Capítulo I

El reconocimiento de voz

Un sonido es, al final, una fórmula matemática que se puede representar en un par de ejes cartesianos. Las curvas esenciales que lo forman son muy simples (senos y cosenos), pero las posibles combinaciones de estas curvas pueden llegar a ser tan complicadas como se quiera. Debido a esta complejidad se produce el fenómeno del habla.

Debido a su naturaleza, el habla es el proceso de comunicación más eficiente y económico entre los seres humanos. Razón por la cual, desde hace cinco décadas, investigadores y desarrolladores en el área de voz se han centrado en el estudio y desarrollo de interfaces con computadoras, de tal manera que la implantación de éstas permitan la realización de tareas y el desarrollo de nuevas tecnologías por medio del reconocimiento de voz.

El reconocimiento de voz es el proceso de convertir una señal acústica a una secuencia de palabras, representadas en forma de texto. Las palabras reconocidas pueden ser el resultado final, así como también pueden servir para entrada a otros sistemas que los usan para hacer alguna acción. La meta de la mayoría de las investigaciones en reconocimiento de voz es desarrollar una máquina que tenga la habilidad de entender habla conversacional, sin restricciones de vocabulario, proveniente de cualquier locutor [Rabiner93]. Cuando se logre esta meta, en un futuro no muy lejano las aplicaciones en reconocimiento de voz se van a volver algo tan normal y útil que van a tener un papel importante en nuestras tareas cotidianas.

1.1 Antecedentes

La investigación de tecnologías en reconocimiento de voz empezaron a finales de los 50's con la llegada del computador digital. Esto combinado con herramientas para capturar y analizar la voz (convertidores de señal análogo a digital y espectogramas de sonidos) permitió a los investigadores buscar otras maneras de extraer características de la voz que mostraran las diferentes propiedades de las palabras.

En los 60's se dieron avances en la segmentación automática de voz en unidades lingüísticas relevantes (fonemas, sílabas y palabras) y en los algoritmos de *pattern-matching* y clasificación. Por los 70's surgió un número de técnicas importantes hasta la fecha, creadas en parte por investigación de *Defense Advanced Research Projects Agency* (DARPA) [Baecker96]. Se hicieron reconocedores que manejaban un dominio de reconocimiento mayor y que estaban basados en el enfoque de reconocimiento de patrones.

La década de los 80´s trajo consigo un cambio del enfoque de reconocimiento de patrones hacia métodos de modelado probabilístico. En los 90´s la innovación tecnológica, tanto mejoras en los algoritmos como avances en poder computacional, ha permitido una notable mejoría en sistemas de reconocimiento de voz. Unido a ésto, las técnicas de hace algunos años han sido refinadas hasta el grado que actualmente se han obtenido muy buenos resultados de reconocimiento, y hay en el mercado sistemas comerciales a precios razonables.

1.2 Características Acústicas

El fonema es la unidad básica del habla, y un conjunto de fonemas determina los sonidos con los cuales se pueden construir palabras. Por otro lado, un alófono es una de las diferentes pronunciaciones para un fonema en particular. El fonema es una abstracción y realmente no puede ser pronunciado excepto en términos de uno de sus alófonos [Dalbor69]. Tomando en cuenta lo anterior se puede decir que en cualquier lenguaje el número de fonemas es obviamente más pequeño que el número de fonos o alófonos.

Tenemos la idea ortográfica de que es muy fácil separar la palabra “mesa” en cuatro sonidos: /m/e/s/a/, pero físicamente es un continuo de principio a final. Como consecuencia, en la vocal /e/ hay elementos de la /m/ y de la /s/. Además, para poder hablar de prisa no llegamos a efectuar los movimientos completos, y los dejamos a medias. Por esas y otras razones, el reconocimiento automático de fonemas no es tan fácil como uno quisiera.

1.2.1 Realización fonética

Los sonidos se producen cuando el aire espirado por los pulmones llega hasta la laringe, donde se encuentran las cuerdas vocales. Estas cuerdas vocales son dos músculos gemelos, elásticos, que vibran cuando el aire espirado pasa por ellas. Es entonces cuando se produce el sonido que llamamos voz. La voz pasa a la cavidad bucal o la cavidad nasal, donde los órganos, principalmente los de la boca, configuran y matizan los diversos sonidos en el habla.

Los fonemas se realizan por medio de sonidos. Los encargados de producir esos sonidos, de realizarlos fonéticamente, son los órganos de fonación.

1.2.2 Clasificación de los fonemas

Los sonidos están divididos en dos grupos principales: consonantes, si la corriente de aire es detenida u obstruida y vocales, si el aire sale libremente.

1.2.2.1 Fonemas consonánticos

Se producen sónidos consonánticos cuando el aire al salir encuentra un obstáculo, ya sea cerrándole totalmente el paso, o dejándole una estrechez por donde pasa con fricción. Para clasificar a las consonantes se tendrá en cuenta el concepto de articulación. Se entiende por articulación de un sonido la posición adoptada por los órganos de la cavidad bucal en el momento de producirse un sonido.

En la producción y clasificación de las consonantes hay que tener en cuenta los siguientes factores:

- Lugar de articulación.
- Manera de articulación.
- Articulaciones sonoras o sordas.

Lugar de articulación

Para el idioma Español hablado en México hay nueve lugares de articulación para las consonantes:

1. *Bilabial*. El labio inferior contra o cerca del labio superior. Ejemplo: [m] en *más*.
2. *Labiodental*. El labio inferior contra o cerca del borde de los dientes frontales superiores. Ejemplo: [f] en *fuente*.

3. *Dental*. Posición de la lengua contra el borde o atrás de los dientes superiores frontales. Ejemplo: [d] en *dar*.
4. *Alveolar*. Posición de la lengua contra o cerca de la zona alveolar. Ejemplo: [n] en *luna*.
5. *Palatal*. Posición de la lengua contra o cerca del paladar duro. Ejemplo: [ch] en *chico*.
6. *Velar*. Dorso de la lengua contra o cerca del velo. Ejemplo: [k] en *calor*.
7. *Bilabio-velar*. El labio inferior cerca del labio superior y al mismo tiempo el dorso de la lengua cerca del velo. Ejemplo: [w] en *hueso*.
8. *Uvular*. Dorso de la lengua contra el uvula. Ejemplo: [rr] en *carro*.
9. *Glotal*. Movimiento de cuerdas vocales. Ejemplo: [x] en *jardín*.

Manera de articulación

Independientemente de cuál sea la zona o punto de articulación, los órganos de articulación trabajan juntos de diferentes maneras para producir los fonemas. A esto se le llama manera de articulación. En Español hay siete maneras de articulación en las consonantes.

1. *Oclusivas*. En algún punto de la articulación del sonido la corriente de aire queda detenida y después es liberada con una pequeña explosión. Ejemplo: [p] en *pasar*.
2. *Fricativas*. La corriente de aire, sin ser detenida, es forzada a través del tracto vocal existiendo un cierre parcial, provocando que el aire salga con turbulencia.
3. *Africativas* (*Oclusiva* + *Fricativa*). La corriente de aire es detenida como en una oclusiva; pero en lugar de ser liberada abruptamente, es liberada con fricción como en un fricativo.
4. *Nasales*. El velo es bajado y la corriente de aire pasa a traves de la cavidad nasal con gran resonancia. Ejemplo: [m] en *más*.

5. *Semivocales*. La cavidad oral es cerrada a la mitad, pero la corriente de aire escapa por ambos lados del lugar de articulación. [l] en *luna*.

6. *Vibrantes*. La posición de la lengua, bajo tensión, conecta el alveolar una vez que la corriente de aire pasa a través del tracto vocal. Ejemplo: [r] en *pero* y [rr] en *perro*.

Articulaciones sonoras o sordas

Todo fonema, sea cual sea su punto o manera de articulación, puede producirse con vibraciones de las cuerdas vocales o sin ellas.

Fonemas Sordos	Fonemas Sonoros
s, j, f, ch, p, t, k	b, d, g, m, n, ñ, l, r, rr, y, w

Tabla 1.1 Clasificación de las consonantes de acuerdo a su articulación

1.2.2.2 Fonemas vocálicos

Con las vocales, la posición de la lengua y la forma en que se posiciona la boca determinan su timbre. Debido a que el aire no es bloqueado o detenido, no hay contacto entre los articuladores superiores e inferiores. En su lugar, la posición de la lengua es significante debido a que cambia el tamaño y forma de la cavidad oral.

Acústicamente, cuando pronunciamos una vocal hay unas bandas de frecuencia que se amplifican más según donde pongamos la lengua: una /a/, por ejemplo, se articula de forma más estrecha en la zona de la faringe y más ancha en la zona de la boca, lo cual lleva a la amplificación de un tramo de frecuencia que está alrededor de los 700Hz, y otro de los 1300Hz. Estas frecuencias se conocen como formantes.

Hay tres posiciones verticales de la lengua, las cuales determinan el grado de abertura de la cavidad oral: Alto, cuando la lengua está cerca del paladar; medio, cuando la lengua se encuentra a la mitad de la boca; y bajo. También hay tres posiciones de la lengua de frente hacia atrás: Anterior, cuando la parte alta de la lengua está muy cerca del alveolar; central, cuando la parte más alta de la lengua está en el centro de la cavidad oral; y posterior, donde la parte más alta de la lengua, el dorso, está muy cerca del velo.

Por esto se puede clasificar a las vocales en un cuadro bidimensional. En español hay cinco posiciones principales.

	Anterior	**Central**	**Posterior**
Alto	/i/		/u/
Medio	/e/		/o/
Bajo		/a/	

Tabla 1.2 Clasificación de los fonemas vocálicos

Las vocales son de mayor duración que las consonantes y están bien definidas espectralmente. Por esto las vocales usualmente son fáciles de reconocer y contribuyen significativamente al proceso de reconocimiento de voz. De acuerdo a la manera en que el tracto vocal se configure se determinan las frecuencias de resonancia del tracto (las formantes) y por ello el sonido se produce. Las vocales se identifican por sus formantes, las cuales son muy fuertes durante todo el fonema. A continuación se muestra un espectograma para la vocal /i/.

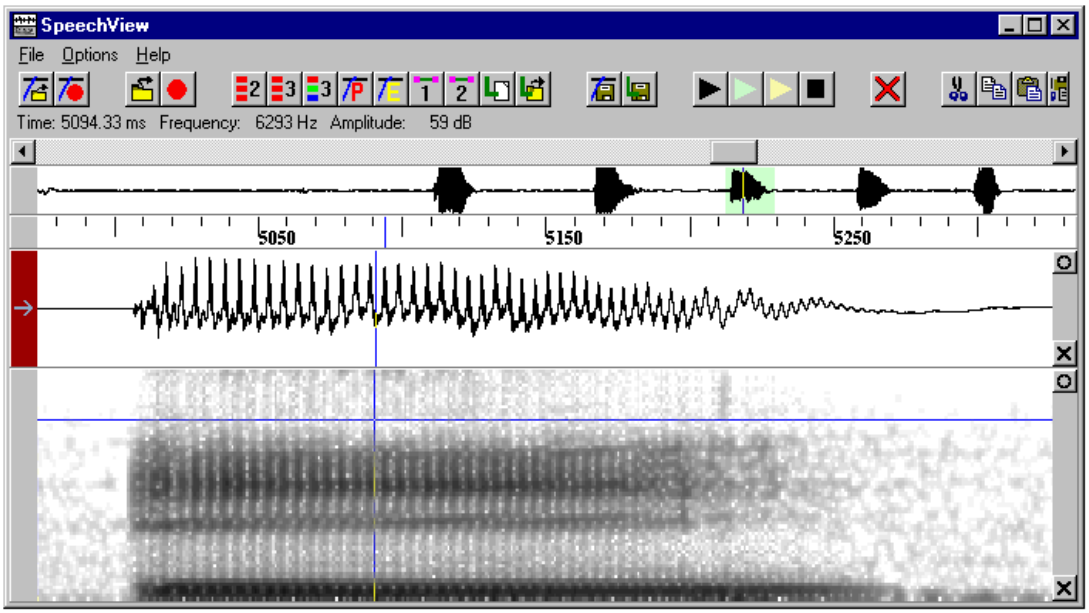

Figura 1.1 Espectograma del fonema /i/

En la figura 1.1 se muestra la representación espectral del fonema /i/, la cual tiene la segunda formante (F2) muy cerca de la formante superior F3. De manera parecida cada vocal tiene su propia configuración de formantes. Esto es muy útil en reconocimiento de voz, ya que las frecuencias a ser medidas nos pueden señalar en determinado momento de que tipo de fonema estamos hablando.

El principal motivo para la clasificación de los fonemas de acuerdo a su manera y lugar de articulación es para poder más adelante hacer experimentos en los cuales se agrupan los fonemas para efectos de mejoras en el reconocimiento. En esta clasificación se van a agrupar aquellos fonemas que compartan características acústicas similares.

1.3 Componentes básicos de un reconocedor

Existen tres componentes básicos en cualquier reconocedor de voz:

1. *Una representación de la señal de voz*. La representación es la forma en la cual el reconocedor convierte la señal de voz, antes de que empieze el analisis para identificar las unidades (palabras, silabas, fonemas, etc). Unas de las representaciones más usadas son los coeficientes *Linear Predictive Coding* (LPC) y los coeficientes *Mel-Frecuency Cepstrum Coefficients* (MFCC).

2. *Un conjunto de modelos*: descripciones de cada unidad a ser reconocida a partir de la representación de la señal de voz usada por el reconocedor. Los modelos también describen las unidades en el vocabulario del reconocedor. En este caso, las unidades son fonemas o alófonos.

3. *Un algoritmo de reconocimiento de patrones* para determinar qué modelo es el más parecido a la porción actual de la señal de voz dada como entrada.

1.4 Tipos de reconocedores

Cuando se diseña un reconocedor es importante entender claramente el propósito para el cual se va a utilizar. Ya que de lo contrario pueden llegar a existir situaciones que tienden a influir negativamente en el desempeño de este reconocedor. Por ejemplo, para construir un sistema de dictado monolocutor, es importante que la fase de entrenamiento se haga con la persona que lo va a utilizar finalmente. Algunas de éstas situaciones o aspectos se mencionan a continuación.

Los sistemas *independientes del locutor* pueden reconocer voz de cualquier persona. Los sistemas *dependientes del locutor* deben ser entrenados para cada usuario individual y típicamente tienen tasas más altas de reconocimiento. El habla proveniente de un solo locutor es más fácil de reconocer que el habla proveniente de una gran variedad de locutores, debido a que la mayoría de los parámetros de representación del habla son sensibles a las características de un locutor en particular. Esto hace que el habla que es reconocido muy bien de un locutor no sea tan bien reconocido de otro locutor. Los sistemas *adaptables al locutor*, un enfoque híbrido, inicia con plantillas independientes del locutor y las adapta a usuarios específicos sobre el tiempo sin entrenamiento explícito.

Los sistemas de voz *continuos* pueden reconocer habla en un ritmo natural mientras que los sistemas de *palabras aisladas* requieren de una pausa deliberada entre cada palabra. El habla continua es más difícil de procesar dado que los efectos de coarticulación causan que una palabra se pronuncie de manera diferente dependiendo de su posición con respecto a otras palabras en una oración [Baecker96]. Hoy en día, casi no existen sistemas de palabras aisladas, debido a la falta de naturalidad que imponen sobre los usuarios.

El tamaño del conjunto de palabras a ser reconocidas también influye fuertemente en el nivel de reconocimiento. Los vocabularios grandes tienden a contener más palabras confundibles que los vocabularios pequeños. El tamaño del vocabulario puede variar de 20 palabras a más de 40,000 palabras. Los grandes vocabularios causan dificultades en mantener exactitud, pero los pequeños pueden imponer restricciones no deseadas sobre la naturalidad de la comunicación. A menudo el vocabulario debe ser restringido por reglas gramaticales.

La gramática del dominio de reconocimiento define las secuencias de palabras permitidas. Una gramática altamente restringida es aquella en la cual es pequeño el número de palabras que pueden seguir legalmente cualquier otra palabra. A la cantidad de restricciones en la elección de una palabra se le da el nombre de *perplejidad* de la gramática. Los sistemas con baja perplejidad son potencialmente más precisos que aquellos que permiten al usuario mayor libertad.

El ruido de fondo cambia según las características del micrófono y esto puede influir dramaticamente en el nivel de reconocimiento. Muchos sistemas de reconocimiento de voz tienen muy buen desempeño en ambientes quietos o silenciosos. Sin embargo, el nivel de reconocimiento baja cuando se introduce ruido o cuando se trabaja en condiciones diferentes a como fue entrenado el sistema. Los sistemas evaluados en esta tesis son diseñados para reconocer el habla a través del teléfono, lo cual es difícil de reconocer, tanto por la variedad y baja calidad de los micrófonos en los teléfonos como por el ruido y el bajo ancho de banda de las líneas telefónicas.

Parámetros	**Rango**
Modo de hablar	Palabras aisladas vs. Habla continua
Estilo de Hablar	Habla espontánea vs. Habla leida
Enrolamiento	Dependiente vs. Independiente del locutor
Vocabulario	Pequeño (menor de 20 palabras) vs. Grande (mayor a 20,000 palabras)
Modelado fonético	Libre vs. Sensible al contexto
Ambiente	Controlado vs. Ruidoso

Tabla 1.3 Parámetros típicos usados para caracterizar la capacidad de los sistemas de reconocimiento de voz

1.5 Aplicaciones

A pesar que el desempeño de los sistemas de reconocimiento de voz está lejos de llegar a ser perfecto, éstos sistemas han probado su utilidad para ciertas aplicaciones. Uno de los medios más populares para aplicaciones de voz es el teléfono, debido a razones de disponibilidad, facilidad de uso y costo accesible. Entre esas aplicaciones podemos encontrar los siguientes:

- Servicios financieros.
- Asistencia de directorio.
- Llamadas por cobrar automáticas.
- Transferencia de llamadas telefónicas.
- Consultas.
- Reservaciones.

Existen otros tipos de aplicaciones que no están basados en el teléfono y requieren de alta calidad en los datos, por ejemplo el dictado automático. También el reconocimiento de voz es aplicado en compañías donde la entrada de datos o comandos por voz es requerida debido a que las manos del operador están ocupadas. Otras aplicaciones similares se pueden encontrar en la inspección de productos, desarrollo de inventarios y el control de robots. El reconocimiento de voz también encuentra una frecuente aplicación en medicina, donde la entrada de voz puede acelerar significativamente la escritura o la rutina de reportes.

1.6 Representaciones de la señal de voz

Los sonidos son variaciones en la presión del aire a través del tiempo. La voz es un subconjunto de los sonidos generados por el tracto vocal. Estos sonidos pueden ser digitalizados por un micrófono o cualquier otro medio que convierta la presión del aire a pulsos eléctricos.

Ya digitalizada la voz, una forma de representar el sonido es por medio de los *waveforms.* En estas representaciones el eje horizontal describe al tiempo y el eje vertical la amplitud. Los *waveforms* representan el sonido original, y una vez que se obtienen se le pueden aplicar procesamiento de señales digitales para conseguir información. Esta forma de representación no es muy utilizada ya que no muestra de manera clara las propiedades acústicas de su contenido.

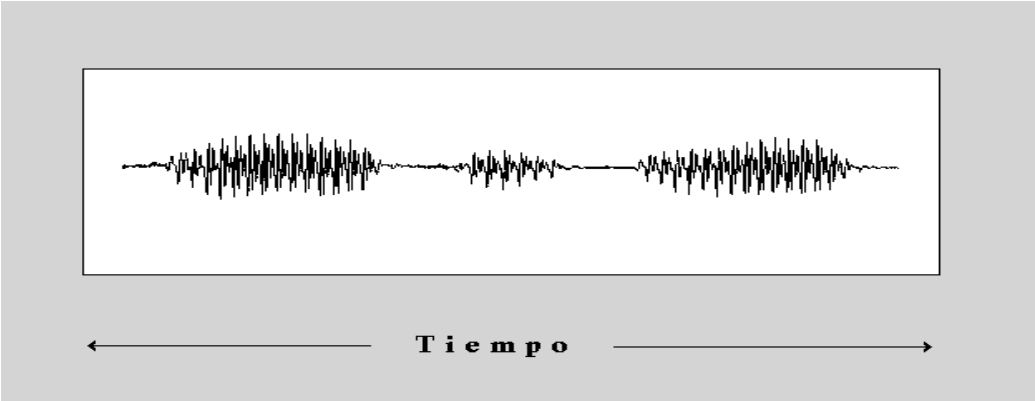

Figura 1.2 Aspecto de un Waveform

Los espectogramas son una representación más adecuada para el análisis del habla por computadora. Estos son una transformación del *waveform* al dominio de frecuencias, revelando características acústicas específicas del habla.

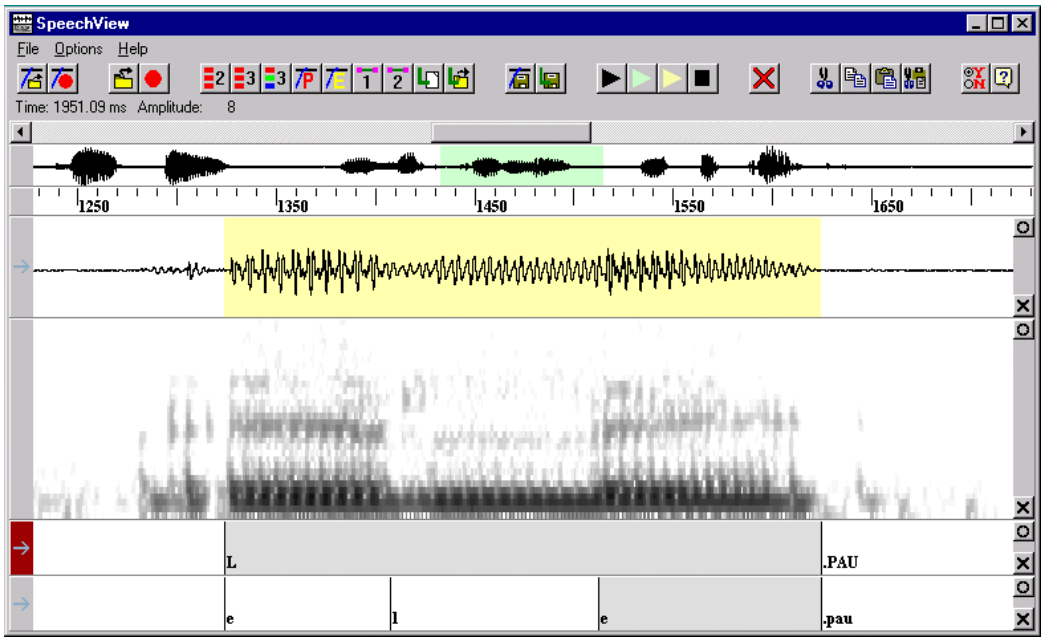

Figura 1.3 Espectograma de la palara “ele”

Fonema	Ejemplo	Clasificación
a	papa	Vocal
e	pepe	Vocal
i	pipa	Vocal
o	boda	Vocal
u	duda	Vocal
p	poca	Oclusiva sorda
b	boca	Oclusiva sonora
t	toca	Oclusiva sorda
d	dos	Oclusiva sonora
k	casa	Oclusiva sorda
g	goma	Oclusiva sonora
f	fila	Fricativa sorda
s	sopa	Fricativa sorda
x	jota	Fricativa sorda
V	nueve	Fricativa sonora
D	dedo	Fricativa sonora
G	igriega	Fricativa sonora
tS	cholo	Africativa sorda
dZ	llave	Africativa sonora
m	mata	Nasal
n	nata	Nasal
nj	baño	Nasal
N	cinco	Nasal
l	lana	Semivocal
j	mayo	Fricativa sorda
w	cuatro	Semivocal
r	pero	Vibrante (flap)
rr	perro	Vibrante (trill)

Tabla 1.4 Ejemplos de algunos fonemas del español hablado en México