INDICE

CAPITULO I

LA SEÑAL DE VOZ:
PRODUCCIÓN, PERCEPCIÓN Y FONETICA ACÚSTICA

1.1 INTRODUCCIÓN

El presente capítulo se basa en la primera sección de [RabJua93]. En éste presentaremos los mecanismos de producción y percepción de la voz en los seres humanos, mostrando el proceso natural y comparándolo con el reconocimiento de voz por computadora. Además indicaremos las diferentes representaciones de la voz en el tiempo y en dominio de frecuencia, de tal manera que puedan ser apreciados mediante gráficos. Se explicará la fisiología y funcionalidad del aparato fonador.

Finalmente, presentaremos las diversas clases de sonidos de voz (fonemas) del español, los cuales son caracterizados de acuerdo a sus propiedades acústicas.

1.2 PRODUCCIÓN DE LA SEÑAL DE VOZ EN LOS SERES HUMANOS

La señal de voz se produce al expulsar el aire de los pulmones y pasar éste por el tracto vocal antes de ser emitido al exterior. *Los sonidos sonoros* (por ejemplo las vocales) se producen con vibración de las cuerdas vocales; la onda periódica producida por las mismas se propaga por el tracto vocal, que produce resonancias a unas frecuencias determinadas, en modo similar a como el cuerpo de una guitarra trata las vibraciones de cualquiera de sus cuerdas. Estas resonancias (que reciben el nombre de formantes) distinguen unos sonidos de otros y son gobernadas por el locutor al dar

diversas formas al tracto vocal con los labios y la lengua fundamentalmente; en la voz normal, se sitúan entre los 200Hz y 3.5kHz. El período de la señal se corresponde con el período de vibración de las cuerdas vocales y define el *tono;* éste puede ser cambiado a voluntad por el locutor. El tono constituye la base de la melodía del canto. El intervalo frecuencial que un locutor puede abarcar con el tono de su voz es una característica individual; el tono medio en los hombres es de 130Hz, mientras que las mujeres tienen un promedio de 220Hz.

La ciencia ha establecido que para que se produzca sonido se requieren tres elementos: un cuerpo elástico que vibre, un agente mecánico que ponga en movimiento ese cuerpo elástico, y una caja de resonancia que amplifique las vibraciones haciéndolas perceptibles al oído, a través de las ondas que las transmiten por el aire. La voz humana posee estas tres condiciones señaladas. El cuerpo elástico que vibra son dos membranas situadas en la garganta llamadas cuerdas vocales, el medio mecánico es el aire. La caja de resonancia está formada por parte de la garganta y por la boca[1].

1.3 ESTUDIO COMPARATIVO DEL PROCESO DE PRODUCCIÓN Y PERCEPCIÓN DE LA VOZ EN LOS SERES HUMANOS Y LAS COMPUTADORAS.

El proceso de producción del habla empieza cuando el hablante formula un mensaje en su mente de lo que desea transmitir hacia el oyente, vía voz. De

[1] [Cast]

manera similar el computador para el proceso de formulación del mensaje genera un texto impreso. En el siguiente paso el emisor realiza la conversión del mensaje a un código de lenguaje, esto corresponde, en el computador a la transformación del texto en una secuencia de fonemas relativos a los sonidos de cada palabra. A través de la prosodia[2] se determina la duración de los sonidos, intensidad y tono asociado a cada sonido. Una vez que se escoge el código del lenguaje el hablante debe ejecutar una serie de movimientos neuromusculares[3] para hacer vibrar las cuerdas vocales. Estos comandos deben controlar simultáneamente los movimientos articulatorios incluyendo: los labios, mandíbula, lengua y el velo del paladar.

Una vez que la señal de voz es generada y propagada hacia el oyente, inicia el proceso de percepción (reconocimiento de voz). Primero el oyente procesa la señal acústica a través de la —membrana basilar[4]— ubicada en el oído interno, la cual realiza un análisis del espectro de la señal de entrada. Un proceso de —transducción neural[5]— convierte la señal espectral en un conjunto de señales que son receptadas por el nervio auditivo, lo que corresponde al proceso de extracción de características. En una manera que aún no es bien entendida, las señales que llegan al nervio auditivo se convierten a un código de lenguaje en los centros de alto procesamiento del cerebro y finalmente se logra la compresión del mensaje.

---

[2] Parte de la gramática tradicional que estudia la correcta pronunciación y acentuación.
[3] Los movimientos neuromusculares conforman un mecanismo que sin influir directamente sobre el modo de desempeñarse de una persona, se incorporan en representaciones cognoscitivas tales como: la percepción, imágenes, ideas y juicios
[4] Ver Anexo I
[5] Ver Anexo II

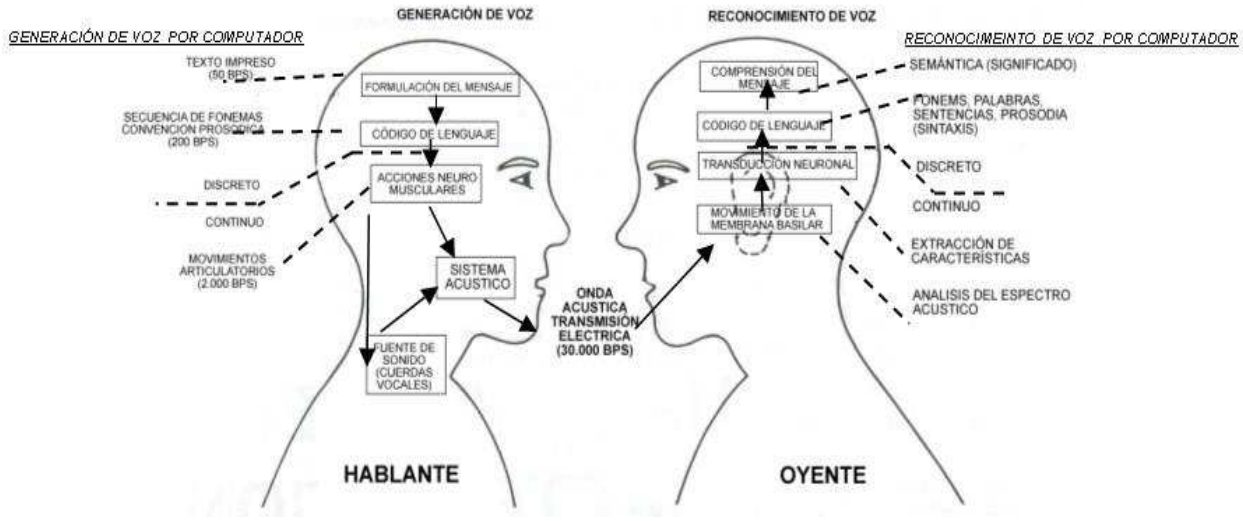


Fig. 1-1: Diagrama esquemático de producción y percepción de la voz

1.4 REPRESENTACIÓN DE LA VOZ EN EL TIEMPO Y EN DOMINIOS DE FRECUENCIA.

La señal de voz varía levemente cuando son examinadas en un corto periodo de tiempo (entre 5 y 100msec), estas características son ciertamente estacionarias. Sin embargo, en periodos de tiempo que están en el orden de 1/5 segundos o más, las características de la señal cambian reflejando diferentes sonidos de voz que han sido pronunciados.

Una ilustración de este efecto podemos observar en la figura 1-2 que muestra el tiempo de una onda de sonido correspondiente a la frase: “Es tiempo…”, pronunciado por una persona de género femenino. Cada línea de la onda corresponde a una sección de 200msec.

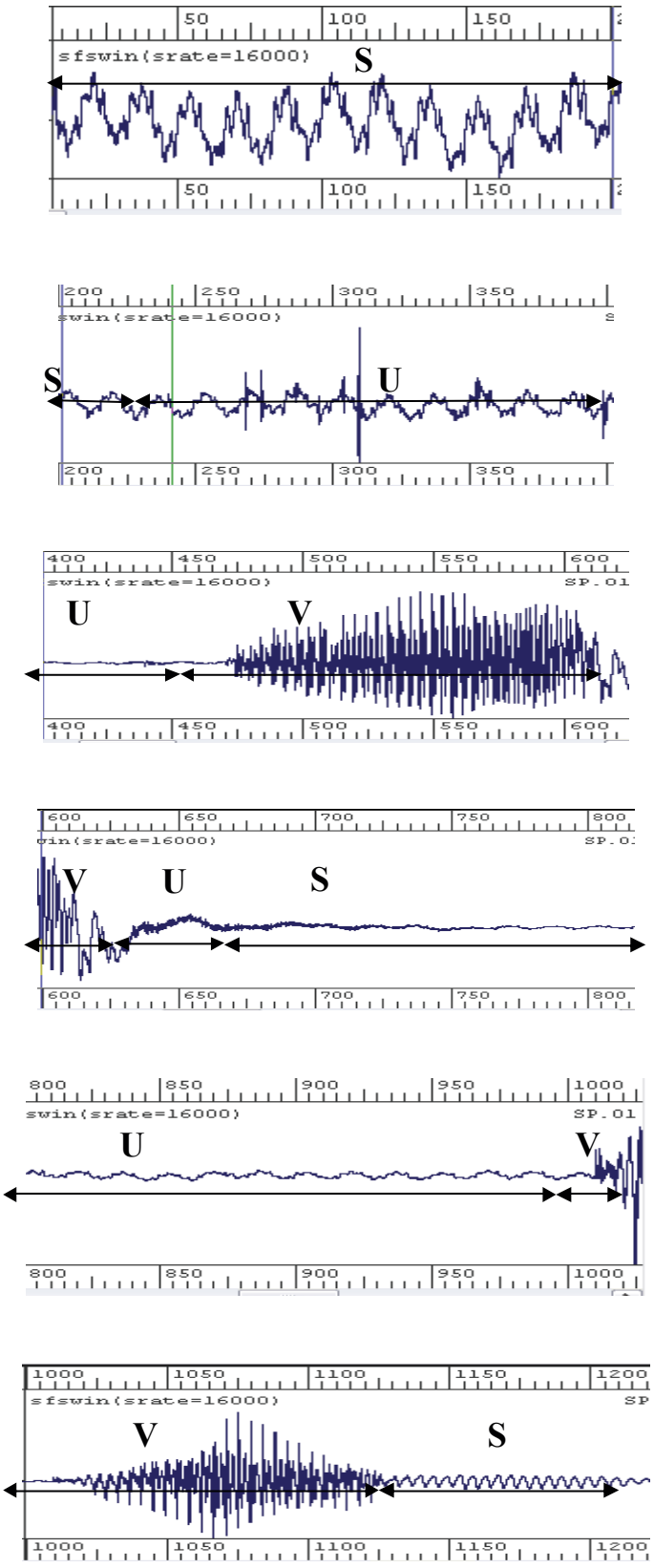


Fig. 1-2: Onda de voz para la pronunciación: "Es tiempo", dividida en secciones de 200ms, gráficos generados por la herramienta SFSWin®.

La señal de voz variando ligeramente en el tiempo puede ser vista en los primeros 200msec de la onda, que corresponde al silencio de fondo, siendo baja en amplitud, para los siguientes 200msec apreciamos un ligero incremento de nivel, seguido de un repentino cambio en la forma de la onda y regularidad, correspondiente a la palabra “ES”, a continuación hay un espacio de silencio, de forma inmediata se produce el proceso de aspiración al que le sigue una variación en la onda representando la palabra “TIEMPO”

Existen varias formas de clasificar los eventos que se producen en una señal de voz, la más simple es por el estado de la fuente de producción de voz -las cuerdas vocales.

La convención para la representación de los diferentes estados son los siguientes:

1. Silencio (S), donde no se produce la voz.
2. Aspiración (U), donde las cuerdas vocales no vibran.
3. Voz (V), se produce la vibración de las cuerdas vocales, generando el habla.

El resultado de aplicar este tipo de clasificación observamos en la figura anterior, inicialmente la onda es clasificada como silencio (S). Un corto periodo de aspiración (U) vemos previo a la generación de voz (V) que corresponde a la vocal inicial de la palabra “Es” seguido de la región de voz tenemos un corto periodo de aspiración, luego una región de silencio, seguido de un relativo periodo extenso de aspiración (U) que corresponde a la pronunciación inicial de la /t/ y, finalmente una región de voz (V) que concierne a la palabra “tiempo”.

La segmentación de la onda definida en regiones (silencio, aspiración y voz) no es exacta, a menudo es difícil distinguir sonidos débiles del silencio, pero sin embargo no es crítico en la segmentación de la señal cuando la precisión es de milisegundos.

Una alternativa de representar la señal y la información asociada a los sonidos es a través del espectrograma, que permiten observar la intensidad de la voz en periodos cortos en diferentes bandas de frecuencia en función del tiempo.

Para obtener un espectrograma se aplica una técnica matemática llamada *análisis de Fourier* a la onda de voz de tal manera que se pueden determinar las frecuencias que están presentes en un momento dado en la señal de voz. El resultado del análisis de Fourier es un espectro, esto se realiza calculando el espectro para una sección de voz típicamente de 5 a 20 milisegundos y así sucesivamente hasta llegar al final de la onda de voz. En general, los espectros cercanos varían levemente reflejando los movimientos del tracto vocal. Existen dos tipos de espectrogramas: de banda ancha y banda estrecha, la elección del ancho de banda es de acuerdo a lo que se quiera analizar:

- Análisis de banda ancha: soluciona la precisión temporal pero pierde precisión en frecuencia.
- Análisis de banda estrecha: soluciona los armónicos, pero emborrona los detalles temporales.

Un ejemplo de este tipo de representación observamos en el siguiente gráfico, un espectrograma de banda ancha en el primer panel, uno de banda estrecha en el segundo panel y la amplitud de la onda de voz para la pronunciación: “No es vencido sino quien cree serlo”, en el tercer panel, la frase es pronunciada por una persona de género femenino.

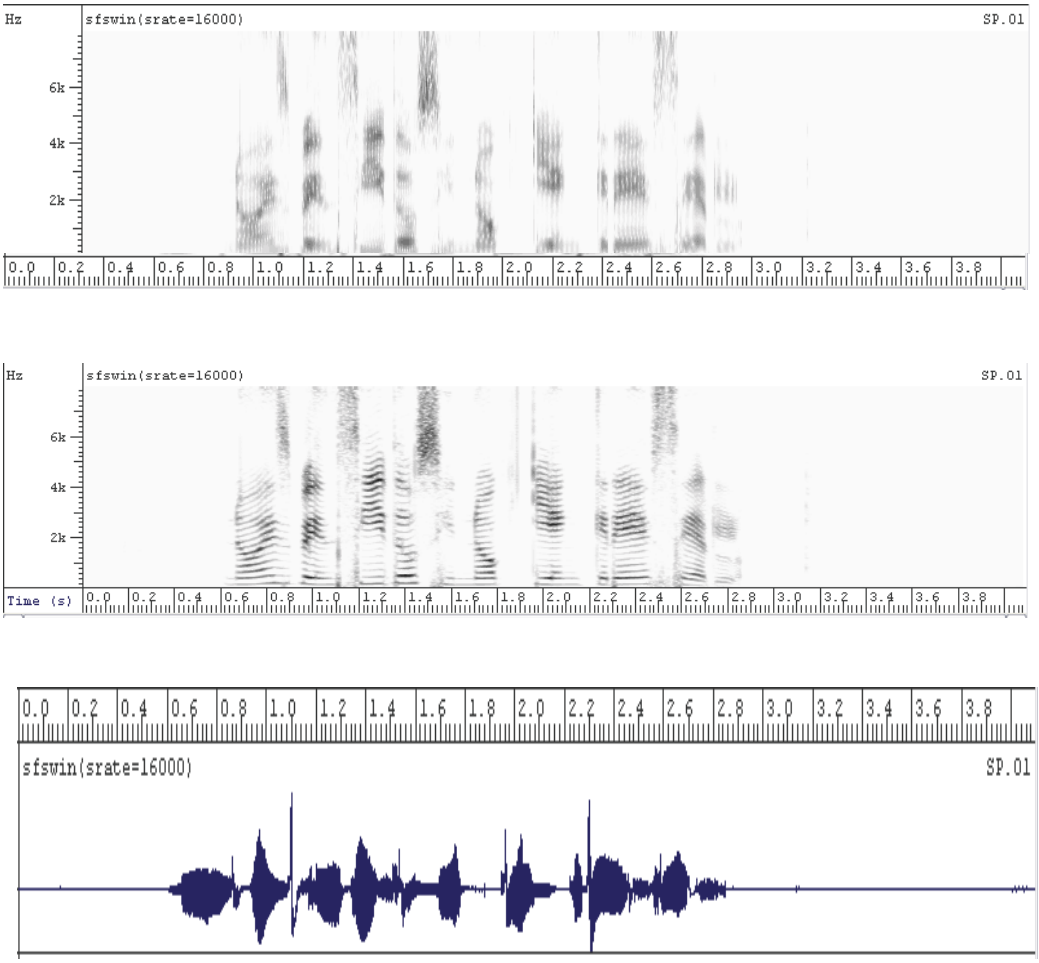


Fig. 1-3: Espectrogramas de banda ancha y estrecha y amplitud de la voz para la pronunciación: “No es vencido sino quien cree serlo.”
Gráficos generados por la herramienta SFSWin®.

En los espectrogramas el eje vertical representa la frecuencia de 0 a 8Khz, mientras que el eje horizontal representa el tiempo y podemos apreciar las variaciones de un espectro a otro.

La intensidad espectral de cada punto en el tiempo está indicada por zonas obscuras para una frecuencia en particular. Los espectrogramas de banda ancha se reconocen fácilmente porque se aprecian estriaciones verticales en las secciones de voz. En cambio en el espectrograma de banda estrecha apreciamos los armónicos individuales correspondientes al tono de voz de la onda. Durante regiones de voz observamos líneas horizontales en el espectrograma. En los periodos de aspiración vemos energía de alta frecuencia mientras que en los periodos de silencio no existe actividad espectral.

1.4.1 TRASPASO DE LA SEÑAL AL DOMINIO DE LA FRECUENCIA

La información más importante del habla se encuentra en sus frecuencias, sin embargo, la señal de voz se toma en tiempo utilizando convertidores analógico/digitales en un proceso denominado muestreo.

Para imitar a la naturaleza debemos aplicar algún mecanismo que traspase a frecuencias la información existente en las muestras de voz que recogemos en el dominio del tiempo. Los mecanismos que se explicarán en capítulos posteriores son: Transformada de Fourier y Método de Predicción Lineal.

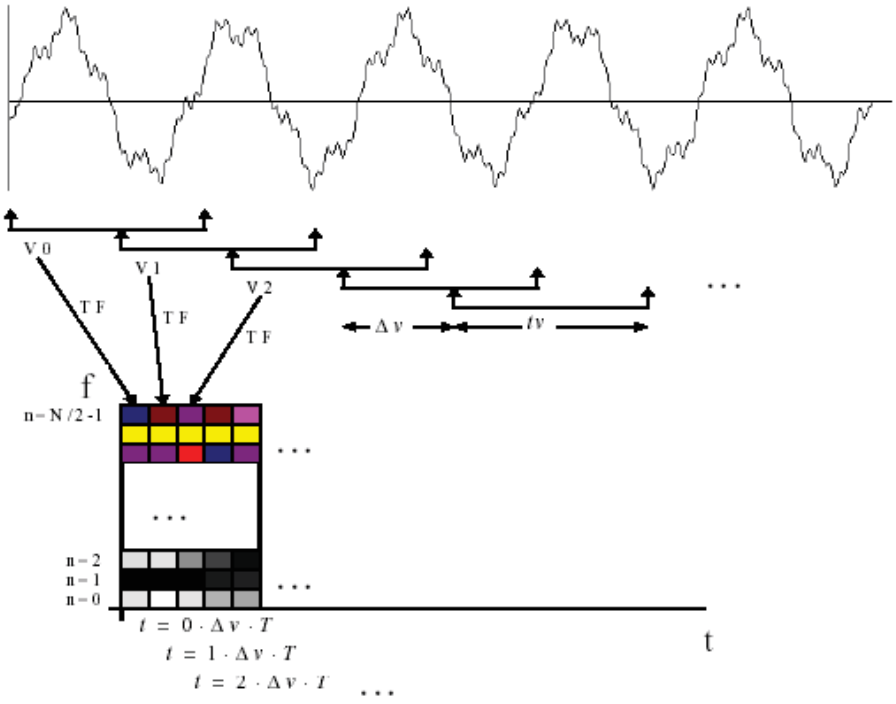


Figura 1-4: Traspaso de la información al dominio de la frecuencia
Fuente: [Ord]

En la figura 1-4 está representada en primer lugar la señal de voz en el dominio del tiempo. Las muestras de la señal se dividen en secciones solapadas[6] de igual tamaño denominadas ventanas y sobre cada conjunto de muestras correspondientes a una ventana se aplica la Transformada de Fourier, u otro método de traspaso a frecuencias.

Si el tamaño de la ventana es de N muestras, entonces se consigue $N/2$ valores de frecuencias. Por lo tanto se obtendrá tantos conjuntos de $N/2$ frecuencias (columnas en el gráfico) como ventanas en el dominio del tiempo.

[6] Que se encuentran cubiertas de forma parcial por otras secciones de voz.

1.5 FISIOLOGIA Y FUNCIONALIDAD DEL APARATO FONADOR

1.5.1 Introducción

El aparato fonador humano es un sistema fuertemente ligado con la función de respiración, que utiliza recursos comunes con la misma.

Se compone de 3 cavidades:

- Cavidades infraglóticas
- Cavidad glótica
- Cavidades supraglóticas

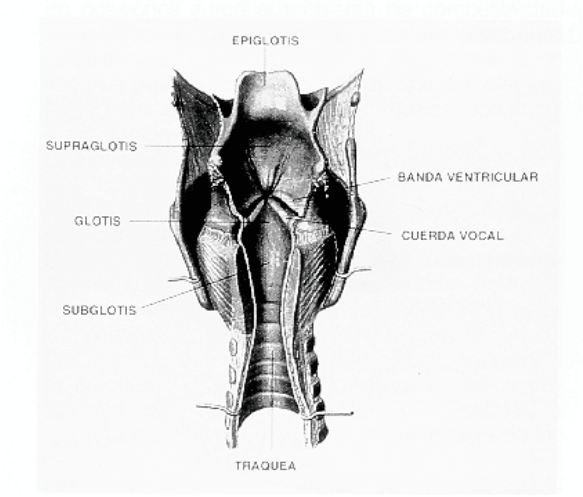


Fig. 1-5: Cavidades que componen el aparato fonador
Fuente: [Asu]

- **CAVIDADES INFRAGLÓTICAS.**

Se denomina cavidades infraglóticas a los espacios existentes por debajo de la laringe, es decir la tráquea, los bronquios y los pulmones, la función de estas cavidades es proporcionar la corriente de aire necesaria para producir el sonido.

- **CAVIDAD GLÓTICA.**

Está formada por la laringe. La característica más importante es la presencia de las cuerdas vocales, que son responsables de la producción de la vibración, acción básica para la generación de la voz.

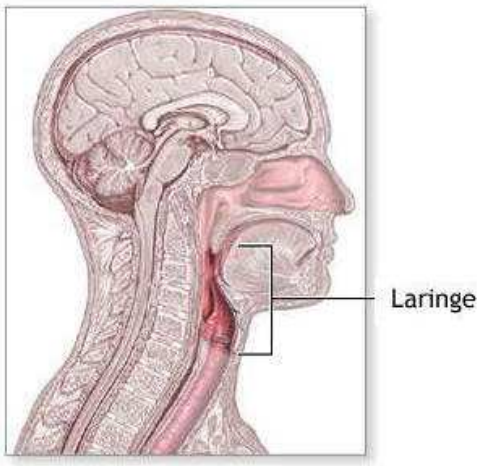


Fig. 1-6: Ubicación de la laringe en el aparato fonador
Fuente: [Fon]

La abertura generada entre las cuerdas vocales se denomina glotis, cuando éstas se encuentran separadas, la glotis adopta una forma triangular. El aire pasa libremente y prácticamente no se produce sonido, éste es el caso de la respiración.

- **CAVIDADES SUPRAGLÓTICAS.**

Se denomina cavidad supraglótica a la porción que incluye las cavidades faríngea, oral y nasal junto con los elementos articulatorios, los mismos que son órganos del canal oral que intervienen en la articulación de los sonidos del lenguaje y son: boca, lengua y paladar

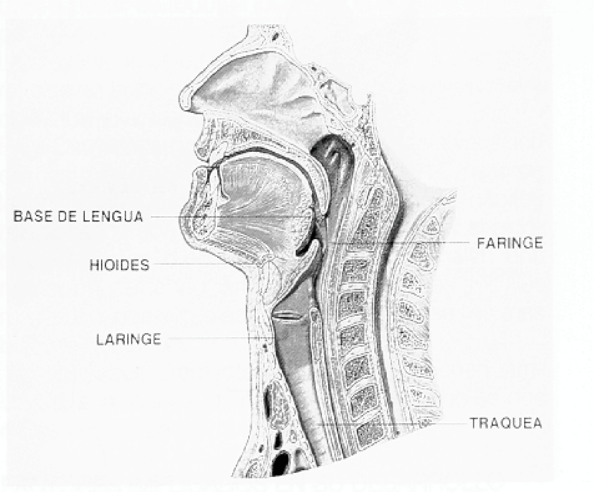


Fig. 1-7: Estructura del aparato fonador
Fuente: [As1]

1.6 FONÉTICA ACÚSTICA

Estudia los fenómenos relacionados con las características físicas del sonido y su transmisión. Este análisis se realiza a 4 niveles:

A) ***Nivel fonológico***.- Se encuentra conformado por el fonema que es la unidad básica e indivisible de nuestro lenguaje, el conjunto de los fonemas se establece por oposición, es decir, si se cambia un sonido en una palabra y la palabra cambia de significado, al sonido se le considera fonema.

B) ***Nivel morfosintáctico***.- Se estudian las palabras estableciendo su género, número, tiempo y las relaciones entre ellas.

C) ***Nivel Semántico***.-.Se refiere a la comprensión del lenguaje, este nivel no sólo trata la relación de las palabras, sino de las oraciones, por lo que el hablante debe contar con un amplio conocimiento del medio para entender y relacionar el mensaje con dichos conocimientos.

D) ***Nivel Pragmático***.- Está relacionado con todos los aspectos que rodean a la conversación como: el tono de voz (agudo - grave), la intensidad, el ritmo (pausado, lento, rápido).

1.6.1 Clasificación de los sonidos de la voz

Los sonidos emitidos por el aparato fonador se clasifican de acuerdo a los siguientes aspectos del fenómeno de emisión:

- Carácter vocálico o consonántico.
- Oralidad o nasalidad
- Carácter tonal (sonoro) o no tonal (sordo)
- Lugar de articulación
- Modo de articulación
- Posición de los órganos articulatorios
- Duración

1.6.1.1 Vocales y consonantes

Las *vocales* son los sonidos generados por las cuerdas vocales sin ningún obstáculo entre la laringe y las aberturas oral y nasal. 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. En el castellano las vocales pueden constituir palabras completas, no así las consonantes.

1.6.1.2 Oralidad y nasalidad

Los fonemas en los cuales el aire pasa por la cavidad nasal se denominan nasales, mientras que aquellos en los que sale por la boca se denominan orales. En castellano son nasales sólo las consonantes **"m", "n", "ñ"**

1.6.1.3 Tonalidad

Los fonemas en los que interviene la vibración de las cuerdas vocales se denominan *tonales* o *sonoros*, todas las vocales son tonales, pero existen varias consonantes que también lo son: **"b", "d", "m"**.

Los fonemas que se generan sin producir vibración en la glotis se denominan *sordos*, algunos de ellos son el resultado de la agitación causada por el aire pasando a gran velocidad por un espacio reducido, como las consonantes **"s", "z", "j", "f".**

1.6.1.4 Lugar y modo de articulación (consonantes)

PUNTO DE ARTICULACIÓN

Indica el lugar de las cavidades supraglóticas donde se produce la articulación del fonema. De acuerdo con los diferentes puntos de articulación se pueden distinguir los siguientes fonemas[7]:

☆ **CONSONANTES:**

Bilabiales.- Contactan los labios superiores e inferiores. /p, b, m/

[7] Gráficos tomados de : [Pho]

/p/

/b/

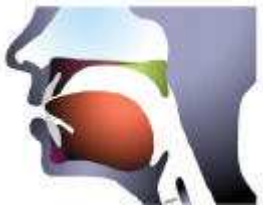
/m/

Labiodentales.- Contacta el labio inferior con los incisivos superiores. /f/

/f/

Linguodentales.- Contacta el ápice de la lengua con los incisivos superiores /t, d/

/t/



/d/

/n/

/l/

Linguointerdentales.- Se sitúa el ápice de la lengua entre los incisivos superiores e inferiores. /θ/

/θ/

Linguoalveolares.- Contacta el ápice o predorso de la lengua con los alvéolos. /l, s, m, n, r/

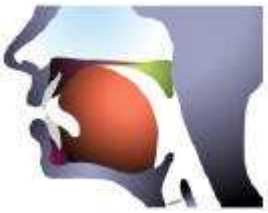
/l/

/s/

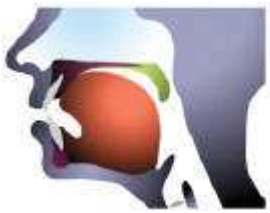
/n/

/r/

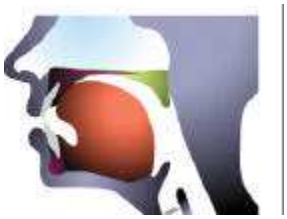
Linguopalatales- Contacta e predorso de la lengua con el paladar duro.



/tf/



/n/



/l/

Linguovelares.- Se aproxima o toca el postdorso de la lengua con el velo del paladar./x,k,g/

/k/

/g/

/x/

1.6.1.5 Posición de los órganos articulatorios (vocales)

En las vocales, la articulación se basa en la modificación de la acción filtrante de los diversos resonadores, esto depende de las posiciones de la lengua (elevación, profundidad o avance), mandíbula inferior, labios y paladar blando.

Las vocales se clasifican tomando en cuenta dos aspectos:

Modo de articulación (Formante 1)

- Cerradas (i, u)
- Medias (e, o)
- Abiertas (a)

Lugar de articulación (Formante 2)

- Anteriores (i, e)

/i/

/e/

- Centrales (a)

/a/

- Posteriores (o, u)

/o/

/u/

Donde los formantes o componentes acústicos F1 y F2 son resonancias del tracto vocal, que por las dimensiones y velocidad de propagación del sonido aparecen generalmente 1 formante por cada kHz.

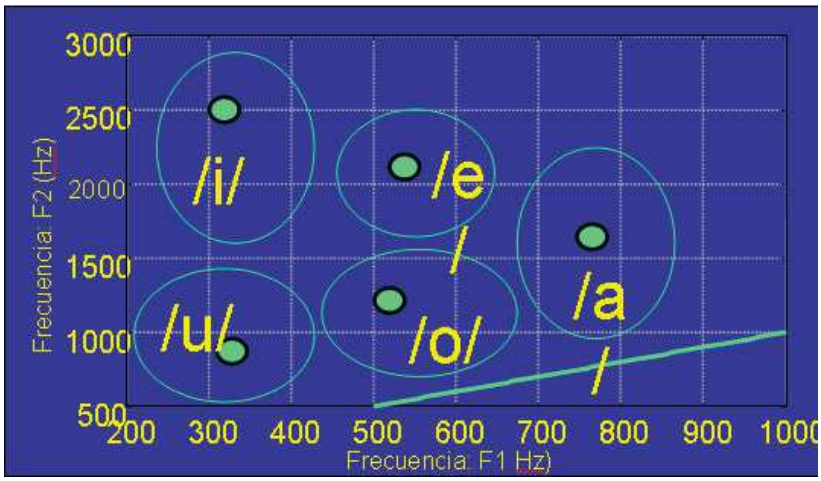


Fig. 1-8: Formantes 1 y 2 en las vocales

Las vocales se diferencian de las consonantes por una estabilidad de sus componentes acústicos. Todos los cambios apreciables en la frecuencia de los formantes, excepto aquellos que aparecen en la unión de dos vocales, contribuyen a la percepción de las consonantes. Las vocales son los sonidos más fácilmente detectables, prácticamente constituyen la mitad de los sonidos de una lengua en porcentaje de aparición.

Fonema	Frecuencia relativa de ocurrencia	Fonema	Frecuencia relativa de ocurrencia	Fonema	Frecuencia relativa de ocurrencia
/e/	14.67%	/s/	8.32%	/θ/	1.45%
/a/	12.19%	/N/	4.86%	/g/	0.94%
/o/	9.98%	/t/	4.53%	/x/	0.57%
/i/	7.38%	/d/	4.24%	/f/	0.55%
/u/	3.33%	/l/	4.23%	/r/	0.43%
Total	**47.55%**	/k/	3.98%	/j/	0.41%
		/ɾ/	3.26%	/ʎ/	0.38%
		/m/	3.06%	/c/	0.37%
		/n/	2.78%	/D/	0.31%
		/p/	2.77%	/G/	0.28%
		/b/	2.37%	/ɳ/	0.25%
		/R/	1.93%	/B/	0.03%
				Total	**52.30%**

Fig. 1-9: Frecuencia relativa de ocurrencia para distintos sonidos.
Fuente: [Acu]

1.6.1.6 Duración

La duración de los sonidos, especialmente de las vocales, no es tomado en cuenta a nivel semántico en el castellano, pero sí en el plano expresivo, a través de la agogia[8]. De forma contraria en el idioma inglés, la duración de una vocal puede cambiar completamente el significado de la palabra que la contiene.

[8] Énfasis o acentuación de los sonidos

1.6.1.7 Fonemas y Grafías del español

FONEMA	GRAFÍA	EJEMPLOS
/a/	a	
/b/	b,v	vaso, bote
/	c,z	cena, caza
/k/	c,qu,k	casa, queso, kilo
/ *tf* /	ch	chico, muchacho
/d/	d	dado
/e/	e	
/f/	f	fama, café
/g/	g,gu	gama, guisa, paga
/i/	i	
/x/	j,g	paja, gitano
/l/	l	ala, mal
/ λ /	ll	llave, calle
/m/	m	mamá
/n/	n	nana
/ η /	ñ	caña
/o/	o	
/p/	p	piedra , capa
/r/	r	para, norte
/ $\overline{r}$ /	rr,r	perro ,reno
/s/	s	soy, dos
/t/	t	tapa, atar
/u/	u	
/ *j* /	y,hi	mayo, la hierba

CAPITULO II

FUNDAMENTOS, ARQUITECTURA Y PROCESO DE RECONOCIMIENTO DE VOZ

2.1 INTRODUCCION

En el presente capítulo revisaremos brevemente la evolución de los sistemas informáticos de reconocimiento de voz, señalando los factores que influyen directamente en la calidad de un reconocedor. Expondremos su arquitectura, la función de cada uno de sus componentes y analizaremos el proceso de reconocimiento de voz y las diferentes técnicas que se aplican en cada fase de éste.

2.2 EVOLUCIÓN DE LOS SISTEMAS DE RECONOCIMIENTO DE VOZ

Esta sección se basa en su totalidad en el primer capítulo de [RabJua93].

2.2.1 Primeros sistemas: Dispositivos electrónicos.

En 1952, en los laboratorios Bell, K. Davis, R. Biddulph y S. Balashek crearon un sistema electrónico que permitía identificar para un solo hablante pronunciaciones de los 10 dígitos realizadas de forma aislada. El fundamento de esta máquina se basaba en medidas de las resonancias espectrales del tracto vocal para cada dígito. Las medidas se obtenían aplicando bancos de filtros analógicos.

En 1959, en la University College de Londres, P. Denes trataba de desarrollar un sistema para reconocer 4 vocales y 9 consonantes. El aspecto más novedoso era el uso de información estadística acerca de las secuencias válidas de fonemas en inglés.

Hasta este momento, todos los sistemas son dispositivos electrónicos. Los primeros experimentos de reconocimiento desarrollados en ordenadores tienen lugar al final de los años 50's y comienzo de los 60's.

2.2.2 Años 60: Sistemas con hardware específico.

Durante estos años se inician 3 proyectos que modifican el curso de la investigación en el área del reconocimiento de voz de manera notable.

El primero de ellos fue realizado por: T. Martin, A. Nelson y H. Zadell en los RCA Laboratories (1964). Como consecuencia de este trabajo los autores diseñaron un conjunto de métodos elementales de normalización en el tiempo, que se basaban en la detección fiable de los puntos de principio y fin de discurso. De esta forma conseguían reducir la variabilidad en las tasas de reconocimiento.

En la Unión Soviética, T. K. Vintsyuk, propone la utilización de métodos de programación dinámica para conseguir el alineamiento temporal de pares de realizaciones de habla.

El tercer trabajo lo realiza D. R. Reddy (Stanford University, 1966) en el campo del reconocimiento de habla continua mediante el seguimiento dinámico de fonemas. La aplicación de sus ideas concluye con el reconocedor de oraciones, dependiente del hablante, para un vocabulario de 561 palabras.

2.2.3 Años 70. Sistemas de reconocimiento de palabras aisladas

Los años 70 representan un periodo muy activo para este campo, distinguiéndose dos actividades principales:

- *El reconocimiento de palabras aisladas.*

- *Primeros intentos de construir reconocedores de habla continua y de grandes vocabularios.*

El reconocimiento de palabras aisladas comienza a ser viable en la práctica, como consecuencia de los trabajos realizados por: V.M. Velichko y N.G. Zagoruyko, quienes contribuyeron al avance del uso de procedimientos de encaje de patrones en el campo del tratamiento de la voz.

El grupo japonés *H. Sakoe y S. Chiba* estableció de manera formal los algoritmos que fundamentados en la —programación dinámica[9], — podían aplicarse a la resolución de este tipo de problemas.

Por último los trabajos de F. Itakura mostraban cómo los principios de las técnicas LPC (Linear Predictive Coding) podían extenderse al reconocimiento del habla.

Es en este momento cuando se advierte que el conocimiento sintáctico, semántico y contextual[10] son fuentes de información que permiten reducir el número de posibles alternativas que todo sistema automático de diálogo hombre-máquina debe considerar.

El sistema Hearsay I, construido por la CMU (Carnegie Mellon University) en 1973 era capaz de emplear información de tipo semántico para reducir el número de posibles alternativas que el reconocedor debía evaluar.

Otro hito durante esta década es el comienzo de los trabajos del grupo investigador de I.B.M., dedicado al dictado automático por voz para grandes vocabularios. Finalmente, en los AT&T Bell Labs (ahora Bell Labs, Lucent

[9] Ver anexo III
[10] Referente al contexto. Unión de cosas que se enlazan y entretejen.

Technologies y AT&T Labs-Research), los investigadores comenzaron una serie de experimentos orientados a conseguir reconocedores realmente independientes del locutor para su uso en aplicaciones telefónicas.

Al final de este periodo, la implementación de sistemas de reconocimiento de voz se ve favorecida por la disponibilidad de tarjetas microprocesador.

2.2.4 Años 80. Construcción de sistemas de habla continua

Los años 80 se caracterizan por la generalización en la construcción de sistemas de reconocimiento de habla continua, siendo capaces de tratar con cadenas de palabras pronunciadas de manera fluida.

El giro metodológico que se produce como consecuencia de pasar de métodos basados en comparación de plantillas a los métodos basados en modelos estadísticos debido a la extensión en el uso de los modelos ocultos de Markov o HMM (Hidden Markov Models).

Estos métodos habían sido desarrollados en la década pasada para tratar con problemas de habla continua, pero su aceptación generalizada no sucedió hasta unos 10 años después. A partir de entonces se han desarrollado numerosas mejoras y actualmente constituyen los mejores modelos disponibles para capturar y modelar la variabilidad presente en el habla.Muchas de las contribuciones durante este periodo y el principio de los años 90`s, provienen de los esfuerzos de la CMU a través de su sistema SPHINX.

2.2.5 Años 90. Mejora y diversificación

La década de los 90`s supone en cierta manera la continuidad en los objetivos ya propuestos, ampliando el tamaño de los vocabularios a la vez

que se diversifican los campos de aplicación; los servicios sobre la línea telefónica son de los que más atención acaparan en la actualidad.

En estos últimos años ha crecido el interés por el estudio de procesos de reconocimiento de voz en condiciones de ruido.

2.3 TECNOLOGÍAS DEL HABLA

Las tecnologías del habla están clasificadas de la siguiente manera:

- **Reconocimiento automático del habla**. Proporciona a las máquinas la capacidad de recibir mensajes orales tomando como entrada la señal acústica recogida por un micrófono. El proceso de reconocimiento automático del habla tiene como objeto decodificar el mensaje contenido en la onda acústica para realizar una acción.
- **Conversión texto a voz**. Proporciona a las máquinas la facultad de generar mensajes orales no grabados previamente, tomando como entrada un texto. Los sistemas de conversión texto-voz realizan el proceso de lectura de forma clara y con una voz lo más natural y humana posible.
- **Codificación de voz y audio**. Se ocupa de la transmisión y almacenamiento de la señal de voz en forma digital, de manera eficiente, sin pérdida de la calidad y con el menor número de bits por muestra.

2.4 FUNDAMENTOS DE RECONOCIMIENTO DE VOZ

El reconocimiento del habla parece natural y sencillo para las personas, en primera instancia se pensó que sería fácil incorporarlo a las máquinas, sin embargo cuando se iniciaron los estudios y primeras investigaciones sobre

el tema se comprobó que no es así, debido a la existencia de diversos factores que determinan la complejidad del reconocimiento del habla:

- **El locutor:** Básicamente se refiere a que una persona no pronuncia siempre de la misma manera, debido a situaciones físicas y psicológicas, además de la diversidad de locutores (niños, mujeres, hombres), diferencias en la edad o región de origen de la persona.
- **La forma de hablar:** El hombre pronuncia las palabras de una forma continua, y debido a la inercia de los órganos articulatorios, que no pueden moverse instantáneamente, se producen efectos coarticulatorios. Ello, unido a las variaciones introducidas por la prosodia, hace que una palabra al principio de una frase sea diferente que cuando se dice en medio, o que sea diferente dependiendo de que es lo que le precede.
- **El vocabulario:** Es el número de palabras diferentes que debe reconocer el sistema, mientras mayor es el número de palabras se incrementa la complejidad del reconocedor por dos motivos:
    - El primero, al aumentar el número de palabras es más fácil que aparezcan palabras parecidas entre sí,
    - Segundo, el tiempo de procesamiento aumenta al incrementar el número de palabras con las que se realiza la comparación.

  Una solución a este problema sería utilizar unidades lingüísticas inferiores a la palabra (alófonos[11], sílabas[12], demisílabas[13]) que en

[11] Diferentes realizaciones de un mismo fonema según el entorno en el que esté situado. El significado de la palabra no cambia por el intercambio de alófonos.
[12] Letra vocal, o conjunto de letras, que se pronuncia en una sola emisión de voz

principio tienen un número limitado, e inferior al de las posibles palabras. Sin embargo, la dificultad de reconocer estas unidades es aun mayor debido a que su duración es muy corta, la frontera entre dos unidades sucesivas es muy difícil de establecer y los efectos coarticulatorios son mucho más fuertes que entre palabras.

- **La Gramática:** Es el conjunto de reglas que limita el número de combinaciones permitidas de las palabras del vocabulario. La existencia de la gramática en un reconocedor mejora la tasa de reconocimiento y elimina ambigüedades.
- **El Entorno Físico:** Es un factor importante al momento de definir el ambiente en el que debe operar el reconocedor, porque no es lo mismo

que un sistema funcione en un ambiente poco ruidoso, como puede ser el despacho de un médico, o por ejemplo, el que debe funcionar a través de la línea telefónica, con la consiguiente reducción de banda.

2.4.1 Principios básicos de un sistema de reconocimiento de voz

Un reconocedor de voz es un programa que trata de entender o decodificar una señal digital. El problema del reconocimiento de voz se puede abordar desde un punto de vista estadístico, debido a la variabilidad en la forma como los humanos producimos los sonidos. Esta variabilidad depende del hablante y del ambiente en que son producidos bs sonidos de voz. Sin embargo, el problema se resuelve encontrando la probabilidad de que una

[13] Ver anexo IV

palabra W haya sido pronunciada, dado que se sabe como una palabra en particular debe sonar,

$$P(W \mid A),$$

es decir, la probabilidad de que se pronuncie W conociendo las medidas de su acústica A. Aplicando el —teorema de Bayes[14]—, tenemos que:

$$P(W \mid A) = \frac{P(A \mid W) P(W)}{P(A)},$$

donde

$P(A|W)$: es la probabilidad de las medidas acústicas A cuando se pronuncia la secuencia de palabras W. Representa el modelo acústico.

$P(W)$: es la probabilidad de la secuencia de palabras W que ha ocurrido. Representa el modelo del lenguaje.

$P(A)$: es la probabilidad de las medidas acústicas A.

La fórmula fundamental del reconocimiento automático del habla es la siguiente:

$$Wm = \max_{W} P(A|W)\ P(W)$$

Es decir, la secuencia de palabras reconocida es aquella que maximiza el producto de dos probabilidades, una $P(A|W)$ que relaciona los datos acústicos con la secuencia de palabras y que denominaremos ***modelo***

[14] [Bay]

acústico y $P(W)$ que únicamente depende de la secuencia de palabras y que denominaremos ***modelo de*** **lenguaje**[15].

2.5 ARQUITECTURA DE UN SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DE VOZ

Los componentes básicos de un sistema de reconocimiento automático se muestran en el siguiente gráfico:

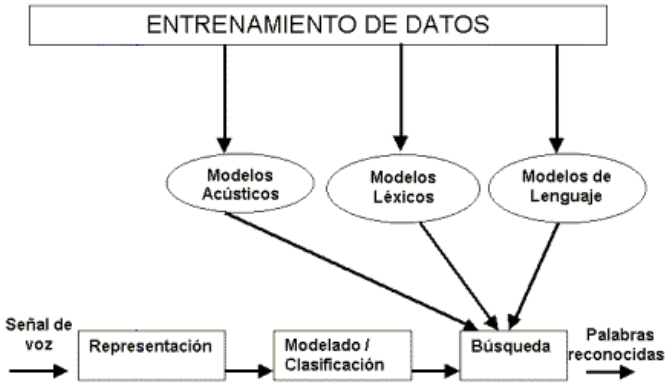


Fig.2-1: Bloques básicos de un sistema de reconocimiento automático del habla
Fuente: [Inv1]

En la figura 2-1 se distinguen dos procesos claramente diferenciados:

- **Entrenamiento:** El sistema aprende a partir de muestras de voz y texto, tomadas de los modelos acústicos $P(A|W)$ y de lenguaje $P(W)$.
- **Reconocimiento:** Fase propiamente dicha de reconocimiento automático del habla en la que la señal acústica es transcrita en una secuencia de palabras de acuerdo con la fórmula fundamental del RAH[16].

[15] [Rah]

[16] Reconocimiento automático del habla.

2.5.1 Modelos Acústicos:

La acústica se define como “La ciencia que estudia la producción, control, transmisión, recepción y efectos del sonido”. Los modelos usados por un reconocedor de voz para la decodificación reflejan esta definición y por ello son conocidos como *modelos acústicos*, digitalmente modelan las características (medidas numéricas) de los sonidos que requiere el reconocedor para el proceso de decodificación.

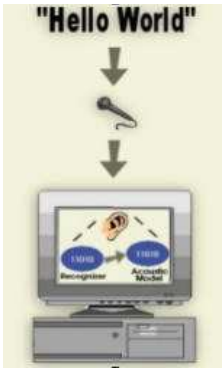


Fig. 2-2.: Interacción del modelo acústico con el reconocedor de voz.

Fuente: [Fun]

La tarea del reconocedor es determinar qué palabras se han pronunciado comparando las medidas acústicas del idioma hablado con las medidas contenidas en el modelo acústico aplicando la técnica probabilística conocida como: Modelos ocultos de Markov.

2.5.2 Modelos Léxicos:

El modelo léxico define un vocabulario de pronunciación (el diccionario) para todas las palabras de la aplicación. Las pronunciaciones se definen normalmente en términos de sílabas. Estas son las unidades que nosotros vamos a intentar modelar a través de un proceso de entrenamiento. Las pronunciaciones

múltiples son permitidas en las palabras. Sin embargo, deben definirse pronunciaciones múltiples para la misma palabra.

Los modelos léxicos incluyen la definición del vocabulario y la pronunciación de las palabras.

2.5.3 Modelos de Lenguaje:

El procesamiento natural del lenguaje (NLP, por sus siglas en ingles) provee otra fuente de conocimiento que necesita el reconocedor, *el modelo del lenguaje*. Mientras los modelos acústicos se construyen a partir del proceso de extracción de características que permiten al reconocedor descifrar los fonemas que comprenden las palabras, el modelo del lenguaje especifica el orden en que es probable que ocurra la sucesión de palabras. Por ejemplo, un saludo típico podría ser una interjección como "Hola", seguido de un sustantivo "Mundo"[17]. El gráfico ilustra el proceso de reconocimiento de voz incorporando el modelo del lenguaje para representar el saludo.

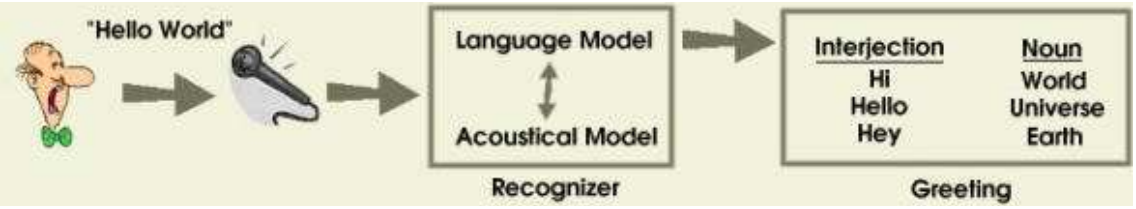


Fig. 2-3.: Modelo del lenguaje.
Fuente: [Pro]

Los Modelos del Lenguaje que se emplean en un sistema de reconocimiento de voz son: N-grams y Networks.

[17] [Mss]

2.6 PROCESO DE RECONOCIMIENTO DE VOZ

El proceso de reconocimiento automático de voz consiste en:

1. Obtener y digitalizar la señal de voz.
2. Extraer un conjunto de características esenciales de la señal.
3. Introducir las características a un clasificador.
4. Realizar un algoritmo de búsqueda para encontrar la secuencia permitida más probable utilizando la salida obtenida y una red de pronunciaciones.
5. Encontrar la(s) palabra(s) que se desea reconocer.

2.6.1 Obtención y digitalización de la señal de voz:

El proceso de reconocimiento de voz empieza cuando una persona habla frente al micrófono, el acto de hablar produce una onda de presión que forma una señal acústica, el micrófono recibe la señal acústica y la convierte en una señal analógica que puede ser entendida por dispositivos electrónicos, finalmente la señal analógica es convertida en una señal digital a través de la tarjeta de sonido.

Fig. 2-4: Captura de la señal acústica.
Fuente: [Pro1]

La calidad de la onda capturada depende de varios elementos:

- la calidad de micrófono utilizado,

- la calidad del convertidor analógico/digital
- y el número de bits utilizados para el almacenamiento de cada muestra.

A nivel informático, el parámetro que podemos modificar es el número de bits por muestra; 8 bits sería una calidad aceptable y 16 bits sería la calidad de un disco compacto de música.

El rango de frecuencias que puede percibir el oído humano varía habitualmente desde los 20Hz hasta 20.000Hz. Para una captura perfecta se debe utilizar una frecuencia de muestreo superior a 40.000Hz.

La frecuencia máxima que se genera en el habla, en general, está por debajo de los 8.000Hz por lo tanto, para capturar un mensaje hablado será suficiente utilizar la frecuencia de muestreo estándar de 22.050Hz.

2.6.2 Extracción de características:

La señal capturada por el micrófono contiene información que el reconocedor aún no puede decodificar. Ciertamente los atributos o características de la forma de pronunciar de una persona son de gran ayuda para la decodificación. Estas características permiten al reconocedor diferenciar los fonemas que hayan sido pronunciados. Normalmente este proceso involucra realizar un análisis espectral de la señal.

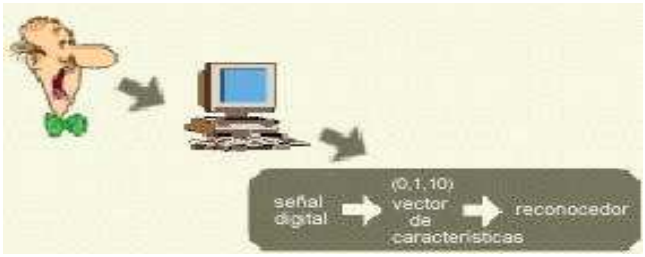


Fig. 2 -5.: Proceso de extracción de características.
Fuente: [Pro2]

Al conjunto de parámetros resultante de esta extracción, se le denomina vector de características.

2.6.3 Introducción de las características al clasificador:

El clasificador aplica un modelo probabilístico al vector de características y vincula a cada uno con alguna unidad lingüística (palabra, fonema u otra unidad específica).

La salida del clasificador proporciona por cada fonema la probabilidad de que dicho fonema es lo que generó el vector de características.

Las dos técnicas más utilizadas en el proceso de clasificación son: cadenas ocultas de Markov y redes neuronales.

2.6.4 Aplicación de algoritmos de búsqueda:

Con la matriz de probabilidades obtenida se procede a realizar una búsqueda para encontrar la secuencia de fonemas con mayor probabilidad de reconocimiento, el algoritmo Viterbi determina la secuencia de palabras más probable usando información estadística de las duraciones máximas y mínimas de cada fonema, con el objetivo de

restringir las opciones. A medida que los límites de las duraciones para cada fonema se afinan, el nivel de reconocimiento se incrementa.

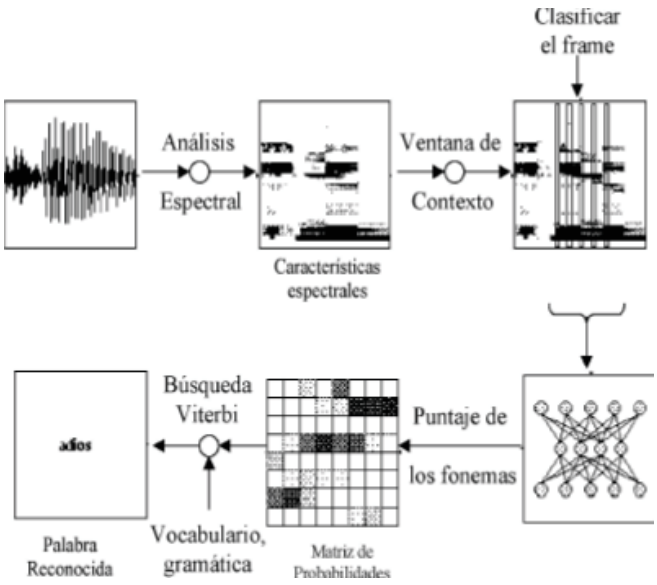


Fig. 2-6: Proceso de reconocimiento de voz
Fuente: [Reco]

CAPITULO III

PROCESAMIENTO DE LA SEÑAL Y METODOS DE ANALISIS APLICADOS AL RECONOCIMIENTO DE VOZ

3.1 INTRODUCCION

Los sistemas de reconocimiento de voz comprenden una colección de algoritmos y una gran variedad de disciplinas tales como: teoría de comunicación, lingüística, psicología, matemática, entre otras. Aunque cada una de estas áreas puede diferir, para cada reconocedor, el denominador común de todos los sistemas de reconocimiento es el procesamiento de la señal, el cual convierte la onda de voz en algún tipo de representación paramétrica (generalmente reduciendo la cantidad de información) para su posterior análisis y procesamiento.

En este capítulo, basado en [Mariñ99] y [RabJua93], explicaremos los fundamentos del procesamiento digital de señales y los métodos dominantes de análisis espectral empleados en el reconocimiento de voz: Transformada de Fourier, Análisis por banco de filtros y codificación por predicción lineal (LPC).

3.2 FUNDAMENTOS DEL PROCESAMIENTO DE SEÑALES

El procesamiento digital de señales —consiste en el análisis y modificación de la información, la cual es medida en una secuencia discreta de números. Este proceso es realizado por un DSP (digital signal processor).

Los DSP son poderosos microprocesadores capaces de procesar mucha información en tiempo real como señales de radio, sonido o video. En

computadores personales, los DSP están presentes en comunicaciones y en el procesamiento de audio profesional en tiempo real[18].

3.2.1 Señales analógicas y digitales

Una señal es analógica porque puede tomar una infinidad de valores a lo largo del tiempo es decir, no hay una cantidad mínima de variación. Sin embargo una señal digital puede variar sólo de una manera determinada[19]. Por ejemplo, si se establece una tensión igual a 5 voltios para representar un "1" lógico y 0 voltios para representar un "0" lógico, entonces la señal idealmente sólo puede tomar dos valores 0 y 5. En el caso anterior la señal analógica podría valer 0.1232, 4.5323254, 5.22212121, etc. Expliquemos gráficamente el concepto de analógico y digital:

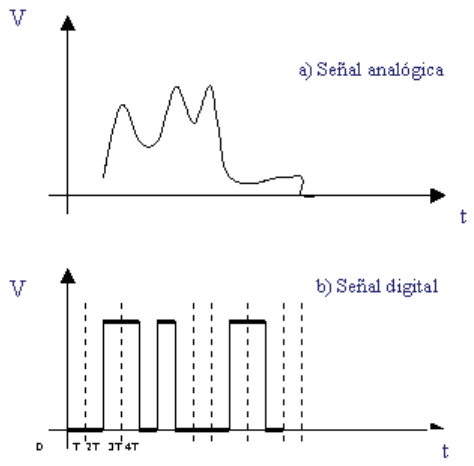


Fig. 3-1: El eje vertical representa los voltajes y el eje horizontal el tiempo. La señal digital de abajo presenta transiciones cada T segundos, es decir cada bit dura T segundos.

[18] [Bas]
[19] [Sig]

Para observar las ventajas de la señal digital sobre la analógica hemos añadido otra señal que representa la original pero modificada por el ruido e interferencias. Resulta imposible arreglar la señal analógica, pero en la señal digital básicamente los efectos indeseables serán casi eliminados por completo, la razón es simple, en caso de la señal digital sabemos a priori cómo tiene que ser la forma de onda y esto nos permite arreglar cierto grado de defectos. Sin embargo, en el caso de la señal analógica, como no sabemos cómo era la señal antes de sufrir los efectos indeseados, éstos no pueden ser corregidos. Entonces las ventajas de la señal dgital vienen porque sabemos qué forma debe tener y si no es así, enseguida detectamos que ha habido algún error.

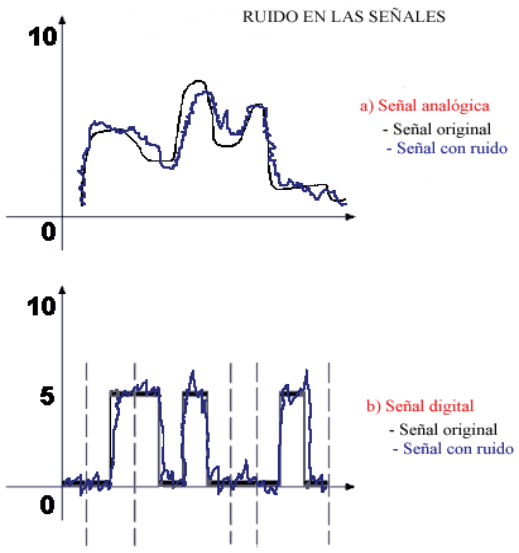


Fig. 3-2: Ruido sobre señales analógicas y digitales

Las técnicas de detección y corrección de errores en señales digitales son extremadamente avanzadas. Por ejemplo, es normal que se pase 1 bit erróneo de cada 100.000.000.

3.2.2 Conversión de señales analógicas a digitales

La conversión de señales analógicas a digitales consta de varios procesos:

- Muestreo
- Cuantización
- Codificación

3.2.2.1 Muestreo

La técnica de muestreo —consiste básicamente en tomar valores de la señal analógica o continua en el tiempo en instantes suficientemente cercanos entre sí de tal manera que se puedan obtener las variaciones de la onda[20].

Frecuencia de muestreo:

La Frecuencia de Muestreo es el número de veces por segundo en los que se toman valores discretos de la señal analógica para pasarla a un formato digital, medida en Hertz (Hz).

$$1Hz = 1/seg.$$

La frecuencia o razón de muestreo determina el rango de frecuencias (Ancho de banda) de un sistema. A mayor frecuencia de muestreo habrá más calidad o precisión.

Teorema de Nyquist:

Desarrollado por H. Nyquist, en este teorema se afirma que "una señal analógica puede ser reconstruida, sin error, de muestras tomadas en

[20] [Mac]

iguales intervalos de tiempo. La razón de muestreo debe ser igual, o mayor, al doble de su ancho de banda de la señal analógica"[21].

Fig. 3-3 Reconstrucción de la señal analógica**.**
Fuente: [Stu]

La teoría del muestreo define que para una señal de ancho de banda limitado, la frecuencia de muestreo, f_m, debe ser mayor que dos veces su ancho de banda B, medida en Hertz [Hz].

$$f_m > 2B$$

Si la señal a ser digitalizada es la voz y conociendo que el ancho de banda de la voz es de 4,000 Hz aproximadamente. Entonces, su razón de muestreo será 2*B= 2*(4,000 Hz), es igual a 8000 Hz, equivalente a 8,000 muestras por segundo (1/8000). Entonces la razón de muestreo de la voz debe ser de al menos 8000 Hz, para que pueda regenerarse sin error.

3.2.2.2 Cuantización

Es el proceso de transformar valores continuos en series de valores discretos.

El siguiente diagrama ilustra cuando una señal analógica es convertida en una representación digital, en este caso con 8 bits y 16 bits de precisión.

[21] [Stu]

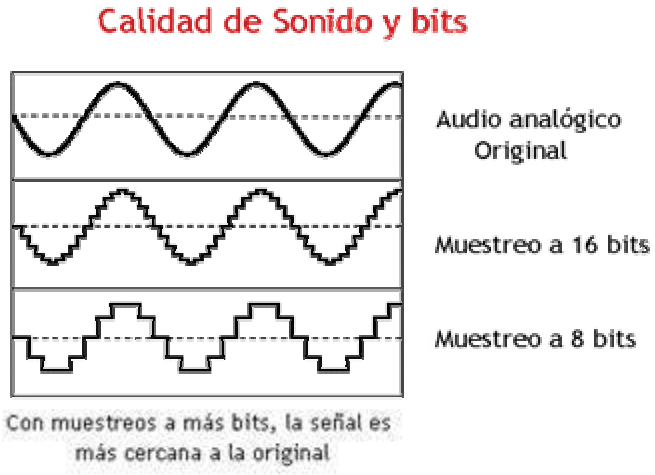


Fig. 3-4 Representación digital de la señal analógica con 16 y 8bits.
Fuente: [Mariñ99]

Mientras que el muestreo representa el tiempo de captura de una señal, la cuantización es el componente amplitud del muestreo.

"Se debe de tomar muestras a tiempos menores y se debe cuantizar a mayores niveles (bits), si sucede lo contrario ocurren **errores de cuantización"**[22].

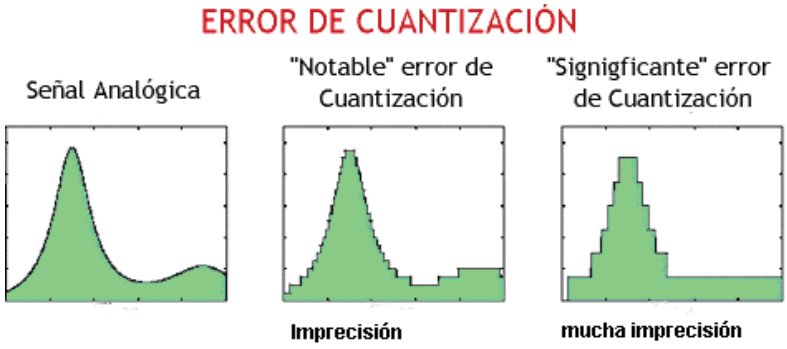


Fig. 3-5 Error de cuantización
Fuente: [Mariñ99]

[22] [Dig]

3.2.2.3 Codificación

La codificación es la representación numérica de la cuantización utilizando códigos ya establecidos y estándares, el código más utilizado es el código binario.

3.2.3 Aliasing

Es la pérdida de datos después de un proceso o transferencia de información, como la señal se muestrea únicamente en intervalos no se conoce lo que está ocurriendo entre muestras, un ejemplo de esto es considerar una "falla " que se produzca entre muestras adyacentes, en este caso no es posible conocer en que punto se produjo el error.

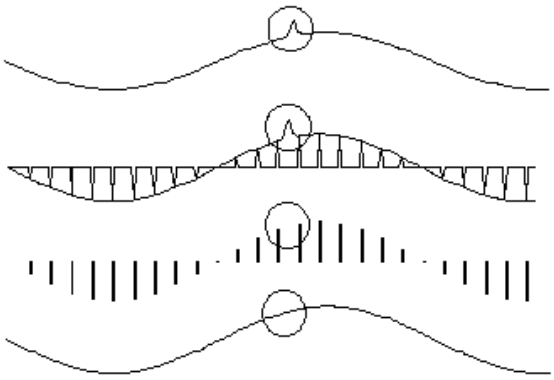


Fig. 3-6: Pérdida de datos por falla en el proceso de muestreo
Fuente: [Bas1]

3.3 MODELOS DE ANÁLISIS ESPECTRAL

3.3.1 Transformada de Fourier

- Las series de Fourier constituyen una importante herramienta en la comunicación porque permiten calcular el ancho de banda de las señales periódicas.

- Simplifican el análisis de complicadas formas de onda mediante su descomposición en señales más elementales (sinusoides).
- Permiten obtener una buena aproximación de la señal manejando únicamente unos pocos términos

El oído humano por medio del caracol, descompone las señales auditivas que le llegan en sus frecuencias fundamentales y ésta es la información básica a partir de la cual se elaboran las señales que le llegan al cerebro. Por tanto podemos afirmar que el proceso de audición se fundamenta en la descomposición en frecuencias de la señal sonora.

Se parte de la base que toda señal genérica, por compleja que sea se puede descomponer en una suma de funciones periódicas simples de distinta frecuencia. En definitiva, la Transformada de Fourier visualiza los coeficientes de las funciones sinusoidales que forman la señal original.

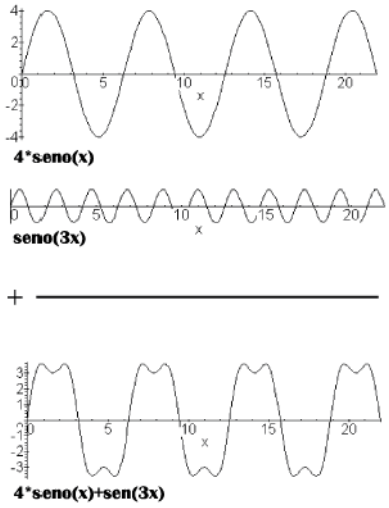


Fig. 3-7: Ejemplo de descomposición de una señal compleja en sumatorio de señales simples.
Fuente: [Enr]

Desde otro punto de vista, la señal inferior puede ser creada sumando las dos funciones sinusoidales superiores.

3.3.1.1 Conceptos básicos

Para explicar el funcionamiento de la transformada de Fourier, partiremos de su formulación básica y explicaremos los conceptos fundamentales.
La Función definida por:

$$F\left(\frac{n}{NT}\right)=\frac{1}{N}\sum_{k=0}^{N-1}m(kT)e^{-j\frac{2\pi nk}{N}} \qquad n=0,1,.....N-1$$

donde,

N: tamaño de ventana

T: período de muestreo

$m(kT)$: muestra del instante kT

n: variable, establece la frecuencia de estudio $0.....N-1$

k: variable

se llama *Transformada de Fourier* de la señal de voz.

El valor del parámetro "n" determina la frecuencia concreta que se va a analizar, es decir, representa una de las frecuencias en las que se va a tratar de descomponer la señal de partida, de esta manera, para hacer el estudio con todas las frecuencias, se usará el rango completo de variación de "n"$=0,1,2,.....N-1$.

Desarrollando la fórmula de la transformada de Fourier para los distintos valores de "n" tenemos:

$$n=0 \;\Rightarrow\; F(0)=\frac{1}{N}\sum_{k=0}^{N-1}m(kT)e^{0}$$

$$n=1 \;\Rightarrow\; F(\frac{1}{N}f)=\frac{1}{N}\sum_{k=0}^{N-1}m(kT)e^{-j2\pi\frac{k}{N}}$$

$$n = 2 \;\Rightarrow\; F(\frac{2}{N} f) = \frac{1}{N}\sum_{k=0}^{N-1} m(kT)\, e^{-j4\pi\frac{k}{N}}$$

$$\vdots$$

$$n = N-1 \;\Rightarrow\; F(\frac{N-1}{N} f) = \frac{1}{N}\sum_{k=0}^{N-1} m(kT)\, e^{-j2\pi\frac{k}{N}}$$

Con las ecuaciones anteriores se obtienen las siguientes conclusiones:

- El sumatorio $\sum_{k=0}^{N-1} m(kT)e^{-j\frac{2\pi nk}{N}}$ indica que la porción de la señal que se analiza se encuentra en el bloque de muestras :

$$[m(0), m(1), m(2), ..., m(N-1)]$$

- El parámetro "n" actúa de índice para obtener las distintas frecuencias de estudio, por ello aparece la secuencia:

$$F\left(\frac{0}{N} f\right), F\left(\frac{1}{N} f\right), F\left(\frac{2}{N} f\right), \;\; ....., \;\; F\left(\frac{N-1}{N} f\right),$$

- Si aumentamos el valor de "N", conseguimos hacer el análisis con un mayor número de frecuencias $(0 \le n \le N-1),$ pero a costa de un mayor tiempo para calcular las operaciones del sumatorio $\sum_{k=0}^{N-1} m(kT)e^{-j\frac{2\pi nk}{N}}$ en el intervalo $(0 \le k \le N-1),$

Ejemplo:

Supongamos una señal de voz sometida a un proceso de muestreo a 10.000 muestras/seg. y un bloque de 100 datos (N=100). Esto implica que se va a realizar el análisis de 10ms de tiempo.

$$n=0 \Rightarrow F(\frac{0}{100}10000)=F(0Hz)=\frac{1}{100}\sum_{k=0}^{99}m(kT)e^{0}$$

$$n=1 \Rightarrow F(\frac{1}{100}10000)=F(100Hz)=\frac{1}{100}\sum_{k=0}^{99}m(kT)*e^{-j2\pi\frac{k}{100}}$$

$$n=2 \Rightarrow F(\frac{2}{100}10000)=F(200Hz)=\frac{1}{100}\sum_{k=0}^{99}m(kT)*e^{-j4\pi\frac{k}{100}}$$

$$\vdots$$

$$n=49 \Rightarrow F(\frac{49}{100}10000)=f(4900Hz)=\frac{1}{100}\sum_{k=0}^{99}m(kT)*e^{-j98\pi\frac{k}{100}}$$

Como se puede observar, el análisis se realiza sobre un ancho de banda aproximado de 5KHz con una resolución espectral[23] de 100 Hz que se obtiene de la siguiente manera:

$$\delta(f)=\left(\frac{1}{NT}\right)=\left(\frac{1}{N}f\right)$$

$$\delta(10.000)=\left(\frac{1}{100}10.000\right)=100Hz$$

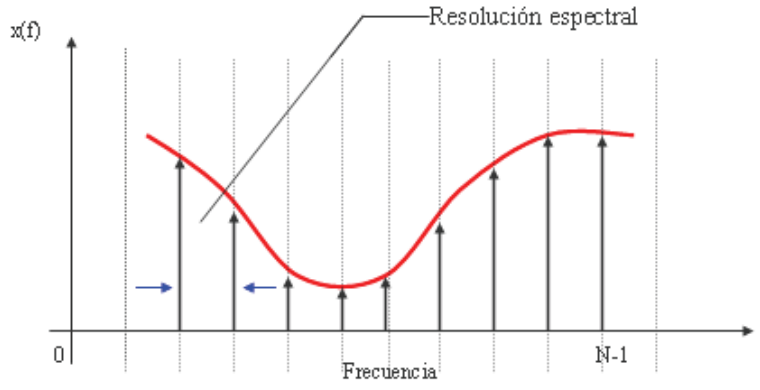


Fig. 3-8: Representación de resolución espectral

[23] Separación mínima detectable entre dos frecuencias contiguas.

3.3.2 Bancos de filtros

Los bancos de filtros pueden entenderse como un modelo sencillo de las etapas iniciales del sistema auditivo humano, su funcionamiento consiste en pasar la señal de voz a través de un conjunto de —filtros paso banda[24]— para calcular la energía de la señal en cada banda.

El diagrama de una estructura de un banco de filtros observamos en la figura 3-10.

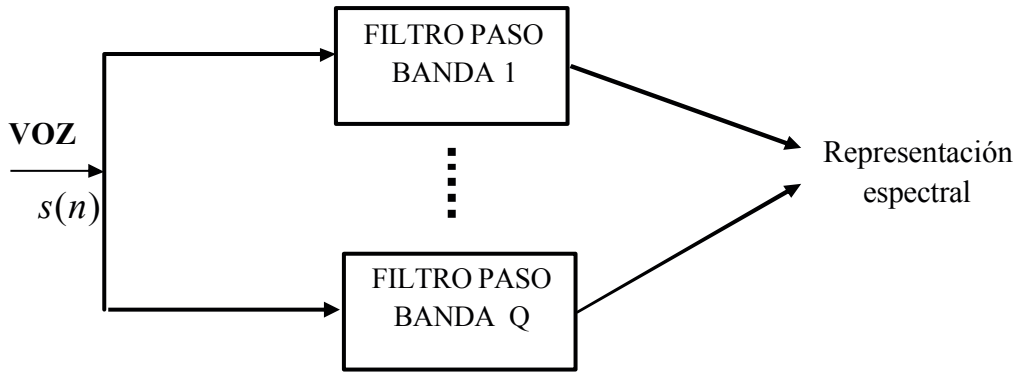


Fig.3-10: Modelo de banco de filtros

Basándonos en el modelo, la señal de voz, $s(n)$ se pasa a través de Q filtros paso banda para medir el rango de frecuencias de interés en la señal (Por ejemplo de 100-8000Hz para señales de banda ancha, 100-3000Hz para señales con calidad de teléfono).

3.3.2.2 Generalizaciones del analizador de bancos de filtros

Una vez que hemos revisado el funcionamiento de los bancos de filtros que se emplean en el reconocimiento de voz, vamos a estudiar la estructura general que se utiliza independientemente del tipo de banco implementado.

[24] Ver anexo V

La generalización incluye un preprocesador de la señal de voz, $s(n)$, que produce una nueva señal, $\hat{s}(n)$, la cual es más "cómoda" para el análisis realizado por el banco de filtro y un postprocesador que actúa sobre los vectores de salida del banco $x(m)$, generando nuevos vectores $\hat{x}(m)$. —El propósito del preprocesador es hacer que la señal de voz sea lo más limpia posible, eliminando el ruido antes de ingresar a los bancos de filtros de tal manera que sea inmune a las fallas de medición de la energía de voz[25]—. Similarmente, el postprocesador limpia la secuencia del vector de características del analizador de bancos de filtros para obtener una mejor representación espectral de la señal de voz y maximizar las oportunidades de reconocimiento de voz.

Fig.3-13: Generalización del Modelo de Banco de filtro

3.3.3 Modelo de codificación por predicción lineal (LPC)

El modelo LPC ha probado ser una técnica muy eficiente debido a la posibilidad de parametrizar la señal de voz con un número pequeño de patrones con los cuales es posible reconstruirla adecuadamente. —Las ventajas que podemos obtener de este modelo son las siguientes:

[25] [RabJua93].

- LPC proporciona un modelo adecuado de la señal de voz y sus parámetros se ajustan a las características del tracto del vocal, especialmente en los sonidos sonoros del habla cuyas propiedades se aproximan más a la señal estacionaria que en los sonidos sordos.
- Los parámetros obtenidos mediante predicción lineal muestran un espectro suavizado que proporciona la información más representativa de la voz[26].

MODELO DEL TRACTO VOCAL

El tracto vocal modelado se manifiesta como un filtro variable en el tiempo cuyos parámetros varían por la pronunciación de las palabras.

El filtro variable en el tiempo tiene dos posibles señales de entrada, y dependerán del tipo de señal: sonora o no -sonora. Para señales sonoras la excitación será un tren de impulsos de frecuencia controlada, mientras que para las señales no sonoras la excitación será un ruido aleatorio. La combinación de estas señales modela el funcionamiento de la glotis. — El tracto vocal manifiesta un número muy grande de resonancias, sin embargo se consideran solo aquellas que cubren un rango de frecuencias entre 100 y 3500Hz y que se las denomina formantes. Esto es debido a que las resonancias de alta frecuencia son atenuadas por la característica frecuencial del tracto vocal que tiende a actuar como un filtro paso bajo.[27]

[26][RabJua93]
[27][New2]

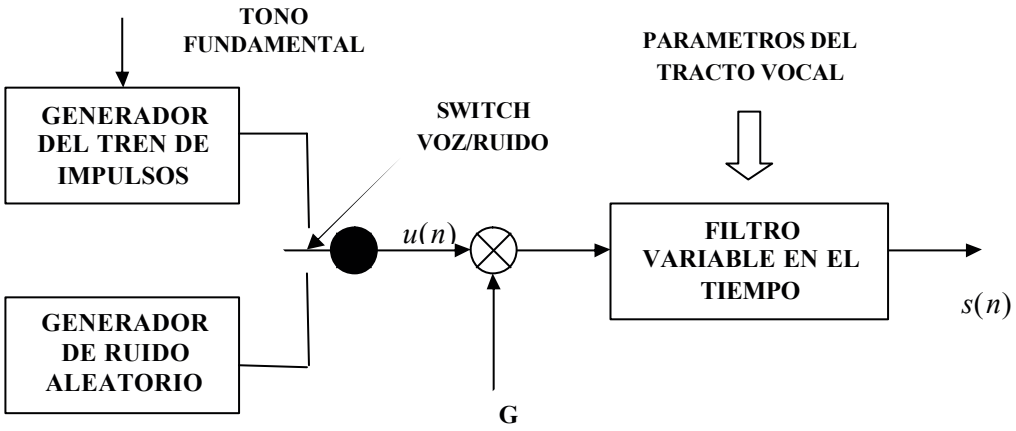


Fig.3-14: Modelo de síntesis de voz basados en el análisis LPC

En las figura 3-15 se representa visualmente la onda de voz y el espectro obtenido tras aplicar la función de transferencia a los coeficientes LPC. Las muestras pertenecen a la grabación de la palabra "generación", la frecuencia de muestreo es de 10Khz.

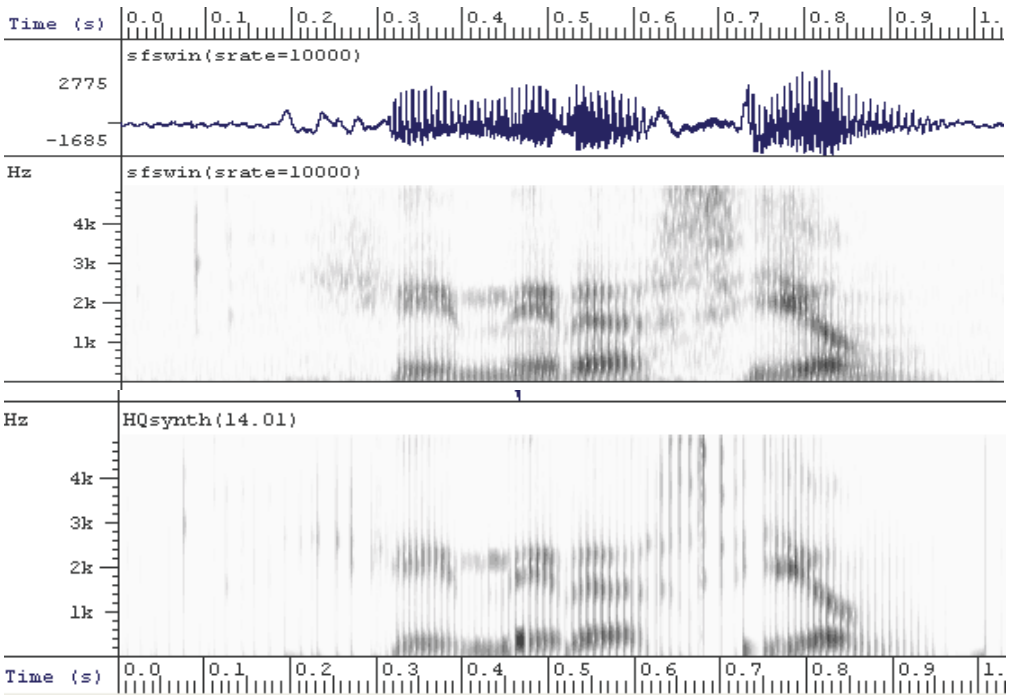


Fig. 3-15: Señal en el tiempo, espectrograma y espectro de voz suavizado obtenido tras aplicar la función de transferencia a los coeficientes LPC.
Gráficos generados por la herramienta SFSWin®.

CAPITULO IV

TÉCNICAS APLICADAS EN EL RECONOCIMIENTO AUTOMÁTICO DE VOZ

4.1 INTRODUCCIÓN

El objetivo de este capítulo es exponer una visión global de varias técnicas utilizadas para el reconocimiento automático de voz por computador, las técnicas que se analizarán son basadas en la primera sección de [RabJua93], y son las siguientes:

- Fonética Acústica
- Reconocimiento de Patrones
- Modelos Ocultos de Markov (HMM)
- Redes Neuronales

4.2 MÉTODO DE FONÉTICA ACÚSTICA

Esta técnica se basa en la teoría Fonético-Acústica la cual afirma que las unidades fonéticas son ampliamente caracterizadas por un conjunto de propiedades que se manifiestan en la señal de voz.

El primer paso en este método es llamado “Fase de segmentación”, porque en éste se divide la señal en regiones donde las propiedades acústicas son representadas a través de unidades o clases fonéticas.

El segundo paso es la determinación de una palabra (s) válida la misma que se obtienen de la secuencia de unidades producidas en el paso anterior.

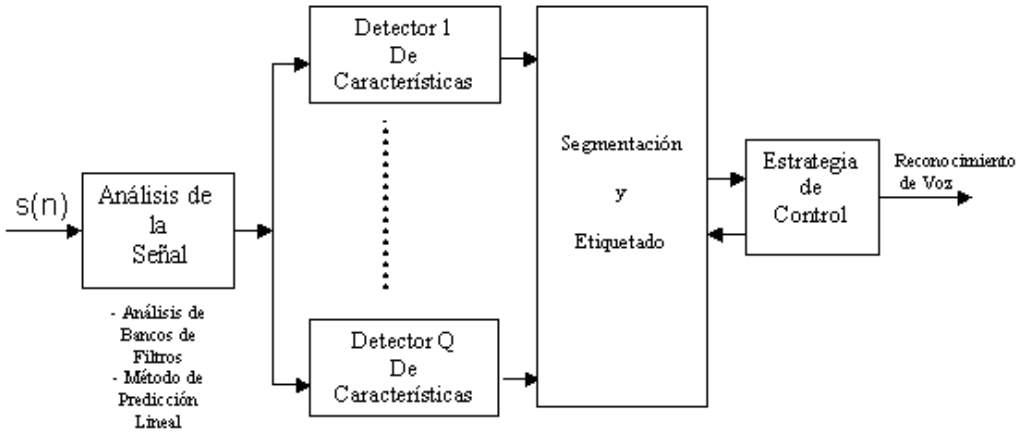


Fig. 4-1 : Diagrama de un sistema fonético – acústico para el reconocimiento de voz

La figura anterior ilustra un diagrama de la técnica de fonética acústica para el reconocimiento de voz. El primer paso en este proceso es el análisis de la señal, en el cual se provee una apropiada representación espectral de las características de la señal de voz en la variación del tiempo, utilizando técnicas de Análisis de bancos de filtros y métodos de predicción lineal.

El siguiente paso en este procedimiento es la etapa de detección de características en la cual se convierte el tamaño de la señal de voz en un conjunto de características que describen el ancho de las propiedades acústicas de diferentes unidades fonéticas; la fase de detección de características usualmente consiste en una conjunto de detectores que operan en paralelo y usan procedimientos lógicos para tomar decisiones en cuanto a presencia, ausencia o valor de una característica.

El tercer paso es la segmentación y etiquetado, ciclo en el que se trata de encontrar regiones estables (donde las características que están sobre la región varían de forma mínima), este período representa el corazón de la técnica fonético- acústica para el reconocimiento de voz, razón por la cual se utilizan varias estrategias de control para limitar el rango de segmentación de puntos.

El resultado del paso de segmentación y etiquetado es la determinación de la mejor palabra o conjunto de palabras obtenidas a través de un proceso léxico.

4.3 MÉTODO DE RECONOCIMIENTO DE PATRONES

Esta técnica ha sido ampliamente utilizada en los reconocedores de voz tradicionales. Su principal ventaja reside en que no es necesario descubrir características espectrales de la voz a nivel fonético, lo que evita desarrollar etapas complejas de detección de formantes, de rasgos distintivos de los sonidos, tono de voz.

El mayor inconveniente de este método es la dificultad para crear una base de datos de patrones del habla que resulte completa, correcta y significativa. Es difícil la obtención de las bases de datos debido a sus grandes tamaños ya que la información espectral que contienen no tiene un significado fonético directo comprensible por parte de los diseñadores.

Esta técnica tiene 4 pasos llamados:

- Análisis de la señal
- Entrenamiento de patrones de voz
- Clasificación de patrones
- Decisión lógica

El proceso de entrenamiento caracteriza las propiedades acústicas de un patrón, este tipo de caracterización de la voz a través del entrenamiento es denominado clasificación de patrones porque la máquina aprende a determinar las unidades de voz que poseen propiedades acústicas similares.

La utilidad de ese método es la etapa de comparación, en la cual cada uno de los patrones aprendidos en la fase de entrenamiento y clasificación son examinados con el fin de armonizarlos.

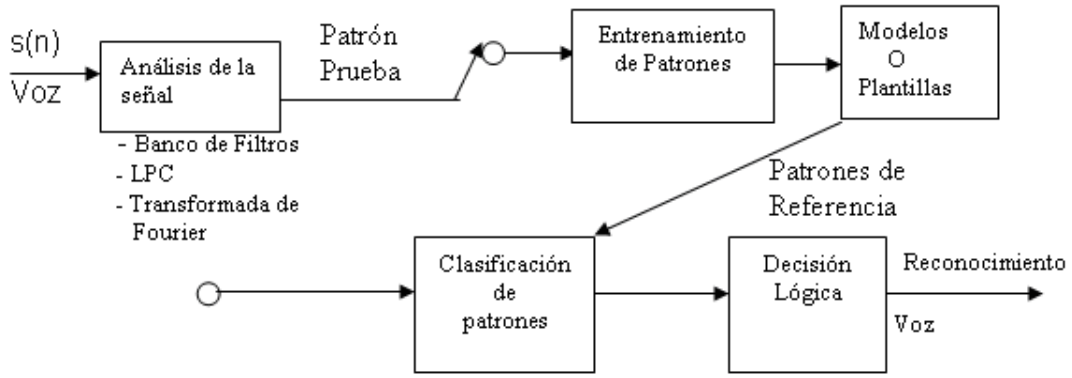


Fig. 4-2: Diagrama de reconocimiento de patrones

En el primer paso que es el análisis de la señal, se toma una secuencia de medidas para definir el "patrón prueba". Para señales de voz la medida de los rasgos son usualmente la salida de algunas técnicas de análisis espectrales tales como: Análisis a través de bancos de filtros, técnicas de predicción lineal y Transformada Discreta de Fourier.

En la fase de entrenamiento uno o más de los "patrones prueba" correspondientes a sonidos de unidades fonéticas semejantes, son usados para crear un patrón representativo llamado "patrón referencia" el mismo que puede ser un modelo o plantilla.

A continuación se realiza la etapa de clasificación en la cual el patrón de prueba desconocido es comparado con el patrón referencia, calculando la distancia que existe entre éstos, para comparar patrones de voz se requiere una distancia local la misma que se define como la distancia espectral entre

dos vectores bien definidos. En la etapa final se realiza la decisión lógica en la cual los patrones de referencia con características similares son usados para decidir qué patrón de referencia o secuencia de patrones de referencia son semejantes al patrón prueba.

La técnica de reconocimiento de patrones es fácil de comprender por su riqueza en la justificación teórica para cada procedimiento utilizado.

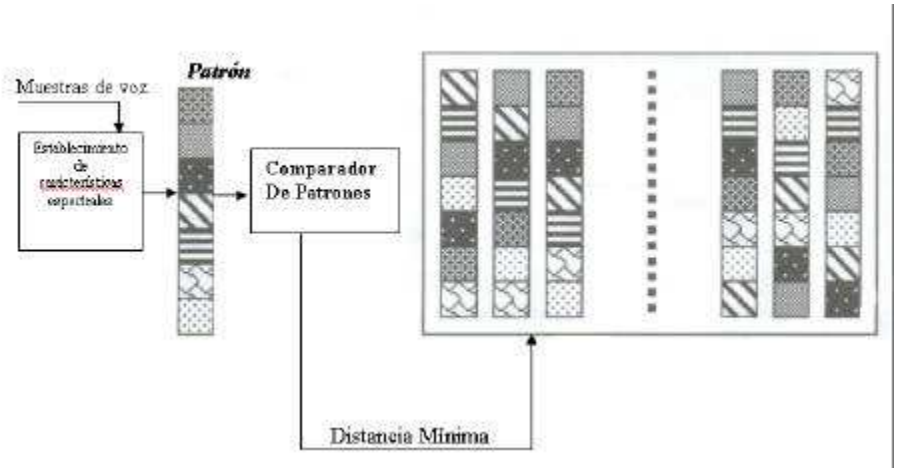


Fig. 4-3: Funcionamiento general de un comparador de patrones

4.4 MODELOS OCULTOS DE MARKOV (HMM).

Un Modelo Oculto de Markov (Hidden Markov Model, HMM) está formado por N estados, en los que se modelan tramos de voz consecutivos. La transición desde un estado a otro está determinada por las probabilidades de conversión entre estados, que son las que representan la variabilidad temporal de la voz. El número de estados para modelos de unidades inferiores a la palabra, se fija normalmente a tres, lo que implica que las unidades más cortas deben durar al menos tres tramas (30 milisegundos).

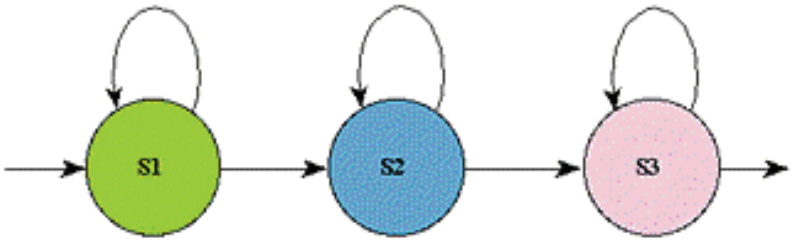


Fig.4-4: Modelo Oculto de Markov de 3 estados

- Un modelo oculto de Markov se puede mirar como una caja negra donde la secuencia de símbolos de salida generados a lo largo del tiempo es visible, pero la secuencia de estados por los que se ha pasado para establecer la serie anterior se desconoce. Es por esta razón que se llaman modelos ocultos.
- Al aplicar los modelos ocultos de Markov en el reconocimiento de la voz, los estados se traducen como modelos acústicos, indicando las ocurrencias de sonidos que son más probables durante los correspondientes segmentos de habla, además las transiciones integran restricciones de tipo temporal acerca de cómo son las secuencias de presentación de esos sonidos.
- Los estados y transiciones pueden ser utilizados para modelar distintas jerarquías del proceso del habla: desde fonemas hasta oraciones pasando por palabras.
- La utilización de este tipo de estructuras debe resolver 3 problemas básicos:
  - Problema de reconocimiento.
  - Problema de decodificación.
  - Problema de aprendizaje o entrenamiento.

4.4.1 Problema de reconocimiento.

Consiste en escoger aquel modelo de entre un grupo de éstos, que mejor represente al conjunto de etiquetas obtenidas a partir de la cuantificación de las correspondientes plantillas espectrales.

4.4.2 Problema de decodificación.

El objetivo es descubrir la secuencia oculta de estados.

4.4.3 Problema de aprendizaje o entrenamiento.

Consiste en construir un modelo de manera que recoja el conocimiento con el que se ha entrenado de forma óptima.

4.5 REDES NEURONALES

4.5.1 Introducción

Una gran parte de los clasificadores usados para el diseño de reconocedores de voz se basan en la utilización de Redes Neuronales.

Las redes neuronales proporcionan un método de aprendizaje automático paramétrico. De esta manera, se puede conseguir hacer separación de clases a partir de un conjunto finito de muestras (patrones) proporcionadas al sistema.

El nombre Redes Neuronales ha sido tomado del modelo biológico que fomenta a este tipo de formalismo matemático de clasificación de patrones.

El tipo de Redes Neuronales que ha sido utilizado con mayor asiduidad en la creación de sistemas de reconocimiento de voz, es el Multicapa con aprendizaje supervisado, basado en el algoritmo de propagación (Backpropagation)

Las Redes Neuronales son redes interconectadas masivamente en paralelo, de elementos simples, los cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico.

4.5.2 Modelo biológico

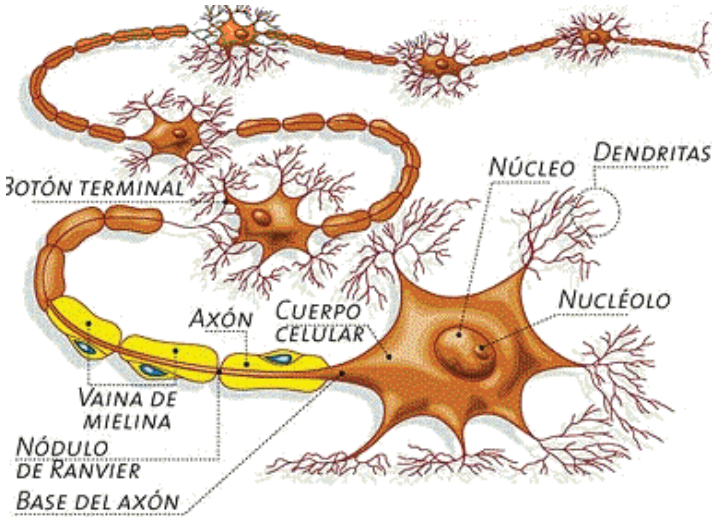


Fig.4-5: Estructura de una neurona
Fuente: [Neu]

Una neurona realiza un proceso, al generar variaciones de potencial eléctrico entre sus entradas y salidas. Por lo tanto es posible considerar como información al potencial eléctrico de la neurona y proceso al mecanismo que permite la modificación de esa información

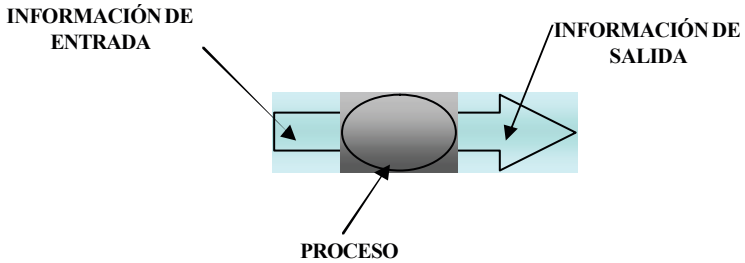


Fig.4-6: Procesamiento de una neurona

Una neurona potencialmente puede estar conectada a miles de neuronas de entrada y miles de conexiones de salida.

4.5.3 Funcionamiento de una neurona

1. El cuerpo de la célula se encuentra en reposo con una carga potencial de -70mV.
2. Al llegar señales de las sinapsis de otras neuronas, el cuerpo de la célula va perdiendo carga negativa
3. Siguiendo el proceso anterior, el cuerpo de la célula llega a situarse con potencial de +40mV
4. Interviene un proceso que descarga esta tensión, devolviendo a la célula a su estado de reposo.

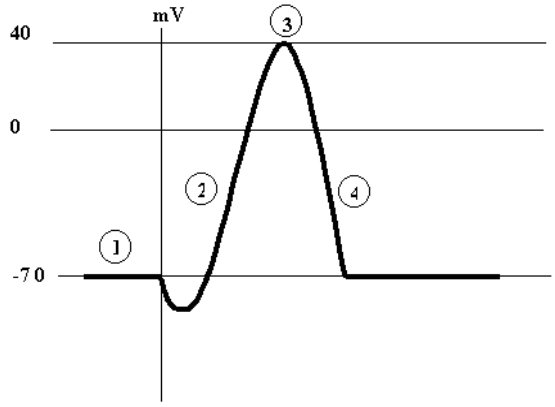


Fig.4-7: Potencial que se propaga a través de axón

La gráfica representada en la Figura 4-7, se corresponde con el potencial que se propaga a través del axón. Este potencial nos marca la evolución de la información a lo largo del tiempo, lo que supone la base del proceso de la neurona.

Los sistemas neuronales biológicos presentan un mecanismo de vital importancia para controlar el flujo de la información que transita a través de las neuronas. Este mecanismo está soportado por el funcionamiento de los neurotransmisores.

Los neurotransmisores actúan químicamente sobre las sinapsis, amplificando o disminuyendo la cantidad de potencial que se transmite de una neurona a otra. De esta manera existen sinapsis inhibidoras y sinapsis excitadoras. Esta característica permite variar la respuesta de las Redes Neuronales, soportando de esta manera la evolución que requieren los mecanismos de aprendizaje.

4.5.4 Características de una neurona artificial simplificada

Las neuronas artificiales más usadas presentan una estructura y pautas de funcionamiento semejantes a sus correspondientes en el modelo natural.

La nomenclatura que se utilizará para describir y analizar neuronas artificiales es la siguiente:

- u_i representar la neurona i-ésima.
- y_i para representar el resultado que genera la neurona y_i (equivalente a la salida que se produce en el axón de la neurona biológica)
- W_{ji} para representar el valor de la inhibición / excitación entre las neuronas u_i y u_j (equivalente al efecto de los neurotransmisores sobre la sinapsis que une u_i con u_j):

$$W_{ji} > 0 \longrightarrow \text{Sinapsis excitadora}$$

$$W_{ji} = 0 \longrightarrow \text{No existe conexión}$$

$$W_{ji} < 0 \longrightarrow \text{Sinapsis inhibidora}$$

- Net_i para representar el valor conjunto de todas las señales que le llegan a la célula u_i. De esta manera podemos establecer que

$$Net_i = \sum_j y_j w_{ji}$$

- $f(Net_i)$ para representar la función de salida o transferencia (equivalente al proceso que realiza la célula biológica, en función de la acumulación de los voltajes que le llegan: Net_i)

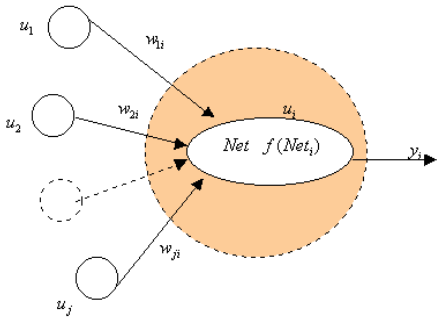


Fig.4-8: Modelo de neurona artificial

Función escalón

Esta función se caracteriza por ofrecer únicamente dos valores de salida. Estos valores se determinan según el factor umbral θ. Cuando $Net_i > \theta$, y_i toma un valor. Cuando $Net_i < \theta$, y_i toma el otro valor establecido.

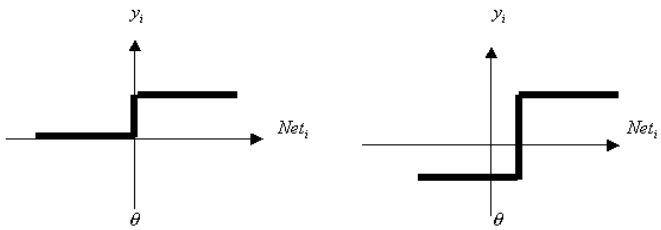


Fig. 4-9: Función escalón

4.5.6 Estructura de una red neuronal artificial

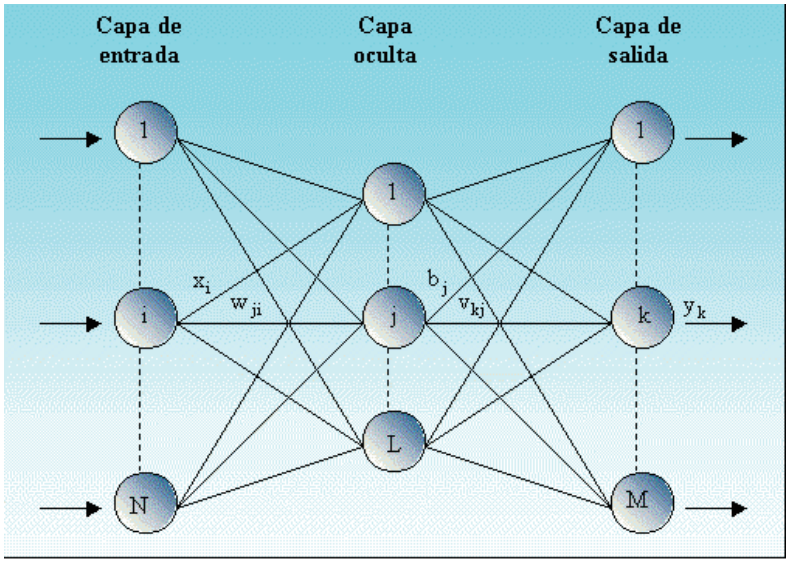


Fig.4-10: Red neuronal artificial
Fuente: [Ima]

- **Entradas:** Valores que alimentan a la Red Neuronal, por ejemplo valores proporcionados por una transformada de Fourier en el reconocimiento de voz
- **Salidas:** Clases reconocidas con la Red Neuronal, por ejemplo un tipo de sonido del español en una aplicación de reconocimiento de voz.
- **Capa de entrada:** Conjunto de neuronas que recogen directamente los valores de entrada a la Red Neuronal
- **Capa de salida:** Conjunto de neuronas que proporcionan los valores de salida de la Red Neuronal.
- **Capas ocultas:** Capas de la Red Neuronal que procesan la información de tal manera que permiten una adecuada separabilidad de las clases que se pretenden reconocer.

4.5.7 Aprendizaje

El proceso de aprendizaje consiste en variar los valores sinápticos w_{ji} siguiendo ejemplos establecidos. En los sistemas biológicos se produce una continua creación y destrucción de conexiones entre las neuronas, en los sistemas que los simulan, la destrucción de una conexión se consigue haciendo que su peso asociado w_{ji} tome el valor cero.

El proceso de aprendizaje ha terminado o la red ha aprendido cuando los pesos permanecen estables; expresando en forma matemática se tiene que la derivada parcial de los pesos con respecto al tiempo tiende a 0.

$\frac{dw_{ji}}{dt} = 0$ Para cada peso de la Red Neuronal

El aprendizaje se puede clasificar como:

- **Supervisado**
  a) Se aplica un patrón de entrada.
  b) Se obtiene la salida que la Red neuronal calcula sobre el patrón de entrada introducido.
  c) Se compara la salida obtenida con la esperada.
  d) Si existe error, significa que hay que variar los pesos de alguna manera para adaptar la Red Neuronal al problema concreto de reconocimiento.
  e) El proceso anterior se repite hasta que se considera aceptable la diferencia entre las salidas que se obtienen y las que se esperan.

- **No Supervisado**

  Emplea las características que las redes recurrentes ofrecen para estabilizar su aprendizaje sin necesidad de ir introduciendo los pares (patrón de entrada, salida esperada) que las redes supervisadas necesitan.

  En el aprendizaje supervisado por corrección de error, se ajustan los pesos en función de la diferencia entre los valores deseados y los obtenidos en la salida de la red; es decir, en función del error producido en la salida.

  $\Delta w_{ji} = \alpha\, y_j\, (d_i - y_i)$ *(Regla delta)*

  d_i *= valor de salida deseado para la neurona* u_j

  α *= factor de aprendizaje (regula la velocidad de aprendizaje)*

  $\Delta w_{ji} + w_{ji}^{\ actual} - w_{ji}^{\ anterior}$ *= modificación del peso* w_{ji}

  $d_i - y_i$ *= Error que se produce en la neurona* u_i

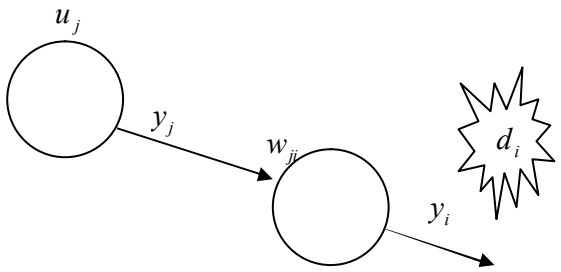


Fig.4-11: Regla Delta

A este tipo de aprendizaje, se le denomina Hebbiano, dada la suposición de Heb (1949): “Cuando un axón de la célula j toma parte en el disparo de la célula i de forma persistente, tiene lugar algún proceso de cambio metabólico en una de las células, o en las dos, de tal modo que la eficiencia

de j, como una de las células que desencadena el disparo de i, se ve incrementada."

4.6 ESTUDIO COMPARATIVO DE LAS TÉCNICAS DE RECONOCIMIENTO DE VOZ

4.6.1 FONÉTICA ACÚSTICA

VENTAJA:

- Las unidades fonéticas son ampliamente caracterizadas por un conjunto de propiedades que se manifiestan en la señal de voz.

DESVENTAJA:

- No existe un procedimiento bien definido para el afinamiento del método en el etiquetado de la voz.

4.6.2 MODELOS OCULTOS DE MARKOV (HMM)

VENTAJAS:

- El método es capaz de obtener un alto porcentaje de reconocimiento con pocos datos para la etapa de entrenamiento.
- Una vez que se ha obtenido el modelo, el tiempo de respuesta al utilizar los reconocedores de voz, es menor que en las otras técnicas de reconocimiento del habla.

DESVENTAJA:

- Cuanto mayor cantidad de material de entrenamiento se use, mayor será la fiabilidad de los modelos entrenados, sin embargo, el hecho de que los conjuntos de entrenamiento sean finitos, implica que algunas unidades aparecerán mucho menos frecuentemente que otras y sus modelos no serán estimados correctamente.

4.6.3 RECONOCIMIENTO DE PATRONES

VENTAJA:

- La técnica de Reconocimiento de patrones provee un amplio rango de unidades de voz (que van desde unidades fonéticas, palabras, frases hasta llegar a oraciones), vocabulario de palabras, acentos.

DESVENTAJAS:

- Los patrones de referencia son sensibles al ambiente y a la voz del hablante, esto se debe a que las características espectrales de la voz son afectadas por el medio de transmisión y el ruido.
- En esta técnica es complicada la incorporación de restricciones sintácticas en la estructura del modelo de reconocimiento de patrones.

4.6.4 REDES NEURONALES

VENTAJAS:

- Son capaces de representar los datos de una manera compacta, por medio de la memorización de las relaciones que los gobiernan, esto es lo que se denomina comprensión de los datos.
- Poseen robustez y tolerancia a fallas como: ruido en la señal de voz.

- Se adaptan en tiempo real para mejorar su rendimiento, esta adaptación es la base del aprendizaje.
- Es la técnica de mayor uso en la actualidad para reconocedores de voz.

DESVENTAJAS:

- Desconocimiento a priori de la estructura de capas y número de nodos necesarios para cada problema.
- Posibilidad de que el sucesivo ajuste de los pesos se esté produciendo alrededor de un mínimo local, lo que significa quedar "anclados" en mínimos locales de las funciones de coste usadas durante el entrenamiento de la red.

CAPÍTULO V
DOMÓTICA: LA TECNOLOGÍA DEL HOGAR

5.1 INTRODUCCIÓN:

En el presente capítulo revisaremos la definición de domótica basándonos en [Dom1],[Sn1] y [Vdom] . Además, analizaremos los estándares de control que utiliza la domótica a nivel internacional.

DEFINICIÓN DE DOMÓTICA

La domótica, palabra que proviene del latín "domo", que significa “casa”, es la tecnología que hace a los edificios inteligentes, en el sentido de que se regulariza, se supervisa y se controla el conjunto de las instalaciones eléctricas y de seguridad del edificio. Esta actividad se realiza en forma integrada y automatizada, mediante el empleo de una aplicación informática, con la finalidad de lograr una mayor eficacia operativa y, al mismo tiempo, un mayor confort y seguridad para las personas que habitan el edificio.

La Domótica está compuesta por un sistema de control programable que integra las funciones de automatización de una vivienda, permitiendo la vinculación de todos los sensores pertenecientes a diferentes subsistemas de la edificación, para que interactúen de acuerdo al programa de control.

Como en una estructura viviente, el sistema domótico está constituido por una red de información que recoge datos a través de sensores instalados en la vivienda, los cuales operan como si fueran el "sistema nervioso" de la edificación. Los

estímulos generados por este "sistema nervioso" son recibidos y analizados por un "cerebro" programable, quien de acuerdo al entorno en que se encuentre la vivienda, genera determinadas respuestas u órdenes, de cuya ejecución se encargarán los actuadores constituidos por las cargas eléctricas que manejan la potencia que se aplica a los diferentes dispositivos con que cuenta la vivienda. El sistema domótico integra todos los servicios del hogar en un solo sistema, permitiendo el acceso desde:

- PC.
- Teclado alfanumérico.
- Touch-screen
- Un teléfono celular
- Internet

En la PC se encuentran cargados el software de programación y monitoreo del sistema. En el monitor se visualizan los planos de la edificación con la distribución de las unidades. Estas unidades pueden ser controladas con el mouse desde la pantalla del monitor. Todo acontecimiento o evento de incumbencia del sistema que suceda en la edificación, queda en un registro con fecha y hora, teniendo la opción de almacenarlo en una base de datos para realizar posteriores consultas y obtener estadísticas. De esta manera el usuario tendrá un total conocimiento de lo que sucede y de lo que ha sucedido en el edificio, en cualquier momento.

Algunas de las acciones que un edificio inteligente puede realizar son:

- Controlar la temperatura de los diversos recintos independientemente o en conjunto.

- Controlar la iluminación, tanto externa como interna y regularla según la presencia del individuo, bien mediante la regulación de las persianas, lámparas, tubos fluorescentes, etc.

- Regular el sistema de riego de plantas y jardines captando la humedad del terreno.

- Detectar inundaciones cortando el suministro de agua automáticamente así como detectar humos y/o gases activando la alarma y avisando al centro de control.

- Permitir al administrador del sistema de control del edificio la conexión/desconexión, de forma local o remota, de todos los componentes anteriormente descritos.

5.3 ESTÁNDARES DE CONTROL

Las redes de control domótico, disponen de una serie de protocolos considerados como estándares dentro de la escena mundial, entre los principales tenemos:

5.3.1 Sistema EIB

El Bus de Instalación (EIB) es un sistema que efectúa comunicación directa, gobierna todas las funciones a través de una única línea de Bus existente, es decir sin precisar de una central. (EIB) es apropiado para:

- Oficinas
- Hoteles
- Escuelas
- Grandes superficies

- Viviendas

5.3.1.1 Ventajas del Sistema EIB

ADAPTABILIDAD

Si se produce una modificación en la utilización del edificio o una ampliación, no se precisa modificar el cableado: Todo está conectado a la única línea del Bus; los componentes del Bus, sensores y actuadores, se programan nuevamente.

REDUCCIÓN DE COSTES DE MANTENIMIENTO

Todos los sistemas están sintonizados entre si y la comunicación funciona óptimamente. De esta forma, el Bus se ocupa de que servicios como: la iluminación, calefacción y climatización estén siempre ajustados a las condiciones ambientales.

AHORRO DE TIEMPO

El esfuerzo en el proyecto y en la instalación se minimiza, porque se reduce considerablemente la cantidad de conductores. Un programa informático apoya este proceso para realizar el proyecto y la instalación (ETS, EIB Tool Software). Y gracias al simplificado cableado se reducen los tiempos de montaje

PREPARADO PARA EL FUTURO

Todos los componentes se pueden conectar sin problemas al Bus disponible, una gran ventaja cuando la instalación debe ser ampliada. Y puesto que el Bus es compatible con sistemas superiores, puede ser acoplado también a otros sistemas de gestión de edificios.

TRABAJO CENTRALIZADO

El sistema trabaja de forma descentralizada. Puede tener estructura lineal, estrellada o ramificada. No es necesario un puesto de control central. Los avisos importantes son considerados prioritarios, lo que garantiza un rápido procesamiento, las prioridades, las direcciones o funciones, pueden introducirse mediante aparato manual de programación o mediante PC.

5.3.2 Sistema BATI BUS

El bus BatiBus es un protocolo abierto que fue desarrollado por las compañías Merlin Gerin, Airelec, Edf y Landis &Gyr, las cuales fundaron en 1989 el Batibus Club International (BCI). El único medio físico del BatiBus es el cable; en el bus se interconectan todos los sensores y actuadores (calefacción, alumbrado, seguridad, etc.) hasta un máximo de7.680 dispositivos. Existe la posibilidad de alimentar los elementos a través del bus. BatiBus está diseñado para edificios de tamaño medio, como pueden ser hogares, residencias, oficinas pequeñas, hoteles o colegios.

5.3.3 Sistema CEBUS

El protocolo de comunicación CEBus (Consumer Electronics Bus) es un estándar vigente en los Estados Unidos que ha sido desarrollado por la Asociación de Industrias Electrónicas (EIA-Electronic Industries Association). El estándar surgió en 1984 cuando la EIA se propuso unificar los protocolos de señalización infrarroja para el control de remoto de electrodomésticos. En 1992 el estándar se había extendido a todo el ámbito de control domótico.

5.3.3.1 Objetivos del estándar CEBUS:

- Facilitar el desarrollo de módulos de interfaz de bajo coste que puedan ser integrados fácilmente en electrodomésticos.
- Soportar la distribución de servicios de audio y vídeo tanto en formato analógico como digital.
- Evitar la necesidad de un controlador central, distribuyendo la inteligencia de la red entre todos los dispositivos. Permitir aumento o disminución de los componentes de la red sin que afecte al rendimiento del sistema ni requiera un gran esfuerzo la configuración por parte del usuario.

5.3.3.2 Medios físicos permitidos:

- Red eléctrica
- Cable trenzado
- Cable coaxial
- Infrarrojos
- Fibra óptica

5.3.3.3 Funcionamiento

Los comandos y los informes de estados se transmiten por el canal de control en forma de mensajes. El núcleo de la especificación se centra en definir este canal de control. El formato de los mensajes CEBUS es independiente del medio de físico utilizado, cada mensaje contiene la dirección de destino sin ninguna referencia sobre el medio físico en el que está situado el receptor o transmisor. De esta manera CEBUS forma una red uniforme a nivel lógico en forma de bus. Soporta una topología flexible, Cualquier dispositivo se puede conectar a cualquier

medio siempre que tenga la interfaz adecuada. Para comunicar segmentos de red que tienen diferente medio físico, se utilizan routers. Para facilitar la difusión de mensajes todos los dispositivos tienen una dirección a la que responden (broadcast address). Además, los conectores pueden ser concentrados en grupos (group address), de esta forma se puede mandar un único mensaje a varios dispositivos al mismo tiempo.

5.3.4 Sistema EHS

EHS define un protocolo de comunicaciones basado en el modelo de referencia OSI/ISO estructurado en 7 niveles, de manera que queda definido desde el cable (o cualquier otro soporte físico) por el que va a circular la información, hasta las reglas sintácticas y semánticas del "idioma" que van a utilizar los diferentes equipos para entenderse entre ellos: Nivel 7 - Aplicación, Nivel 6 - Presentación, Nivel 5 - Sesión, Nivel 4 - Transporte, Nivel 3 - Red, Nivel 2 - Enlace, Nivel 1 – Físico.

5.3.4.1 Medios de Transmisión EHS

Las necesidades de una vivienda a otra pueden variar sustancialmente, por esta razón es necesario contar con medios físicos variados de tal forma que se pueda elegir el más adecuado El caso más evidente lo tenemos entre una casa en fase de proyecto o una ya construida. En esta última será prioritaria la facilidad de instalación, por lo que tendrá preferencia la utilización de la línea eléctrica como medio físico de transmisión (no precisaremos realizar ninguna instalación adicional). Sin embargo, en una casa en fase de construcción podemos realizar un

tendido de par trenzado por toda la vivienda, lo que no nos supondrá un coste excesivo, y obtendremos en general mejores prestaciones.

EHS define varios medios físicos de transmisión:

- Par trenzado
- cable coaxial
- Línea de red eléctrica
- Infrarrojos (IR).

Cada uno de ellos está definido por la especificación EHS para cubrir tres servicios:

1. Servir de soporte para la transmisión de mensajes entre las diferentes unidades.

2. Suministrar alimentación a las diferentes estaciones conectadas a la red domótica.

3. Opcionalmente, transmitir información en tiempo real (audio y vídeo).

CAPITULO VI
HERRAMIENTAS DE RECONOCIMIENTO DE VOZ

6.1 INTRODUCCION

En el presente capítulo describimos las herramientas de reconocimiento de voz integradas en el SICV:

- IBM VIA VOICE DICTATION RUNTIME V8.0
- IMB VIA VOICE TTS V6.4
- MICROSOFT AGENT V.2.0

Estas herramientas fueron elegidas por su capacidad de escalabilidad y adaptabilidad a sistemas de reconocimiento de voz orientados a tareas específicas, es decir, buscamos una herramienta que nos permitiera la interacción de forma directa con el motor de reconocimiento de voz, con la finalidad de poder crear nuestros propios comandos de voz.

Los programas de reconocimiento de voz que hoy en día están ampliamente difundidos en el mercado internacional son: ViaVoice Standard Edition de IBM y Dragon Naturally Speaking Preferred de Scansoft. Para determinar cual de los dos es el mejor en cuanto a nivel de acierto, inmunidad al ruido y precio, descargamos del Internet el artículo: PCACTUAL_MAYO_274-277[28]. Este realiza una evaluación a los dos productos y concluye que el software de IBM es más adecuado para un ambiente de hogar, porque ofrece un mayor índice de acierto y la diferencia de precios es significativamente menor.

Por último, exponemos la tecnología Microsoft Agent que provee una nueva forma de interacción con el usuario mediante la incorporación de personajes animados en aplicaciones, los cuales pueden moverse por toda la pantalla,

[28] [Pca]

hablar y esperar órdenes de voz, brindándole al usuario una interfaz conversacional.

6.2 INTEGRACIÓN DE LOS MODULOS DE IBM Y MICROSOFT AGENT AL SISTEMA DE ILUMINACION

El sistema de iluminación interactúa con el Agente que es el encargado de gestionar las peticiones del usuario, éste contiene la lista de los comandos de voz que se crean dinámicamente y se añaden al vocabulario de comandos del motor de reconocimiento de voz, que además es el encargado de decodificar las palabras que hayan sido pronunciadas por el usuario para realizar la acción de encender o apagar una bombilla.

La respuesta del sistema al usuario se hace a través del Agente el cual utiliza los servicios de síntesis de voz para notificar de forma hablada al usuario sobre la acción que se está realizando.

El nivel en que está ubicado cada módulo se observa en la siguiente figura:

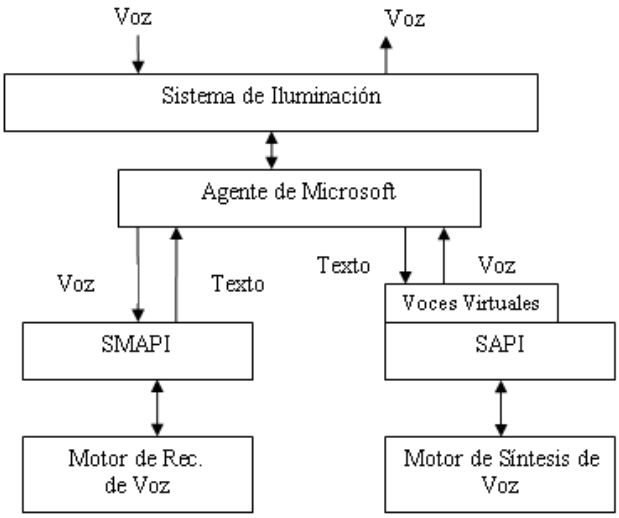


Fig. 6-1: Interacción del Sistema de iluminación con el Agente y los motores de reconocimiento y síntesis de voz de IBM.

6.3 IBM VIA VOICE DICTATION RUNTIME V8.0

La herramienta —IBM ViaVoice Dictation Runtime v8.0[29]— soporta dictado contínuo, reconocimiento de palabras aisladas y ejecución de comandos para la realización de una determinada acción.

Las aplicaciones pueden acceder al motor de reconocimiento de Via Voice por medio de la API (*Speech Manager API*: SMAPI), que comprende una parte del motor, es decir es un recurso más al que puede acceder una aplicación.

En el siguiente gráfico observamos el nivel en el que se ubica el SMAPI:

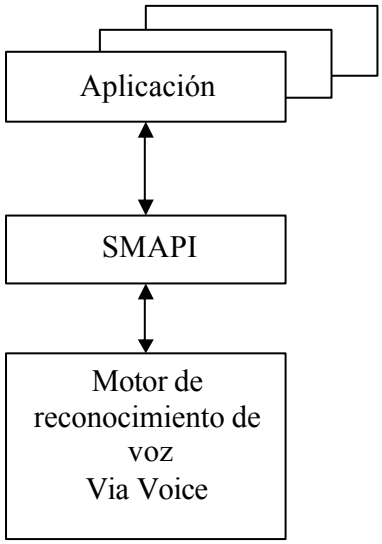


Fig. 6-2: Nivel en el que se ubica el SMAPI

6.3.1 Recursos del motor de reconocimiento de voz

El motor de Via Voice trabaja con un conjunto de recursos que utiliza para procesar las palabras que han sido pronunciadas y convertirlas en texto:

[29] [Pgm]

- Lenguajes de usuario
    - Dominios
        - Vocabularios
        - Modelos de palabra
        - Pronunciaciones
    - Modelo de voz

Lenguajes de usuario:

Los lenguajes soportados por el Via Voice son los siguientes:

- Ingles U.S.
- Francés
- Alemán
- Italiano
- Español

Cabe mencionar que múltiples lenguajes pueden ser instalados y coexistir en una misma máquina y aplicación.

Dominios:

Cada lenguaje incluye diferentes *dominios.* Un dominio es un conjunto de vocabularios, pronunciaciones y modelos de palabra, diseñados para soportar una aplicación específica.

Via Voice se ejecuta con un dominio general del lenguaje seleccionado, el cual contiene de 20.000 a 30.000 palabras representativas de un ambiente de oficina.

Para aplicaciones de comandos de voz, el Via Voice dispone de un conjunto de herramientas para la creación de vocabularios específicos de acuerdo a la aplicación a implementarse.

6.3.2 Vocabularios y modelos de palabras:

Un vocabulario es una lista válida de palabras o frases, provee información estadística de las secuencia de palabras. Tanto los vocabularios como los modelos de palabra son usados por el motor en la selección de la frase o palabra que haya sido pronunciada.

Existen 3 tipos de vocabularios que son manejados de manera diferente por el motor de Via Voice:

- **Vocabularios de Comando:**

  Usados para el reconocimiento de palabras y frases de una lista que puede crearse dinámicamente en tiempo de ejecución.

  La decodificación de las palabras está basada en los sonidos que la componen. Después que una palabra o frase ha sido reconocida, el motor se detiene y espera para que la aplicación responda a esta palabra. El motor se detiene porque es probable que el estado de la aplicación cambie en respuesta a una orden y esto provocará un nuevo juego de vocabularios válidos.

- **Vocabularios de Gramática:**

  Son usados en el reconocimiento contínuo de voz, cuando una aplicación requiere comandos más complejos, como por ejemplo: "enviar correo a nombre@hotmail.com".

  Después que una palabra o frase es reconocida en un vocabulario de gramática, el motor se detiene y espera para que la aplicación responda. De igual forma la aplicación puede cargar un nuevo conjunto de vocabularios almacenados en un archivo mientras el motor está detenido.

- **Vocabularios de Dictado:**

  En vocabularios de dictado el modelo de palabra usado es una base de datos compuesta de una secuencia de palabras que ocurren con mayor frecuencia en un dominio específico. Durante el dictado, el modelo de palabra asiste al motor en la selección de la mejor secuencia de palabras dictadas y diferencia palabras que son acústicamente idénticas.

Pronunciaciones:

Las pronunciaciones son las representaciones fonéticas de una palabra. Las palabras pueden tener múltiples representaciones. Via Voice contiene un vocabulario predefinido de 20.000 a 30.000 pronunciaciones de palabras.

Modelos de voz:

El motor de reconocimiento de Via Voice usa un modelo voz independiente del hablante en la decodificación del habla, lo que significa que los usuarios no tienen que entrenar al sistema para lograr precisión en el reconocimiento de palabras. En Via Voice los usuarios pueden opcionalmente crear un modelo de voz personal grabando una secuencia predefinida de oraciones y el sistema generará un modelo de voz con características propias de cada persona, tales como el acento y tono de voz.

El siguiente gráfico ilustra la relación de los recursos de voz con el proceso de reconocimiento. En este ejemplo usamos el lenguaje español, el motor está utilizando el dominio general del español para dictado contínuo y el modelo de voz activo es Juan.

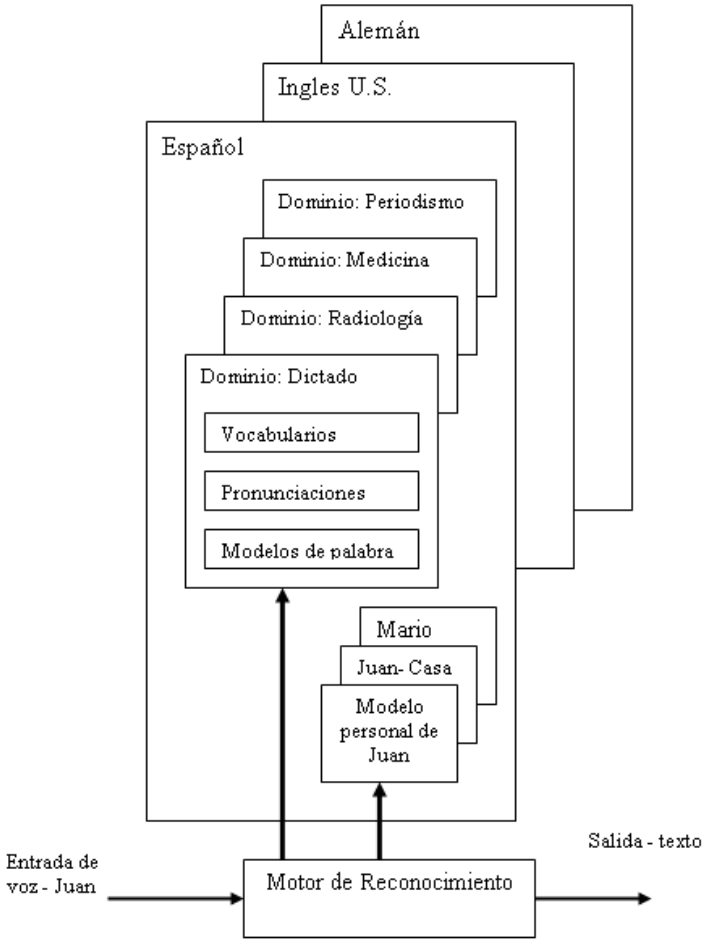


Fig.6-3: Relación de los recursos de voz con el proceso de reconocimiento de voz

6.4 ARQUITECTURA DEL MOTOR DE RECONOCIMIENTO DE VOZ

La siguiente figura ilustra la arquitectura lógica del Via Voice. Observamos los componentes principales del motor y como la aplicación interactúa con él a través del SMAPI.

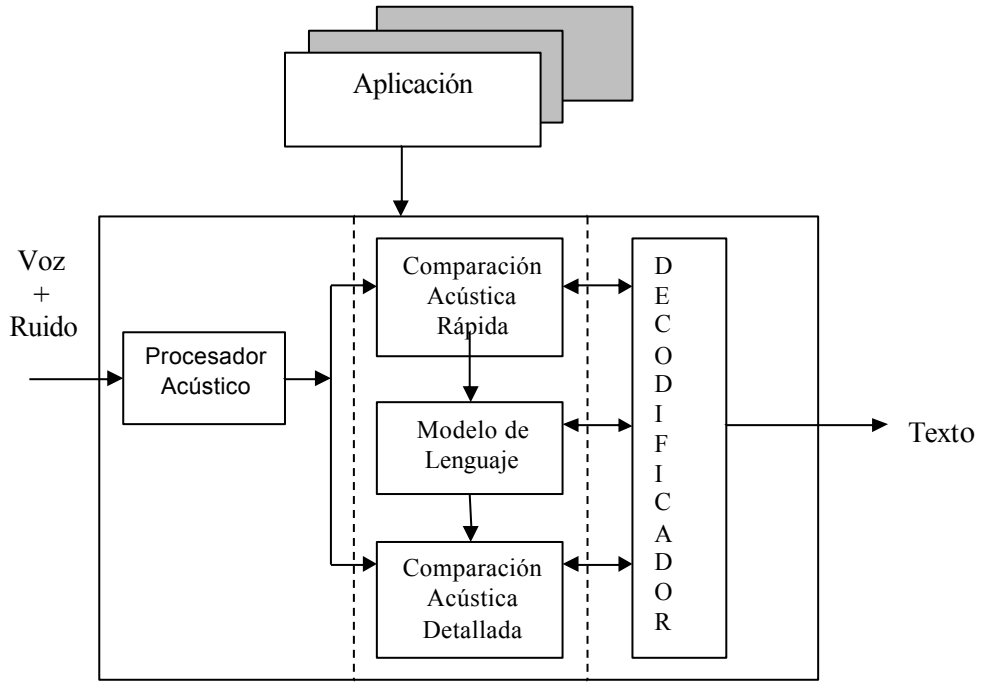


Fig. 6-4: Arquitectura del motor de voz

6.4.1 Procesador acústico

El procesador acústico toma la señal de voz y la convierte en un formato apropiado para su posterior uso. Está compuesto por dos componentes: el procesador de señal y el etiquetador.

En el motor de Via Voice la señal de voz es recogida por el micrófono para ser analizada por el procesador. La señal de audio es capturada a 11kHz, que contiene el habla y el ruido del ambiente. El procesador debe adaptarse al entorno acústico para reducir el impacto del ruido. El análisis realizado por el procesador de la señal produce un conjunto de números, que representan una centésima de segundo de un segmento de voz. Estas características capturan la señal de voz en forma compacta (tal como energías en diferentes bandas de frecuencia). Así el procesador de señal

analiza el audio de entrada y reduce la señal original de 11.000 muestras/segundo a 100 vectores por segundo.

El etiquetador convierte la salida generada por el procesador en un conjunto de cadenas que identifican varias categorías de sonido. Usa plantillas o prototipos para la clasificación de los diferentes sonidos de voz. Estos prototipos corresponden a los fonemas del lenguaje hablado. La salida del etiquetador refleja la comparación realizada entre los vectores de características contra los prototipos almacenados para determinar la categoría del sonido en cada centésima de segundo.

6.4.2 Emparejamiento de palabras

La siguiente fase del proceso de reconocimiento de voz consiste en identificar las palabras candidatas y el rango de posibilidades basadas en el análisis acústico y contextual. Este proceso incluye los siguientes pasos:

6.4.3 Comparación acústica rápida

Realiza una primera aproximación de la palabra pronunciada contra todas las palabras del vocabulario, produciendo una pequeña lista de por ejemplo 100 palabras candidatas.

6.4.4 Modelo del Lenguaje

El Modelo del Lenguaje analiza la probabilidad de la secuencia de palabras independientemente de su forma acústica, el modelo del lenguaje contribuye en la predicción de palabras futuras basadas en palabras que se hayan pronunciado. Un modelo de lenguaje representa un dominio específico, (por ejemplo: medicina, periodismo, etc.)

En vocabularios de comandos dinámicos, el modelo del lenguaje es uniforme, es decir, cada palabra o frase tiene la misma probabilidad de ocurrencia.

6.4.5 Comparación acústica detallada

Realiza una comparación acústica sobre el conjunto de palabras candidatas. Este proceso demanda de mayor procesamiento del computador. Utiliza el modelo acústico y de lenguaje para producir una nueva lista de palabras candidatas, por ejemplo de las 100 palabras anteriores se obtienen una nueva lista de 25 palabras que son las más probables que se han pronunciado.

6.4.6 Búsqueda

El decodificador es el último componente del motor, es el encargado de ejecutar la búsqueda de la secuencia de palabras una vez que se han obtenido los puntajes de ocurrencia en el modelo acústico y del lenguaje. Es decir calcula la probabilidad de las cadenas de palabras usando una pila de búsqueda para encontrar la oración completa más probable que ha sido pronunciada dados los puntajes de las palabras del modelo acústico y de lenguaje.

6.5 IBM VIA VOICE- TTS V6.4

El motor de —IBM ViaVoice- TTS v6.4[30]— realiza la conversión de texto a voz toma el texto como entrada de una aplicación y la transforma en voz. Las aplicaciones acceden a los recursos del motor a través de un conjunto de

[30] [Mant]

API's. Las interfaces de programación del motor están disponibles a través de las voces virtuales y del API de Microsoft.

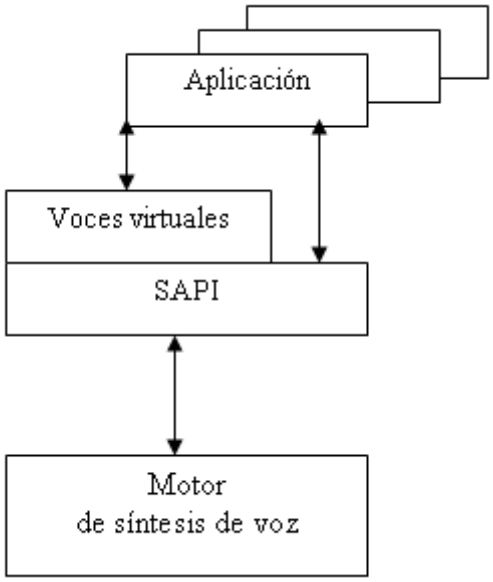


Fig.: 6-5: Motor de síntesis de voz

6.5.1 Recursos del motor de texto a voz

El motor de texto a voz usa los siguientes recursos para transladar el texto a voz sintetizada:

- Diccionarios de usuario
  - Palabras especiales
  - Abreviaciones
  - Raíces

Diccionarios de usuario

La pronunciación de una palabra (que es una representación fonética) es la fuente más importante para el motor de síntesis de voz en el procesamiento del texto. Las pronunciaciones están representadas en símbolos fonéticos.

El motor usa reglas de pronunciación para generar los sonidos que componen la palabra. Sin embargo, de forma alternativa, las pronunciaciones pueden estar especificadas en diccionarios de usuario permitiendo la aplicación de reglas de sonido normales.

Palabras Especiales:

Los diccionarios de palabras especiales se utilizan para representar las pronunciaciones de palabras compuestas, cartas (mayúsculas / minúsculas) y números.

Las pronunciaciones de palabras en los diccionarios especiales pueden contener anotaciones y representaciones fonéticas. Por ejemplo, la pronunciación para “guardabosque”, puede ser ingresada como “guarda bosque”.

Abreviaciones:

El diccionario de abreviaciones se emplea para especificar siglas y abreviaturas que son pronunciadas letra a letra, (por ejemplo: “IBM” se pronuncia como: “I”, “B”, “M”) y abreviaciones que son expandidas (por ejemplo “mn” a “milla náutica”)

Raíces:

El diccionario de raíces se utiliza para especificar pronunciaciones de palabras ordinarias, como: verbos, sustantivos y adjetivos.

6.6 ARQUITECTURA DEL MOTOR DE SÍNTESIS DE VOZ

El motor de síntesis de voz toma la cadena de caracteres y la convierte en onda de voz, está formado por dos componentes principales:

- Componente de procesamiento de texto

- Componente de generación voz

El componente de procesamiento de voz convierte el texto de entrada en una representación de voz. El componente de voz toma esta representación de voz para generar la onda de voz (el habla).

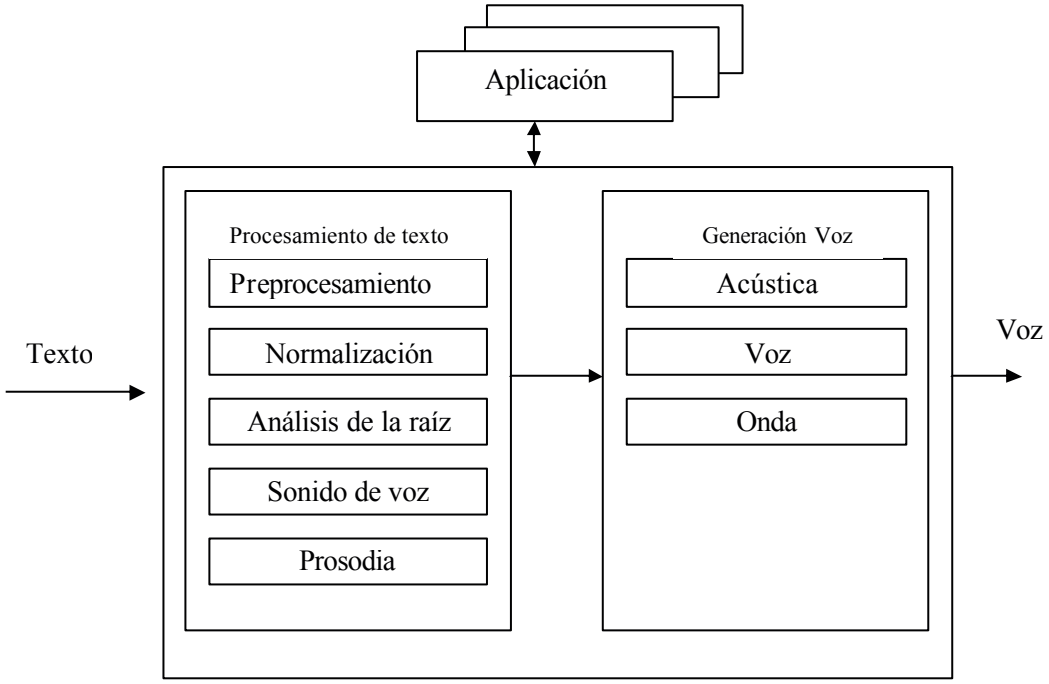


Fig. 6-6: Arquitectura del motor de síntesis de voz

6.6.1 Componentes del procesamiento de texto

Preprocesamiento:

El bloque de preprocesamiento examina el texto de entrada para determinar: anotaciones, abreviaturas, números, horas, fechas y signos de puntuación.

Normalización:

En la fase de normalización del texto, las cadenas de caracteres que representan abreviaciones, iniciales, numerales son convertidas a palabras. Por ejemplo: “=” se convierte en “igual” y “32” es convertido a “treinta y dos”.

Análisis de la raíz:

El siguiente paso, consiste en analizar las palabras en sus raíces, para lo cual se consulta el diccionario de raíces del que se obtienen las pronunciaciones de cada palabra.

Prosodia:

La prosodia genera una representación de la melodía de la voz aplicando un conjunto de reglas, cuando esto se ha realizado el componente de procesamiento del texto ha generado una abstracción lingüística de la voz que el componente de voz puede usar para convertirlo en valores acústicos.

6.6.2 Componentes de generación de voz

El componente de voz consiste en tres procesos básicos: primero se generan los valores acústicos para producir los fonemas y los patrones de prosodia especificados en el componente de procesamiento de texto. Segundo, se añaden las características de la voz y, finalmente el sintetizador crea la onda de voz que es enviada a la tarjeta de sonido.

6.5 MICROSOFT AGENT V.2.0

Microsoft Agent v2.0[31] es un API que provee un conjunto de servicios que soporta animación de personajes en 3D, reconocimiento y síntesis de voz. Ha sido implementado como un servidor de automatización OLE (Modelo Objeto Componente COM). Al ser un servidor puede iniciar sesiones y ejecutar servicios de entrada y salida de voz en varias aplicaciones al mismo tiempo.

Como un servidor COM, Microsoft Agent automáticamente se inicia y espera por peticiones de clientes para conectarse, y si no existen se detiene.

Microsoft Agent también incluye un control ActiveX, este control hace que sea fácil el acceso a los diferentes servicios del Agent desde los lenguajes de programación que soportan esta tecnología. Hemos utilizado este último por las facilidades que brinda al desarrollador al momento de la implementación.

6.6 DOCUMENTACIÓN TECNICA DE MICROSOFT AGENT V.2.0

VER ANEXO VI

[31] [Msag]

CAPITULO VII

METODOLOGIA Y ELABORACION DEL SOFTWARE DE ILUMINACIÓN DE UNA CASA POR COMANDOS DE VOZ

7.1 METODOLOGIA: PROCESO UNIFICADO DE RATIONAL (RUP)

El RUP[32] es un proceso de —desarrollo de software[33]—, que es:

1. **Manejado por casos de uso**: Un caso de uso es una funcionalidad que el sistema provee al usuario. Los casos de uso constituyen una orientación para las actividades que se realizan durante todo el proceso de desarrollo, incluyendo el diseño, implementación y pruebas del sistema.

2. **Centrado en arquitectura:** La arquitectura comprende los elementos más relevantes del sistema; los principales factores que la influyen son: plataformas de software, sistemas operativos, manejadores de bases de datos, protocolos de comunicación.

3. **Iterativo e Incremental:** Es recomendable dividir el proyecto en partes más pequeñas o mini proyectos. Cada mini proyecto es una iteración que resulta en un incremento. Las iteraciones deben estar controladas, es decir, deben seleccionarse y ejecutarse de forma planificada.

4. **Basado en componentes:** La creación de software complejo, implica dividir el sistema en componentes con interfaces bien definidas, que posteriormente serán ensamblados. Esta característica en un proceso de

[32] [Trans]

[33] Conjunto de actividades indispensables para convertir los requisitos de un usuario en un sistema de software

desarrollo permite que el sistema se vaya creando a medida que se desarrollan sus componentes.

5. Modelado por un único lenguaje: UML[34] es adoptado como único lenguaje de modelamiento para el desarrollo de todos los modelos.

7.2 FASES EN EL CICLO DE DESARROLLO

Fase 1: Preparación hicial En esta fase se identifican y priorizan los riesgos más importantes, se planifica en detalle la fase de elaboración y se estima el proyecto de manera aproximada.

Fase 2: Preparación Detallada

En esta fase se realiza la captura de la mayor parte de los requerimientos funcionales, manejando los riesgos que interfieran con los objetivos del sistema, acumulando la información necesaria para el plan de construcción y obteniendo suficiente información para hacer realizable el caso del negocio.

Fase 3: **Construcción**

En esta fase la línea base de la arquitectura crece hasta convertirse en el sistema completo. La descripción evoluciona hasta convertirse en un producto preparado, listo para operar, éste es frecuentemente llamado versión beta.

Fase 4: Transición

[34] Es un lenguaje de modelamiento para la especificación, visualización, construcción y documentación de los artefactos de un proceso de sistema intensivo. Ver anexo VII

Su objetivo principal es realizar la entrega del producto operando, una vez realizadas las pruebas de aceptación por un grupo especial de usuarios y habiendo efectuado los ajustes y correcciones que sean requeridos.

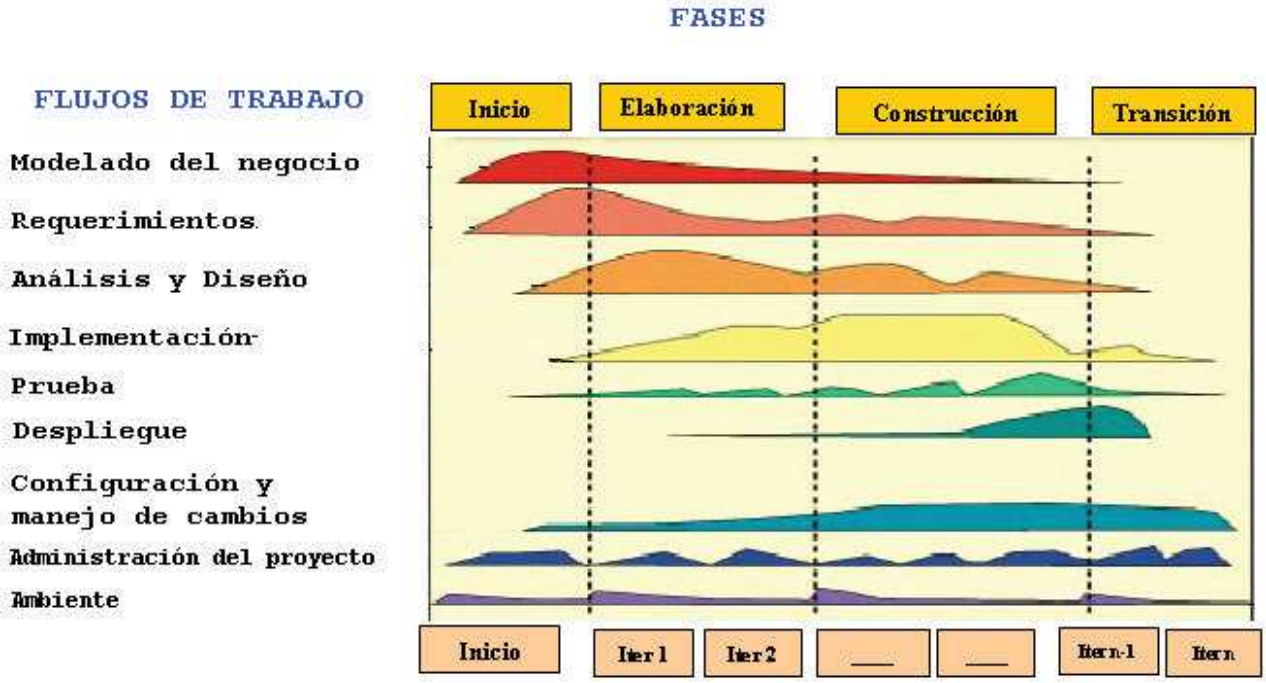


Fig. 7-1: Fases de la Metodología RUP
Fuente: [Ajl]

7.3 DESARROLLO DE LA DOCUMENTACIÓN TÉCNICA Y DE USUARIO

7.3.1 MANUAL TÉCNICO

- Permite realizar el mantenimiento del software de una manera fácil.
- Es un elemento imprescindible que certifica la calidad del sistema desarrollado. (Anexo VIII)

7.3.2 MANUAL DE INSTALACIÓN

- Provee al usuario la guía necesaria para efectuar la instalación del sistema.(Anexo X)

7.3.3 MANUAL DE USUARIO

- Permite que el conjunto de usuarios puedan usar la aplicación de manera efectiva.
- Establece una solución para prescindir de la dependencia de personal. (Anexo XI)

7.3.4 PRUEBAS

- Representa una revisión final de las especificaciones, diseño y de la codificación de la aplicación desarrollada. (Anexo XII)

7.4 CONCLUSIONES

- El presente trabajo nos permitió conocer claramente cuáles son las tecnologías disponibles para la iluminación de una casa.

  Este trabajo ha sido la oportunidad de investigar en el área de reconocimiento de voz utilizando Microsoft Agent y herramientas del IBM Via Voice, para aplicarlo en ambientes domésticos que desean tener un control de la iluminación que genere ahorro energético.

- El ritmo actual de vida ha provocado un fenómeno cultural sin precedentes. Estamos inmersos en la Sociedad de la Comunicación de Información. Nuestros sentidos y nuestro lenguaje nos entregan grandes cantidades de datos procesados; con poco esfuerzo podemos transformar información presentada en dibujos a palabras, y viceversa. Por tal motivo, el presente trabajo incorporó la tecnología de reconocimiento de voz en la automatización y control de la iluminación de una casa.
- El proceso de investigación nos introdujo en temas que tenían una vasta dispersión de los conocimientos involucrados en cada área, de modo que nuestra labor fue encontrar y captar publicaciones especializadas en campos relacionados a tratamiento de señal, fonética acústica y matemáticas, logrando así entender el funcionamiento de los sistemas informáticos de reconocimiento de voz.

- La tecnología de reconocimiento de voz implementada en los hogares es un mercado de crecimiento actual muy rápido en el área académica y comercial, debido a que las personas desean facilitar el acceso a los sistemas informáticos a través de la utilización de interfaces denominadas “hombre / máquina”, que con el hecho de utilizar sencillas instrucciones vía voz pueden interactuar de forma adecuada con el computador.

7.5 BIBLIOGRAFIA:

REFERENCIAS BIBLIOGRÁFICAS

[Brist86]
G. Bristow. “Electronic Speech Recognition, Techniques, technology & Applications”, McGraw-Hill, 1986.

[Flore87]
Ramón G. Flores, Federico Velasco Coba, “Análisis de Fourier”, ADDISON, 1987.

[Mariñ99]
José B. Mariño, “Tratamiento digital de la señal. Una introducción experimental”, Alfaomega, 1999

[Moor90]
R. Moore, “Speech Processing”, McGraw-Hill, 1990.

[RabJua93]
L.R. Rabiner, B.H. Juang, “Fundamentals of Speech Recognition, Prentice Hall”, 1993

REFERENCIAS DE INTERNET

A

[Asu]
http://gps-tsc.upc.es/veu/personal/asuncion/curso/Tema2.ppt

[As1]
http://gps-tsc.upc.es/veu/personal/asuncion/curso/Tema2.ppt

[Acu]
http://tamarisco.datsi.fi.upm.es/ASIGNATURAS/FRAV/apuntes/acustica.pdf

[Ajl]
http://www.ajlopez.com/

[Anex]
http://www.diac.upm.es/asignaturas/sistaudio/TF Ignacio/equipos4/anexo2.html3#uno

B

[Bas]
http://www.bores.com/courses/intro/basics

[Bas1]
http://www.bores.com/courses/intro/basics/1_alias.htm

[Bay]

http://www.itch.edu.mx/academic/industrial/sabaticorita/_private/08Teorema%20de%20bayes.htm

C

[Cast]
www.memo.com.co/fenonino/aprenda/castellano/castellano36.html

[Cour]

http://www.bores.com/courses/intro/

[Coag]

http://www.microsoft.com/spanish/MSDN/estudiantes/ia/conocimiento/agent.asp

[Crea]

http://www.creangel.com/

D

[Dig]
http://www.tecnun.es/asignaturas/tratamiento%20digital/tema5.pdf

[Dom1]
http://cablemodem.fibertel.com.ar/casasinteligentes/domotica.htm

[Down1]

http://www.wizzardsoftware.com/voice/voicetools/download_section.asp

[Down2]

http://www.wizzardsoftware.com/voice/voicetools/download_section.asp

[Dtw]
http://alek.pucp.edu.pe/~dflores/tesis/dtw.html

E

[Enr]
http://www.upseros.net/fotocopiadora/ficheros/Sistema%20Operativo%20UNIX/
Enrique%20Torres/resumen%20ordenes.pdf

[Enc]
http://www.nlm.nih.gov/medlineplus/spanish/ency/images/ency/fullsize/8762.jpg

F

[Fon] http://www.filos.unam.mx/LICENCIATURA/Cuetara_Palacios_2003/El_aparato_fonador.ppt

[Fun]
http://www.isip.msstate.edu/projects/speech/software/tutorials/production/fundamentals/current/section_05/s05_01_p01.html.

[Fram]

http://www.uiowa.edu/~acadtech/phonetics/spanish/frameset.html

[Fones]
http://www.auburn.edu/forlang/Spanish/FLSP0301mats/slides/fonesp-04.ppt

G

[Guid]
http://compy.ww.tu-berlin.de/IBMVoice/proguide/pgm guide.htm#ToC_26

I

[Inv1]
http://www.gtc.cps.unizar.es/~eduardo/investigacion/voz/rah.html

[Ima]
http://bibliopsiquis.com/psicologiacom/vol5num2/2833/Image249.gif

[Indx]

http://www.isip.msstate.edu/projects/speech/index.html

M

[Mss]
http://www. isip.msstate.edu

[Mac]
http://www.digitalizo.com.ar/mac/mp3.htm

[Mant]
http://compy.ww.tu-berlin.de/IBMVoice/proguide/pgmguide.htm#ToC_26 puede consultar el manual técnico.

[Msag]

http://www.microsoft.com/msagent/default.asp

N

[New2]
http://www. net-research.net/swen/contraintelige/newpage21.htm

[Neu]
http://icarito.tercera.cl/infografia/chumano/nervioso03/img/neurona-55k.jpg

O

[Ord]
www.upseros.net/fotocopiadora/ficheros/Sistema%20Operativo%20UNIX/Enrique%20Torres/resumen%20ordenes.pdf

P

[Pca]
http://www.outsource-sl.com/fabricantes/IBM/PCACTUAL_MAYO_274-277.pdf

[Pho]
http://www.uiowa.edu/~acadtech/phonetics/spanish/frameset.html

[Pgm]

http://compy.ww.tuberlin.de/IBMVoice/proguide/pgmguide.htm#ToC_26

[Pgmi]

http://compy.ww.tu-berlin.de/IBMVoice/proguide/pgmguide.htm#ToC_26

[Pro]
http://www.isip.msstate.edu/projects/speech/software/tutorials/production/fundamentals/current/section_06/s06_01_p01.html

[Pro1]
http://www.isip.msstate.edu/projects/speech/software/tutorials/production/fundamentals/current/section_03/s03_01_p01.html

[Pro2]
http://www.isip.msstate.edu/projects/speech/software/tutorials/production/fundamentals/current/section_03/s03_01_p01.html

[Prin]

http://www.ub.es/labfon/princip.htm

R

[Rah]
http://www.gtc.cps.unizar.es/~eduardo/investigacion/voz/rahframe.html

[Rahf]

http://www.gtc.cps.unizar.es/~eduardo/investigacion/voz/rahframe.html

[Reco]
http://cslu.cse.ogi.edu/tutordemos/nnet_recog/intro.gif

S

[Sig]
http://eo.ccu.uniovi.es/llamaquique/virtual/docencia/websignal/introd/introducion.htm

[Stu]
http://studies.ac.upc.es/

[Sn1]
http://www.ub.es/geocrit/sn/sn-146(136).htm

T

[Tec]
http://www.cienciadigital.net/abril2001/tecnologia.html

[Trans]

http://www3.labc.usb.ve/EC4514/AUDIO/Sistema%20Auditivo/Mecanismo_de_transduccion.html

[Trans1]
http://www.inf-cr.uclm.es/www/cbravo/is/tema10_2xh.pdf

V

[Vdom]
http://jamillan.com/v_domotica.htm