 10 repeticiones de cada palabra para el entrenamiento y se realizaron las pruebas con esas mismas repeticiones, se utilizaron las funciones de entrenamiento correspondientes a la red neuronal ART. El resultado obtenido se observa en la tabla IV-17.

Tabla IV-17. Reconocimiento de palabras fonéticamente distintas con red neuronal ART.

Prueba	ART
% 1	5

4.8.2. Reconocimiento de palabras fonéticamente distintas con red neuronal ARTMAP (pruebas con mismos vectores empleados durante el entrenamiento de la red).

Para la realización de esta prueba se consideró el corpus de las palabras fonéticamente distintas, se emplearon 10 repeticiones de cada palabra para el entrenamiento y se

realizaron las pruebas con esas mismas repeticiones, se utilizaron las funciones de entrenamiento correspondientes a la red neuronal ARTMAP. El resultado obtenido se observa en la tabla IV-18.

Tabla IV-18. Reconocimiento de palabras fonéticamente distintas con red neuronal ARTMAP.

Prueba	ARTMAP
% 1	8.75

4.8.3. Reconocimiento de vectores (aleatorios con distribución normal) con redes neuronales (prueba con diferente número de capas e incluyendo ART y ARTMAP).

Para la realización de esta prueba se consideró un corpus conformado con ayuda datos generados de manera aleatoria con distribución normal (media=0 y desviación std.=1), se consideraron vectores de la misma dimensión a la utilizada en las pruebas anteriores, se utilizaron las funciones de entrenamiento traingda y trainscg, de tal manera que, se usaron redes neuronales de 2 capas (45 neuronas en la primera y 16 en la segunda), 3 capas (25 neuronas en la primera, 45 en la segunda y 16 en la tercera) y 4 capas (25 en la primera, 45 en la segunda, 45 en la tercera y 16 en la cuarta), y también se realizó la prueba usando la red ART y la ARTMAP. En la tabla IV-19 se comparan los resultados de estas pruebas.

Tabla IV-19. Reconocimiento de palabras fonéticamente distintas con redes neuronales (datos aleatorios con distribución normal).

	traingda	trainscg	traingda	trainscg	traingda	trainscg	ART	ARTMAP
Prueba	**[45 16]**	**[45 16]**	**[25 45 16]**	**[25 45 16]**	**[25 45 45 16]**	**[25 45 45 16]**		
% 1	100	100	100	100	100	100	100	100
% 2	100	100	100	100	100	100	100	100
% 3	100	100	100	100	100	100	100	100
Promedio	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Dado que las pruebas realizadas otorgaron un pobre desempeño se propuso una serie de pruebas adicionales considerando a las vocales como corpus, de modo que se pudiera observar el comportamiento con señales más pequeñas y al mismo tiempo con un corpus muy pequeño.

4.9. Reconocimiento de las vocales con redes neuronales

4.9.1. Reconocimiento de las vocales con redes neuronales

Para la realización de estas pruebas se consideró un corpus conformado por las vocales, se emplearon 10 repeticiones de cada vocal para el entrenamiento y posteriormente se realizaron las pruebas con 5 repeticiones, diferentes a las usadas en el entrenamiento, de cada una de ellas.

Se utilizaron las funciones de entrenamiento traingda y trainscg, así como también, se usaron redes neuronales de 2 capas (45 neuronas en la primera y 5 en la segunda), 3 capas (25 neuronas en la primera, 45 en la segunda y 5 en la tercera) y 4 capas (25 en la primera, 45 en la segunda, 45 en la tercera y 5 en la cuarta). Se dividieron las señales en 4, 1 y 2 segmentos para realizar las pruebas. En la tabla IV-20, IV-21 y IV-22, respectivamente, se comparan los resultados de estas pruebas.

Tabla IV-20. Reconocimiento de las vocales con distintas redes neuronales (considerando 4 segmentos).

	traingda	trainscg	traingda	trainscg	traingda	trainscg
Prueba	**[45 5]**	**[45 5]**	**[25 45 5]**	**[25 45 5]**	**[25 45 45 5]**	**[25 45 45 5]**
% 1	40	56	52	40	40	56
% 2	48	48	40	48	44	56
% 3	40	48	52	56	52	48
Promedio	**42.67**	**50.67**	**48.00**	**48.00**	**45.33**	**53.33**

Tabla IV-21. Reconocimiento de las vocales con distintas redes neuronales (considerando 1 segmento).

	traingda	trainscg	traingda	trainscg	traingda	trainscg
Prueba	**[45 5]**	**[45 5]**	**[25 45 5]**	**[25 45 5]**	**[25 45 45 5]**	**[25 45 45 5]**
%1	52	44	56	52	52	64
%2	44	36	40	48	48	44
%3	52	40	56	52	48	40
Promedio	**49.33**	**40.00**	**50.67**	**50.67**	**49.33**	**49.33**

Tabla IV-22. Reconocimiento de las vocales con distintas redes neuronales (considerando 2 segmentos).

	traingda	trainscg	traingda	trainscg	traingda	trainscg
Prueba	**[45 5]**	**[45 5]**	**[25 45 5]**	**[25 45 5]**	**[25 45 45 5]**	**[25 45 45 5]**
%1	52	44	36	44	52	44
%2	52	52	60	44	48	52
%3	52	48	44	48	48	52
Promedio	**52.00**	**48.00**	**46.67**	**45.33**	**49.33**	**49.33**

4.10. Análisis de resultados

Para poner en contexto las pruebas realizadas a lo largo de este capítulo, primeramente, se incluye un análisis de las pruebas realizadas con el sistema Sphinx 3 y después se realiza un análisis de los resultados obtenidos empleando redes neuronales.

4.10.1. Análisis de resultados con el sistema Sphinx 3

- Al realizar pruebas con el sistema Sphinx 3 haciendo uso del corpus del laboratorio de procesamiento de voz de la FI UNAM notamos que el desempeño no era lo suficientemente bueno, por lo que se buscó una manera de mejorarlo. Nuestra propuesta contempló aumentar a por lo menos 3 repeticiones de cada una de las palabras del corpus. Para ello se grabaron nuevas frases y se conformó un nuevo corpus agregando estas frases al corpus que se tenía. Los resultados de las pruebas usando el nuevo corpus, de manera general, arrojaron una mejoría en el desempeño, por lo que podemos afirmar que funcionó la propuesta que se planteó.

- Al reconocer palabras del corpus adicional se obtuvieron mejores resultados que al reconocer frases del corpus original, considerando que en ambos casos se entrenó el sistema con el nuevo corpus conformado. En la tabla IV-23 se comparan los resultados obtenidos al reconocer frases de los dos corpus.

Tabla IV-23. Resultados obtenidos al reconocer frases de los distintos corpus.

Corpus Reconocimiento	"SER"	"WER"
Original	87.5%	41.5%
Adicional	72.5%	40.6%

Esto se puede deber a varias causas: por ejemplo, que en las frases adicionales el número de hablantes es menor que en el corpus original, o también, a que las grabaciones de las frases adicionales estuvieron más cuidadas, es decir, en un ambiente con menos ruido, por lo que resultó en la obtención de mejores vectores característicos de las señales. Esto es mucho más notorio si vemos el resultado del "SER" en donde la mejoría es del 15%, a diferencia del "WER" donde la mejoría es solo del 0.9%.

- Si observamos el resultado del reconocimiento al entrenar con el corpus adicional y reconocer frases únicamente de este corpus, los resultados son aún mejores. En la tabla IV-24 se comparan los resultados al entrenar con el corpus nuevo y con el adicional.

Tabla IV-24. Resultados del reconocimiento del corpus adicional.

Entrenamiento	Reconocimiento	"SER"	"WER"
Corpus Nuevo	Corpus Adicional	72.5%	40.6%
Corpus Adicional	Corpus Adicional	60.0%	23.6%

La mejora se nota sobre todo en el "WER", donde la mejoría es de un 17%. Por lo que de acuerdo con este resultado podemos apuntar a que las grabaciones del corpus adicional son mejores.

4.10.2. Análisis de resultados empleando redes neuronales

- Habiendo realizado pruebas con el sistema Sphinx 3, si bien los resultados mostraron una mejoría en el desempeño del sistema, estos resultados aún distan de ser un resultado digno de un buen reconocedor. Por lo que se planteó otra hipótesis, la cual consistió en contar con un corpus de palabras fonéticamente confusas, mismas que, al ser reportadas como reconocidas por el sistema Sphinx 3 nos alertaran de un posible

error de reconocimiento. De manera que, al presentarse este caso se lleve a cabo nuevamente el reconocimiento pero solamente para el segmento de voz que contiene la palabra en cuestión. Para este nuevo reconocimiento se planteó el uso de redes neuronales y adicionalmente se consideró un corpus conformado con palabras fonéticamente distintas. Los resultados que obtuvimos, de manera general, no son los esperados ya que el desempeño es pobre sobre todo si los comparamos con los resultados que tomamos como "baseline" los cuales se basan en la técnica "VQ", con la cual trabajé en mi tesis licenciatura.

- Se corroboró la eficacia del uso de un sistema "VQ", el cual tomamos como "baseline", como ya se mencionó, ya que tanto para las pruebas con palabras fonéticamente distintas como para las pruebas con palabras fonéticamente confusas, el reconocimiento es muy alto. Lo cual podemos observar en la tabla IV-25.

Tabla IV-25. Reconocimiento usando "VQ".

Corpus	Reconocimiento con "VQ"
Fonéticamente Distintas	96.25%
Fonéticamente Confusas	82.5%

Además se verificó, como era de esperarse, que al realizar el reconocimiento de las palabras usadas durante en el entrenamiento el resultado es del 100% de reconocimiento, en ambos casos, corpus de palabras fonéticamente distintas y fonéticamente confusas.

- Después de realizar y observar los resultados de varias pruebas con los corpus de palabras fonéticamente distintas y fonéticamente confusas, así como, con diferentes configuraciones de redes neuronales. Notamos que, a diferencia de lo que considera la teoría, la función Levenberg-Marquardt no resultó en una mejora del desempeño, ya que es demasiado lenta en comparación con las demás y al reducir el número máximo de iteraciones su desempeño resultó bastante pobre, lo cual no era lo esperado.

- Observamos que al aumentar el número de capas en la configuración de la red neuronal no necesariamente mejora el reconocimiento, tanto en el caso del corpus de las palabras fonéticamente distintas como en el de palabras fonéticamente confusas, además, aun considerando el mejor de los casos el desempeño es muy pobre comparado con los resultados que se obtienen con "VQ".

- Al observar el pobre desempeño, que en general se obtuvo con estas pruebas, se planteó probar con algunas variantes en las especificaciones para el reconocimiento con redes neuronales una de ellas consistió en usar una red neuronal por cada segmento, considerando 4 segmentos y posteriormente hacer uso del método de K-neighborg para determinar al ganador. Los resultados que se obtuvieron muestran que el desempeño empeoró, ya que para el corpus de las palabras fonéticamente distintas, de un reconocimiento de entre 68.33% - 55.83% bajó a un reconocimiento de entre 46.98% - 38.13%, y de manera similar para el corpus de las palabras fonéticamente confusas, de un reconocimiento entre 40.83% - 31.25% bajó a un reconocimiento de entre 22.92% - 19.69%.

- Otra variante con la que se probó, consistió en obtener los vectores característicos considerando la palabra completa, es decir como si se tratara de un solo segmento. Después de realizar las pruebas, podemos decir que los resultados que obtuvimos no mostraron una diferencia significativa, sin embargo, se mantienen similares a las primeras pruebas con redes neuronales. Lo que se nota al observar los resultados para el corpus de las palabras fonéticamente distintas, donde de un reconocimiento de entre 68.33% - 55.83% bajó a un reconocimiento de entre 64.17% - 54.17%, y de manera similar para el corpus de las palabras fonéticamente confusas, de un reconocimiento de entre 40.83% - 31.25% bajó a un reconocimiento de entre 38.75% - 30.83%.

- A manera de verificar el desempeño del reconocimiento con redes neuronales, se probó usando en el reconocimiento las mismas repeticiones de las palabras empleadas durante el entrenamiento. Al realizar estas pruebas notamos una mejoría, como era de esperarse, ya que para el corpus de las palabras fonéticamente distintas de un

reconocimiento de entre 68.33% - 55.83% aumentó a un reconocimiento de entre 73.54% - 67.92%, y de manera similar para el caso de las palabras fonéticamente confusas de un reconocimiento de entre 40.83% - 31.25% aumentó a un reconocimiento de entre 53.75% - 43.33%. sin embargo, estos resultados no se comparan con los obtenidos haciendo uso de "VQ", donde se alcanzó un 100% de reconocimiento, tanto para el corpus de las palabras fonéticamente distintas como para el corpus de las palabras fonéticamente confusas.

- Otras pruebas que planteamos, con el fin de mejorar el desempeño del reconocimiento, consistieron en utilizar "Adaptive Resonance Theory", en este caso ART y ARTMAP. Los resultados que se obtuvieron mostraron un desempeño sumamente pobre, menor al 10% en ambos casos, lo cual nos hizo dudar de este método. Por lo que propusimos verificar estos métodos, empleando datos generados aleatoriamente con una distribución normal, y para este caso se alcanzó un reconocimiento del 100% en ambos casos, inclusive para las configuraciones de redes que se habían probado anteriormente. Esto no hace sino confirmar la complejidad de este tipo de señales y la gran dificultad que aún se tiene para obtener un buen sistema de reconocimiento de voz.

- Se probó el reconocimiento de las vocales con redes neuronales, y lo que pudimos notar es que los resultados no son muy diferentes a los casos de reconocimiento de los corpus de las palabras fonéticamente distintas o fonéticamente confusas. Ya que, de un reconocimiento de entre 68.33% - 55.83%, para el caso del corpus de las palabras fonéticamente distintas, o del reconocimiento de entre 40.83% - 31.25%, para el caso de las palabras fonéticamente confusas, el resultado del reconocimiento de las vocales no mejoró, por el contrario, se situó entre los resultados obtenidos para los corpus de las palabras fonéticamente distintas y de las palabras fonéticamente confusas, ya que se obtuvo un reconocimiento de entre 53.33% – 42.67%.

- Finalmente se realizaron pruebas de reconocimiento de las vocales con algunas variantes. De tal manera que en la tabla IV-26 podemos comparar los resultados

obtenidos considerando 4 segmentos y 11 vectores característicos por segmento, 1 segmento y 16 vectores característicos, y 2 segmentos y 16 vectores característicos.

Tabla IV-26. Reconocimiento de las vocales con distintas configuraciones.

Configuración	Reconocimiento
4 Segmentos – 11 vectores por segmento	53.33% - 42.67%
1 Segmento – 16 vectores por segmento	50.67% - 40.00%
2 Segmentos – 16 vectores por segmento	52.00% - 45.33%

Podemos ver que no varían mucho los resultados, aunque podemos destacar que con 2 segmentos resultan más consistentes. Además, adicionalmente, en todas las pruebas de las vocales un error que se presentó de manera recurrente, fue el que se reconociera la “o” como si fuera “u”.

V. Conclusiones

- Se logró realizar un análisis de los fundamentos de producción de la voz, anatómica y fisiológicamente, haciendo énfasis en las modificaciones que sufre la voz con relación a la edad.

- Se puso en funcionamiento el sistema de reconocimiento de voz Sphinx 3 con el corpus del laboratorio de procesamiento de voz de la FI UNAM, otorgándonos los resultados que tomamos como punto de partida.

- Se logró conformar un nuevo corpus, el cual sirvió para incrementar el porcentaje de reconocimiento en Sphinx 3.

- Se realizó un "manual" o guía de uso para el sistema Sphinx 3, lo cual permitirá que los desarrollos posteriores ahorren tiempo en la puesta en marcha del sistema Sphinx 3 y se aboquen a trabajar en otros tópicos. Esto resulta particularmente importante ya que no existía.

- En relación a la conveniencia de emplear redes neuronales para llevar a cabo el reconocimiento de voz, podemos afirmar que, si bien este proceso es bastante rápido en un equipo relativamente reciente, lo cual hace factible su inserción en el sistema de reconocimiento Sphinx 3, es evidente que la eficacia del método es muy baja dados los resultados que se obtuvieron y aún más cuando los comparamos con otros métodos como es el caso del reconocedor por "VQ". Además, resulta poco viable ya que a priori esperábamos que las ventajas que ofrecen las redes neuronales, tal como, la capacidad para efectuar un modelado no lineal, su alta capacidad de discriminación y su capacidad de adaptación redundaran en un mejor reconocimiento de voz, sin embargo, no se ven reflejadas en un buen desempeño al momento de efectuarlo.

- De acuerdo con los resultados obtenidos y detallados en este trabajo, se ha comprobado la insuficiencia de las redes neuronales como método para realizar el reconocimiento de las palabras fonéticamente confusas.

- Con base en los resultados obtenidos, queda como reto trabajar en otras arquitecturas de redes neuronales para resolver este problema de inteligencia artificial o aprendizaje de maquinas, y abordarlo en un trabajo doctoral. En este caso no fue el propósito diseñar arquitecturas de redes neuronales, sino probar las existentes y seleccionar algunas para realizar pruebas. Mismas con las que se demostró que son insuficientes.

- Queda también como una propuesta de trabajo a futuro, ya que en este caso por falta de tiempo no fue posible, el probar este reconocedor por redes neuronales, dándole como entrada al mismo vectores característicos de otro tipo, es decir, que se obtengan de manera diferente a los “LPC’s”, tal como, Cepstral, Mel frequency Cepstral (MFC), aplicándole a las señales la transformada en tiempo corto de Fourier (STFT). Esto, para observar como es el desempeño del reconocedor y determinar si resulta satisfactorio para alguno de ellos.

VI. Referencias

[1] J. Deller, J. Proakis y J. Hansen, Discrete-Time Processing of Speech Signals, 3ª ed., Prentice Hall, 1993.

[2] J. Reyes, Reconocimiento Continuo del Español Hablado en México. Tesis de Maestría en Ingeniería Eléctrica, Facultad de Ingeniería. UNAM, 2010.

[3] G. Saon y C. Jen-Tzung, «Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances,» *Signal Processing Magazine, IEEE,* vol. 29, nº 6, pp. 18-33, Nov. 2012.

[4] X. Huang, A. Acero y H. W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, 1ª ed., Prentice Hall, 2001.

[5] Huggins-Daines y David, «The CMU Sphinx Group Open Source Speech Recognition Engines,» 2001. [En línea]. Available: http://sourceforge.net/projects/cmusphinx/.

[6] E. Gouvéa, «Robust Group's Open Source Tutorial. Learning to use the CMU SPHINX Automatic Speech Recognition System,» The Sphinx Group, Pittsburg. CMU, 2001.

[7] A. Chan, E. Gouvéa, R. Singh, M. Ravishankar, R. Rosenfeld, Y. Sun, D. Huggins-Daines y M. Seltzer, «The Hieroglyphs: Building Speech Applications Using CMU Sphinx and Related Resourses,» 2007. [En línea]. Available: http://www-2.cs.cmu.edu/archan/documentation/sphinxDocDraft3.pdf.

[8] H. Printz y P. Olsen, «Theory and practice of acoustic confusability,» *Computer Speech and Language,* pp. 1-34.

[9] J. Anguita, J. Hernando, S. Peillon y A. Bramoullé, «Detection of confusable words in automatic speech recognition,» *Signal Processing Letters, IEEE,* vol. 12, nº 8, pp. 585-588, Aug. 2005.

[10] C. Ittichaichareon, S. Suksri y T. Yingthawornsuk, «Speech Recognition using MFCC,» *International Conference on Computer Graphics, Simulation and Modeling,* 2012.

[11] T. Adam y M. Salam, «Spoken English Alphabet Recognition with Mel Frequency Cepstral Coefficients and Back Propagation Neural Networks,» *International Journal of Computer Applications,* vol. 42, nº 12, 2012.

[12] F. Miyara, La voz humana. Material utilizado en la asignatura "Procesamiento Digital de Señales de Voz", Universidad Nacional de Rosario.

[13] G. J. Tortora y B. Derrickson, Principios de Anatomía y Fisiología, 11ª ed., Oxford University Press, 2006.

[14] B. Torres y F. Gimeno, Anatomía de la Voz, Paidotribo, 2008.

[15] IPA, «IPA: International Phonetic Association,» [En línea]. Available: http://www.langsci.ucl.ac.uk/ipa/.

[16] K. F. Lee, H. W. Hon y R. Reddy, «An overview of the SPHINX speech recognition system,» *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 38, nº 1, pp. 35-45, Jan. 1990.

[17] K. F. Lee, H. W. Hon, M. Y. Hwang, S. Mahajan y R. Reddy, «The SPHINX speech recognition system,» *Acoustics, Speech, and Signal Processing. 1989. ICASSP-89., 1989 International Conference on,* vol. 1, pp. 445-448, 23-26 May. 1989.

[18] M. Seltzer, «Sphinx III Signal Procesing Front End Specification,» 1999. [En línea]. Available: http://www.cs.cmu.edu//mseltzer/sphinxman/s3_fe_spec.pdf.

[19] R. Sharma, «Application of Neural Networks Models in Recognition Field: A survey,» *International Journal of Scientific & Engineering Research,* vol. 4, nº 2, 2013.

[20] E. Turban, R. Sharda, D. Delen y D. King, Business Intelligence: A Managerial Approach, Cap. 6: Neural Networks for Data Mining, 2ª ed., Pearson - Prentice Hall, 2010.

[21] M. T. Hagan y H. B. Demuth, Neural Network Design, 1ª ed., PWS Publishing Company, 1995.

[22] A. K. Jain, J. Mao y K. Mohiuddin, «Artificial neural networks: a tutorial,» *Computer,* vol. 29, nº 3, pp. 31-44, Mar. 1996.

[23] H. Y. Hen y H. Jenq-Neng, Handbook of Neural Network Signal Processing, CRC press, 2002.

[24] L. Cruz y M. Acevedo, «Reconocimiento de Voz usando Redes Neuronales Artificiales Backpropagation y Coeficientes LPC,» SEPI-Telecomunicaciones ESIME IPN Unidad Profesional "Adolfo Lopez Mateos", 2008.

[25] M. I. Garrido y J. A. Herrera, «Reconocedor de palabras confusas en idioma español basado en MSBC-VQ,» de *II Simposio "La investigación y Desarrollo en la Facultad de Ingeniería"*, 2005.

[26] O. F. Navarrete, Sistema Automatizado de Iluminación de una casa mediante comandos de voz. Tesis de licenciatura en Ingeniería Eléctrica Electrónica, Facultad de Ingeniería. UNAM, 2009.

[27] A. Acero y R. Stern, «Environmental robustness in automatic speech recognition,» *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on,* vol. 2, pp. 849-852, 3-6 Apr. 1990.

[28] C. M. D. Acevedo y M. G. Nieves, «Integrated System Approach for the Automatic Speech Recognition using Linear predict Coding and Neural Networks,» *Electronics, Robotics and Automotive Mechanics Conference, 2007. CERMA 2007,* pp. 207-212, 25-28 Sept. 2007.

[29] K. Agaram, S. Keckler y D. Burger, «A characterization of speech recognition on modern computer systems,» *Workload Characterization, 2001. WWC-4. 2001 IEEE International Workshop on,* pp. 45-53, 2 Dec. 2001.

[30] Y. Bengio, R. De Mori, G. Flammia y R. Kompe, «Global optimization of a neural network-hidden Markov model hybrid,» *Neural Networks, IEEE Transactions on,* vol. 3, nº 2, pp. 252-259, Mar. 1992.

[31] N. Botros, M. Z. Deiri y P. Hsu, «Automatic voice recognition using artificial neural network approach,» *Circuits and Systems, 1989., Proceedings of the 32nd Midwest Symposium on,* vol. 2, pp. 763-765, 14-16 Aug. 14-16 Aug. 1989.

[32] G. Carpenter, S. Grossberg y D. Rosen, «ART 2A: An Adaptive Resonance Algorithm for Rapid Category Learning and Recognition,» *Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference on,* vol. 2, pp. 151-156, 8-14 Jul 1991.

[33] G. Carpenter y S. Grossberg, «ART3: Hierarchical Search Using Chemical Transmiters in Self-Organizing Pattern Recognition Architectures,» *Neural Networks,* vol. 3, pp. 129-152, 1990.

[34] F. Casacubierta, R. García, J. LListerri, C. Nadeu, J. Pardo y A. Rubio, «Desarrollo de corpus para investigación en tecnologías del habla (Albayzín),» *Procesamiento del Lenguaje Natural,* vol. 12, pp. 35-42, 1992.

[35] J. O. Cuétara Priede, Fonética de la Ciudad de México. Aportaciones desde las tecnologías del habla, Tesis de Maestría en Lingüística Hispánica, Posgrado en Lingüística. UNAM, 2004.

[36] M. I. Garrido y J. A. Herrera, «Reconocedor de palabras confusas en idioma español basado en el modelo oculto de Markov,» de *III Simposio "La Investigación y Desarrollo en la Facultad de Ingeniería"*, 2006.

[37] S. K. Hasnain, M. Maqsood, M. Shahzad y S. Bashir, «Development of Speech Recognition System,» *Technology Forces: Journal of Egineering and Sciences,* vol. 2, nº 1, 2008.

[38] J. A. Herrera Camacho, Apuntes del curso de Procesamiento Digital de Voz, Facultad de Ingeniería. UNAM, 1999.

[39] S. Hu, D. Mulvaney y S. Datta, «Modification of Sphinx 3 for embebed system implementarion,» *Multimedia, Signal Processing and Communication Technologies (IMPACT), 2011 International Conference on,* pp. 137-140, 17-19 Dec. 2011.

[40] X. Huang y K. F. Lee, «On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition,» *Speech and Audio Processing, IEEE Transactions on,* vol. 1, nº 2, pp. 150-157, Apr. 1993.

[41] J. Huerta, S. Chen y R. Stern, «The 1998 Carnegie Mellon University Sphinx-3 Spanish Broadcast News Transcription System,» Department of Electrical and Computer Engineering and School of Computer Science. CMU.

[42] D. Jurafsky y H. M. James, Speech and Languaje Processing. An Introduction to Languaje Processsing. Computational Linguistics and Speech Recognition, Prentice Hall, 2000.

[43] K. F. Lee y H. W. Hon, «Large-vocabulary speaker-independent continuous speech recognition using HMM,» *Acoustics, Speech and Signal Processing, 1988. ICASSP-88., 1988 International Conference on,* vol. 1, pp. 123-126, 11-14 Apr. 1988.

[44] O. Nieto, Diseño de un reconocedor de comandos de voz para el DSP TMS320C6711, Tesis de Maestría en Ingeniería Eléctrica, Facultad de Ingeniería. UNAM, 2006.

[45] J. Proakis y D. Manolakis, Tratamiento digital de señales, 4ª ed., Pearson - Prentice Hall, 2007.

[46] L. Rabiner, «A tutorial on hidden Markov models and selected applications in speech recognition,» *Proceedings of the IEEE,* vol. 77, nº 2, pp. 257-286, Feb. 1989.

[47] L. R. Rabiner y J. B.H., Fundamentals of Speech Recognition, Prentice-Hall, 1993.

[48] W. Rozzi y R. Stern, «Speaker adaptation in continuous speech recognition via estimation of correlated mean vectors,» *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on,* vol. 2, pp. 865-868, 14-17 May. 1991.

[49] R. Sanchez, Reconocimiento adaptable de palabras aisladas utilizando redes neuronales, Tesis de Maestría en Ciencias de la Computación, UNAM, 1997.

[50] F. Teles y L. Lee, «A Neural Architecture Based on the Adaptive Resonant Theory and Recurrent Neural Networks,» *International Journal of Computer Science & Applications,* vol. 4, nº 3, pp. 45-56, 2007.

[51] G. J. Tortora, A Photographic Atlas of the Human Body with selected cat, sheep, and cow dissections, 2nd ed., 2004.

[52] E. Trentin y M. Gori, «Robust combination of neural networks and hidden Markov models for speech recognition,» *Neural Networks, IEEE Transactions on,* vol. 14, nº 6, pp. 1519-1531, Nov. 2003.

[53] E. Uraga y C. Gamboa, VOXMEX Speech Database: Design of a Phonetically Balanced Corpus, Departamento de ciencias de la computación, IIMAS. UNAM.

[54] A. K. Veera, «Speech Recognition Based on Artificial Neural Networks,» Helsinki University of Technology.

[55] F. F. Zeng y P. C. Shi, «Neural network design based on isolated words,» *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on,* vol. 2, pp. 769-772, 10-13 July 2011.