UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ciencias y Humanidades

RECONOCIMIENTO DE VOZ CON PROCEDIMIENTOS EN EL REGIMEN DE FRECUENCIAS EN UN SISTEMA CON APOYO DE *HARDWARE*

Guatemala
2002

RECONOCIMIENTO DE VOZ CON PROCEDIMIENTOS EN EL REGIMEN DE FRECUENCIAS EN UN SISTEMA CON APOYO DE *HARDWARE*

UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ciencias y Humanidades

RECONOCIMIENTO DE VOZ CON PROCEDIMIENTOS EN EL REGIMEN DE FRECUENCIAS EN UN SISTEMA CON APOYO DE *HARDWARE*

Trabajo de investigación presentado para optar al grado académico de Licenciatura en Ingeniería Electrónica

Presentado por: Jorge Bolívar Díaz Schneider.

Guatemala
2002

Vo.Bo. Dr. Manuel López.
Asesor

TRIBUNAL EXAMINADOR

Dr. Roberto Molina

MSc. Ricardo Quan.

Dr. Manuel López.

FECHA DE APROBACIÓN: martes 18 de junio de 2002.

CONTENIDO

LISTA DE TABLAS

LISTA DE ILUSTRACIONES

LISTA DE GRÁFICAS

CAPITULO I

MARCO CONCEPTUAL

A. Antecedentes

El reconocimiento de voz ha constituido durante la última década una de las áreas de mayor investigación en la ciencia moderna. El interés por lograr el reconocimiento de patrones de voz humana resurgió gracias a los grandes avances suscitados en el campo del procesamiento de datos, ya que para lograr un rápido y efectivo reconocimiento, deben procesarse cantidades considerables de datos, en un espacio reducido de tiempo.

Los sistemas de reconocimiento de voz se aplican mayormente en programas de procesamiento de palabras que se valen de las herramientas multimedia de las computadoras personales y permiten al usuario dictar las palabras en lugar de escribirlas en el teclado. Sin embargo, existe un gran número de aplicaciones en las que los sistemas semiautomáticos de instrucciones verbales optimizarían significativamente los procesos involucrados. En este sentido, el trabajo del sistema de reconocimiento de voz es servir de apoyo en situaciones en las que se prefiere que una persona mantenga sus manos ocupadas en la tarea que está efectuando, sin tener que distraerlas de la misma para conseguir herramientas o material de apoyo. Esta necesidad se pone de manifiesto por ejemplo en quirófanos, donde a menudo se necesita el ajuste de iluminación durante intervenciones quirúrgicas. También en laboratorios clínicos que realizan exámenes que involucran conteos celulares, y en diversas industrias que operan con sistemas hombre máquina y que requieren del operador el ajuste de los mismos, según las variaciones en la tarea que se realiza.

En todos estos ejemplos, un sistema semiautomático fundamentado en el reconocimiento de instrucciones verbales cortas contribuiría a solucionar las dificultades propias de cada situación.

La idea de reconocer patrones de voz a partir de su espectro de frecuencias tomó forma luego de observaciones empíricas que dieron lugar a pensar que, en un espectro de

frecuencias, existe suficiente información categórica y de fácil análisis que debería hacer posible la discriminación entre instrucciones.

Actualmente se utilizan como herramientas de análisis de instrucciones verbales los modelos matemáticos de procesos estocásticos, la probabilidad estadística, la predicción lineal, etc. Algunas de las cuales utilizan el análisis espectral como herramienta para determinar las características principales de la instrucción analizada (ver desarrollo en marco teórico sobre principales herramientas de reconocimiento de patrones de voz). El reconocimiento basado únicamente en espectros de frecuencias reviste especial importancia al tomar en cuenta la facilidad con que permite identificar las frecuencias que constituyen cada instrucción.

El análisis espectral ha sido empleado también en el desarrollo de técnicas eficientes de filtrado que permiten la eliminación de componentes de ruido no deseados, que son inherentes a la voz humana (Deller, 2000). El espectro permite discernir con claridad cuáles frecuencias están siendo atenuadas por el filtro y cuáles son preservadas.

Con base en lo expuesto, se argumenta que no se ha desarrollado una técnica de reconocimiento de patrones de voz que utilice exclusivamente el espectro de frecuencias correspondiente.

B. Justificación

Las observaciones empíricas previamente desarrolladas esbozan la posibilidad de realizar un reconocimiento de voz con base en los espectros de frecuencia generados por cada instrucción, ya que los patrones obtenidos son bastante característicos. Sin embargo, para poder concluir en forma categórica sobre los alcances y posibilidades del reconocimiento con base en espectros de frecuencia se hace necesario realizar un análisis más extenso, que es precisamente uno de los objetivos de este trabajo de investigación. Es necesario determinar si espectros de instrucciones diferentes son lo suficientemente diferenciables como para lograr una tasa satisfactoria de reconocimiento.

El reconocimiento de voz basado en el análisis de espectros de frecuencia manifiesta una serie de ventajas cuyas aplicaciones prácticas pretende explorar el presente trabajo. En primer lugar, el estudio se limita al espectro de la voz humana, lo que resulta en un menor número de muestras para analizar que si se trabajara la instrucción en términos de amplitud y tiempo. Por otra parte, la información de componentes frecuenciales, al incluir las principales características de una instrucción verbal, permite la posibilidad del desarrollo de técnicas matemáticas más sencillas para lograr el reconocimiento (ver desarrollo de técnicas de reconocimiento en marco teórico).

Se pretende fomentar el desarrollo de sistemas de reconocimiento rápido y confiable de instrucciones verbales cortas, capaz de adaptarse a diversas aplicaciones (como seguridad, domótica (estudio, diseño e implementación de sistemas de automatización de viviendas), medicina, talleres especializados, bodegas, etc.), que compartan la necesidad de utilizar este tipo de instrucciones para optimizar su funcionamiento. Se desea sentar las bases para dichos procedimientos de reconocimiento de voz en principios de procesamiento de señales digitales que permitan la caracterización exitosa de patrones frecuenciales de voz. Resulta importante también determinar el grado en que estas técnicas permiten:

1. realizar el análisis y reconocimiento en el régimen de frecuencias.
2. facilitar la caracterización e identificación de patrones.
3. lidiar con el efecto de ciertas condiciones de variabilidad en la pronunciación de las instrucciones verbales (tales como variación en amplitud, en duración de pronunciación y en el inicio de pronunciación).

C. Objetivos

1. Objetivo General

Determinar la viabilidad del desarrollo de técnicas de reconocimiento de voz basadas exclusivamente en los espectros frecuenciales propios de cada instrucción verbal.

2. Objetivos Específicos

2.1 Determinar el grado de similitud entre espectros de muestras de una misma instrucción, así como el grado de diferenciabilidad con espectros de muestras de instrucciones diferentes.

2.2 Determinar si es posible la diferenciación de instrucciones verbales, por medio de un conjunto de variables que describan los respectivos espectros frecuenciales.

2.3 Dar a conocer las principales condiciones que permiten realizar el reconocimiento de patrones de voz por medio del análisis de sus espectros frecuenciales correspondientes.

2.4 Esbozar las áreas y principios en los cuales deberán fundamentarse futuros procedimientos de reconocimiento de voz basados en espectros frecuenciales.

2.5 Determinar el grado de influencia que las condiciones de degradación propias del proceso de generación de la voz tienen en el análisis del espectro frecuencial correspondiente.

El estudio se limitará al análisis de 10 instrucciones verbales cortas, con una duración aproximada de 0.5 segundos. El espectro frecuencial de cada instrucción se obtendrá por la aplicación directa de la transformada rápida de Fourier. El análisis se hará del espectro de amplitud vs. frecuencia.

CAPITULO II

MARCO TEÓRICO

Los conceptos desarrollados en el siguiente marco teórico están basados en el análisis y conclusiones expuestos por John Deller Jr., John Hansen y John Proakis en su libro: *Discrete-Time Processing of Speech Signals*, así como en los conceptos expuestos en el libro: *Computer Speech Processing*, editado por Frank Fallside y William Woods.

A. Conceptos fundamentales de la ciencia del habla

Sobre el problema de reconocimiento de voz, Deller (2000) afirma que:

> «El primer paso para resolver el problema de reconocimiento de voz es comprender su complejidad.».

Un análisis profundo y detallado de los principios que hay detrás de la comunicación hablada resulta necesario y constituye la base del correcto desarrollo de sistemas de reconocimiento de voz.

El habla es un proceso dinámico para transmitir información. Dicha información es de compleja naturaleza. Se identifican 3 tipos de información que se transmiten en forma simultánea:

1. Información lingüística: Este tipo de información se refiere al significado del mensaje hablado.
2. Información sociolingüística: Este tipo de información es propio de los mensajes hablados y permite determinar la procedencia del hablante. Indica si proviene de una región geográfica, cultura o clase socioeconómica en particular. Se manifiesta generalmente en forma de acento y modismos característicos.
3. Información personal: Indica aspectos adicionales de la identidad del hablante, a menudo permite identificar características como el estado de ánimo de la persona, etc. En general, esta información se manifiesta en el tono y vocabulario que cada persona emplea al comunicarse.

Para lograr la comunicación, el hablante debe producir una señal de habla en forma de onda de presión de sonido, la cual viaja desde la boca del hablante (o emisor) hasta el oído del receptor. Las señales propias del habla se componen de una secuencia de sonidos que sirven como una representación simbólica de una idea que el hablante desea transmitir. El arreglo de estos sonidos está gobernado por reglas asociadas con cada idioma, al estudio de dichas normas y su uso en comunicaciones orales se le denomina **lingüística**. Debe hacerse una diferencia entre la lingüística y la **fonética**, la cual se encarga del estudio de las características de la producción del sonido humano, específicamente orientado a la descripción, clasificación y transcripción del habla. Las diferentes técnicas de reconocimiento de voz se fundamentan en ambas ciencias. Sin embargo, en esta sección se presentan principios relacionados con la fonética. Se deja la descripción de los principios lingüísticos para cuando se introduzcan los sistemas avanzados de reconocimiento de voz.

1. Comunicación hablada

El proceso de comunicación hablada consiste en la transferencia de información de una persona a otra. Primero, el emisor concibe una idea, la cual desea transmitir al receptor. El emisor convierte esta idea a través de una serie de procesos neurológicos y movimientos musculares en una onda acústica de presión de sonido, la cual viaja a través de un medio (típicamente aire) y es recibida por el sistema auditivo del receptor. El receptor procesa entonces esta onda de sonido y la convierte de nuevo en señales neurológicas que interpreta su cerebro. Además, el sistema auditivo del emisor permite, a través de una retroalimentación, que el habla se ejecute en la mejor forma posible (el emisor puede darse cuenta si la onda acústica producida corresponde con la idea que deseaba transmitir). Es evidente que el sistema auditivo humano juega un papel importante en el proceso de comunicación hablada.

2. Anatomía y Fisiología del "sistema de producción del habla"

La forma de onda del habla es una onda acústica de presión de sonido que se origina del movimiento voluntario de estructuras anatómicas (denominadas **articulado-**

res) que constituyen lo que se denomina el "sistema de producción del habla". Las partes del sistema se detallan en la figura 1.

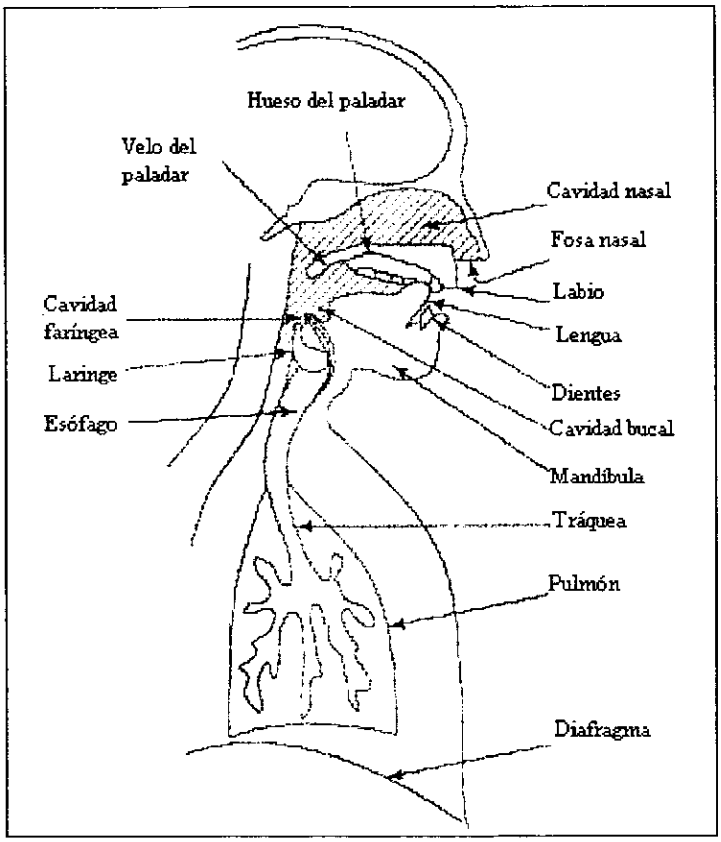


Figura No. 1 Diagrama esquemático del "sistema de producción del habla" (corte sagital). Reproducido de: John Deller et al. *Discrete-Time Processing of Speech Signals*, IEEE Press. E.E.U.U., 2000. Pág. 102.

Para comprender mejor el funcionamiento del sistema de producción del habla es de utilidad visualizarlo como una serie de operaciones de filtrado acústico. En base a esta asociación, las tres cavidades principales del sistema (la cavidad faríngea, nasal y bucal) constituyen el filtro acústico principal, el cual es excitado por los órganos que se encuentran debajo de él (pulmones y laringe).

Es posible visualizar el funcionamiento del sistema de producción del habla con el modelo mostrado en la figura 2.

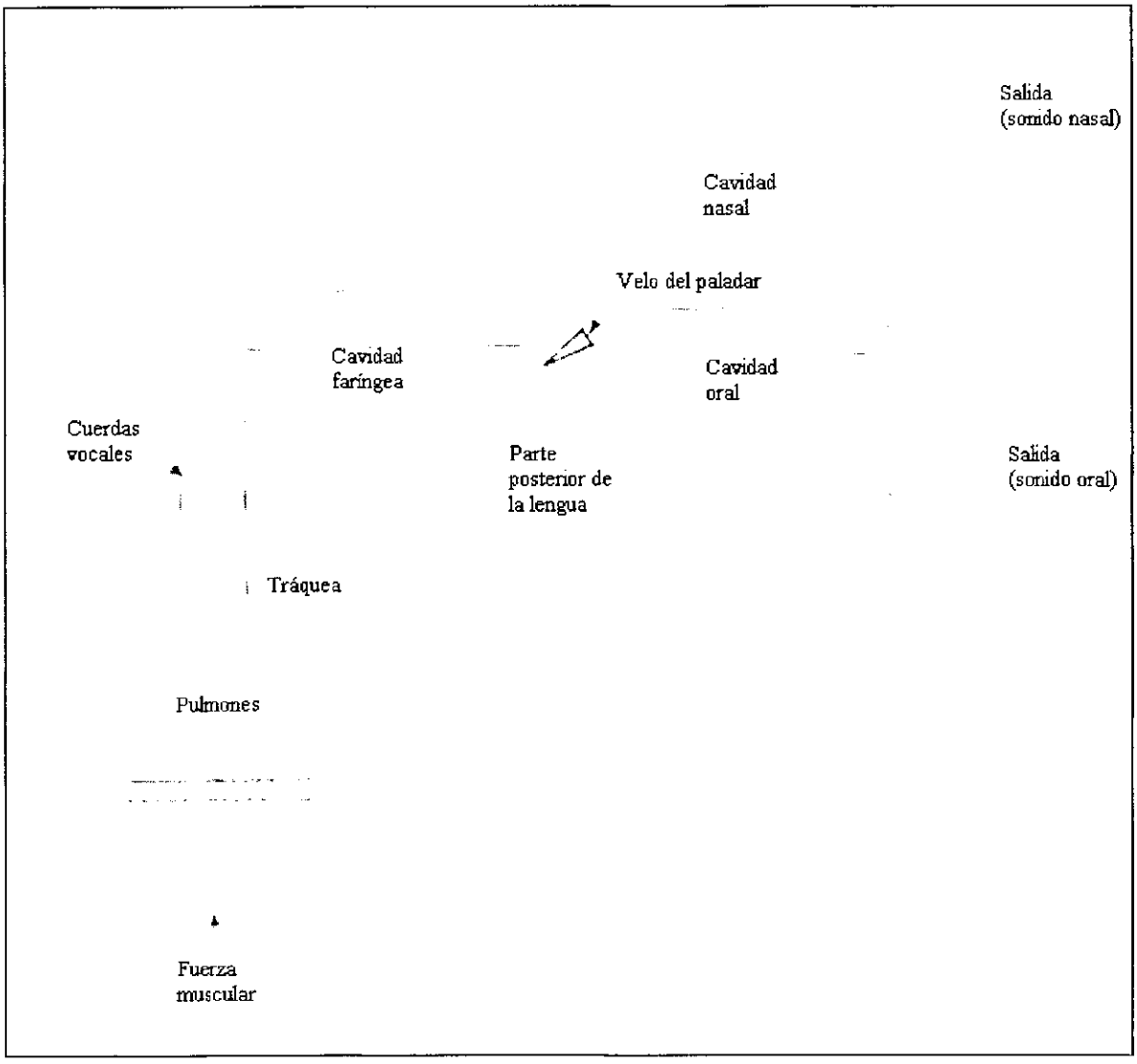


Figura No. 2 Diagrama de bloques del funcionamiento del "sistema de producción del habla". Reproducido de: John Deller et al. *Discrete-Time Processing of Speech Signals*. IEEE Press. E.E.U.U., 2000. Pág. 103.

Las tres cavidades principales constituyen la estructura resonante del habla humana. En el adulto promedio, la longitud total del "tracto vocal" (que comprende desde la laringe hasta los labios) es de 17 cm. en los hombres, y de 14 cm. en las mujeres. En el niño promedio, el tracto mide aproximadamente 10 cm. El reposicionamiento de los articuladores del tracto vocal ocasiona que el área transversal del mismo varíe desde cero (cierre total) hasta aproximadamente 20 cm^2. El tracto nasal constituye una salida auxiliar para la transmisión de sonido, con una longitud aproximada de 12 cm. en adultos.

El acople entre las cavidades nasal y bucal es regulado por el tamaño de la apertura en el velo del paladar. Una de las principales estructuras anatómicas que interviene en la producción del habla es la laringe. Ésta provee una excitación periódica al sistema, en cuyo caso los sonidos emitidos se conocen como "vocalizados". Dicha excitación periódica se debe al paso de una corriente de aire a través de dos pliegues elásticos de músculos y membranas mucosas que se extienden desde el cartílago tiroides hasta los aritenoides, denominados cuerdas vocales. El paso de la corriente de aire hace que las cuerdas vocales vibren. La apertura entre las cuerdas vocales se conoce como glotis y es regulada de forma voluntaria. Cuando la apertura entre ambas cuerdas es lo suficientemente grande, el aire pasa libremente sin generar una excitación periódica, en cuyo caso se producen sonidos denominados "no vocalizados". Es importante recalcar que la laringe no solo consta de las cuerdas vocales, sino que consiste en un órgano complicado (desde el punto de vista anatómico), formado por varios ligamentos y cartílagos, de los cuales los principales se muestran en la figura 3.

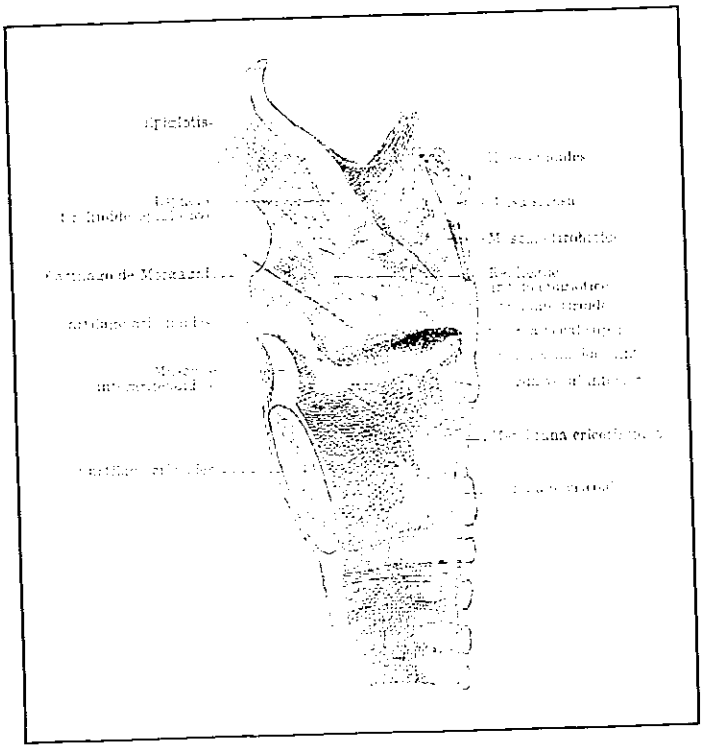


Figura No. 3 Mitad lateral izquierda de la laringe, vista por su cara interna (corte sagital). Reproducido de: Rouviere, H. Compendio de Anatomía y Disección. Tercera Edición Española. Salvat Editores, S.A. España, 1971. Pág. 210.

Un análisis anatómico profundo escapa al objetivo del presente trabajo. Sin embargo es importante notar que gracias a diversas terminaciones nerviosas y músculos se logra la adecuada posición de las cuerdas vocales y demás cartílagos de la laringe que intervienen en el proceso del habla humana.

El habla no es una cadena de sonidos discretos "bien formados", por el contrario, constituye una serie de sonidos "objetivo" (a menudo muy breves) con transiciones intermedias, lo que se refleja en características espectrales variantes en el tiempo. Es posible dividir al habla en segmentos de sonido que poseen propiedades acústicas similares, en intervalos relativamente cortos de tiempo. Los sonidos que constituyen el habla pueden dividirse en dos amplias categorías: vocalizados (típicamente las vocales) y no vocalizados (las consonantes). Los primeros se caracterizan por vibraciones periódicas de las cuerdas vocales y casi ninguna restricción al paso del aire a través del tracto vocal (patrones de frecuencia bastante bien delimitados). Los segundos se caracterizan por la ausen-

cia de vibración en las cuerdas vocales y constituyen únicamente el resultado de diferentes constricciones del tracto vocal que crean un flujo turbulento al forzar una corriente de aire a través de ellos (tienen patrones de frecuencia poco definidos, con "características de ruido"). De esta manera, todo cambio en la onda del habla es una consecuencia directa del movimiento voluntario de los articuladores del sistema, los cuales raramente se mantienen fijos durante un tiempo prolongado. Las limitantes naturales del sistema auditivo y del tracto vocal restringen el ancho de banda de la comunicación humana hablada el que oscila entre 7 y 8 kHz. aproximadamente.

Los articuladores son también responsables de regiones de énfasis producidas al hablar, que se manifiestan como frecuencias resonantes en el espectro. Estas frecuencias son producto de las diversas cavidades acústicas formadas a lo largo del tracto vocal por los articuladores. Desde el punto de vista de un modelo sistemático, estas frecuencias pueden ser vistas como las frecuencias resonantes de los diferentes filtros que forman el sistema productor del habla. En general, a estas frecuencias resonantes que tienden a formar el espectro del habla se les conoce como "formantes". En principio, existe un número infinito de formantes en un sonido dado, aunque en la práctica es común limitarse a 3 ó 5, dada la escasa información adicional que presentan formantes adicionales.

3. Excitación del sistema productor del habla

Se identifican dos tipos elementales de excitación: vocalizada y no vocalizada. Sin embargo, para propósitos de diseño de modelos y clasificación suelen delimitarse cuatro tipos adicionales de excitación: mixta, "plosiva", suspirada y silenciosa. Cuando la excitación es vocalizada, los sonidos se producen al forzar aire a través de la glotis. Se ajusta la tensión en las cuerdas vocales para que vibren en forma oscilatoria. La obstrucción periódica del flujo de aire produce soplos casi periódicos de aire que excitan el tracto vocal. Este tipo de sonidos resulta ser de los más interesantes, ya que involucra el uso de un órgano especializado (la laringe). La secuencia de eventos que ocurren en la laringe durante un ciclo de interrupción del flujo de aire se muestra en la figura 4. Ejemplos de sonidos vocalizados son las vocales.

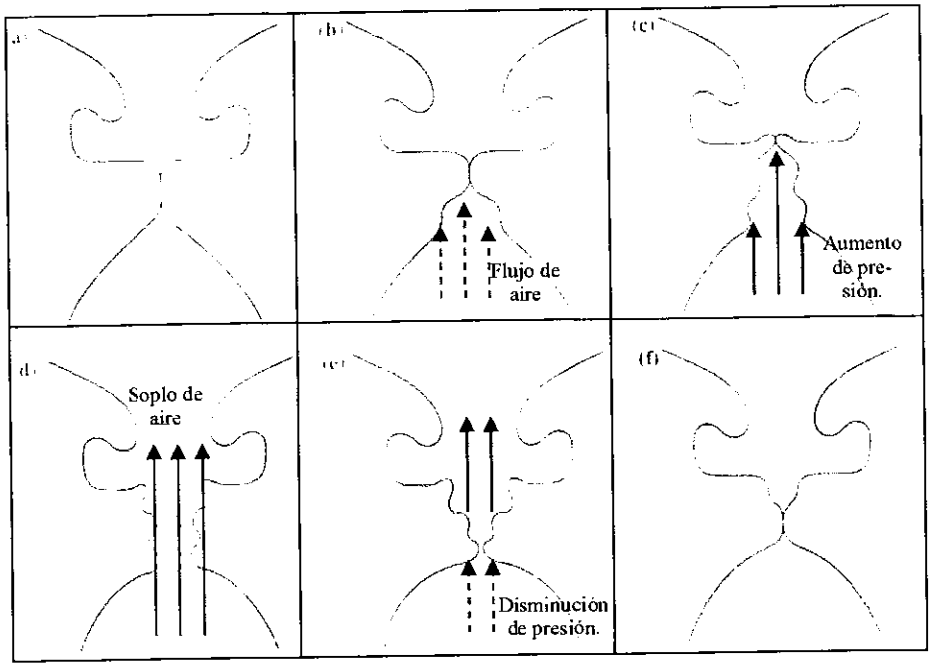


Figura No. 4 Secuencia de cortes longitudinales de la laringe que ilustra un ciclo de interrupción del flujo de aire proveniente de los pulmones. (a) Estado relajado, no hay flujo de aire. (b) Deformación de pliegues vocales por presión ejercida por flujo de aire. (c) Incremento de deformación debido al aumento de presión. (d) Expulsión de aire (soplo) causado por exceso de presión que vence la resistencia de los tejidos de la laringe. (e) Inicio de cierre de apertura entre pliegues (disminuye presión después de la liberación del soplo). (e) Restauración de condición normal de los pliegues (ausencia total de presión subglotal). Reproducido de: John Deller et al. *Discrete-Time Processing of Speech Signals*, IEEE Press. E.E.U.U., 2000. Pág. 111.

Se denomina período fundamental T_0 al tiempo entre aperturas sucesivas de los pliegues vocales, de donde a la razón de vibración se le denomina frecuencia fundamental F_0 o *"pitch"*. Dos factores afectan la frecuencia fundamental de un sonido: la acentuación tónica y la entonación empleada por el orador. El primero se caracteriza por cambiar el valor de la frecuencia fundamental y su intensidad, para manifestar cambios en el énfasis de ciertas sílabas o palabras. La entonación se refiere al contorno de la frecuencia fundamental en el tiempo, que permite señalizar la estructura gramatical de un idioma. Los

rangos típicos de frecuencia fundamental oscilan entre 50 y 250 Hz. en hombres y entre 120 y 500 en mujeres.

Por otro lado, cuando la excitación es no vocalizada, los sonidos son producidos por constricciones a lo largo del tracto vocal. Al forzar aire a través de esta constricción, se produce un flujo turbulento que caracteriza a este tipo de sonidos. Ejemplo de sonidos no vocalizados son la mayoría de consonantes, específicamente el sonido de la letra "s". En el caso en que los sonidos muestran características de excitación tanto vocalizada como no vocalizada, se denominan mixtos.

4. "Fonémica" y Fonética

La unidad teórica básica utilizada para definir la manera en que el habla comunica significados lingüísticos es el fonema. Los fonemas están constituidos por vocales, "semivocales", diptongos y consonantes (nasales, pausas, fricativas, africativas). Cada fonema se considera un código constituido por un conjunto único de "gestos de articulación", que incluyen el tipo y ubicación de la excitación, así como la posición y movimiento de los articuladores del tracto vocal.

Para distinguir entre fonémica y fonética conviene diferenciar primero "fonos" de fonemas. El concepto de fonema constituye una unidad de sonido ideal, que no es reproducida con exactitud por todos los hablantes. A los sonidos producidos que se asemejan grandemente al fonema (o unidad ideal) se les llama "fonos"; es decir, los fonos constituyen los sonidos que realmente son producidos por los diferentes hablantes. La fonética se restringe entonces al estudio de los fonemas, mientras que la fonémica al estudio de los fonos. Es claro que para cada fonema existe un conjunto de fonos asociados, que incluyen todas las variaciones del fonema original permisibles. A estos fonos se les denomina "alófonos" del fonema. Los alófonos representan entonces la libertad permisible al hablar determinada lengua.

5. Clasificación de fonemas

Existen diversos métodos para la clasificación de fonemas, generalmente se agrupan en base a propiedades relacionadas con la forma de onda en el tiempo, las características frecuenciales, la forma de articulación, y el lugar de articulación o tipo de excitación. Para efectos de esta investigación se clasificarán los fonemas según la forma de excitación del tracto vocal, como fonemas vocalizados y consonantes o no vocalizados.

a. Fonemas vocalizados

1) Vocales

Constituyen los fonemas de mayor longitud, que puede variar de 40 a 400 mseg. según su tipo. Las frecuencias formantes de cada vocal están determinadas por la variación del área de la sección transversal a lo largo del tracto vocal, que a su vez son determinados por la posición de la parte posterior de la lengua y del grado de constricción que la posición de ésta ofrece al flujo de aire. La forma de onda de una vocal es casi periódica, debido al movimiento cíclico de los pliegues vocales en la laringe, que sirven como excitación del sistema.

La experimentación ha permitido formular 6 reglas que relacionan el efecto que las características del tracto vocal tienen en las frecuencias formantes de las vocales, lo que permite asociar una característica con cada uno de los formantes:

1. Regla de la longitud: Las frecuencias que forman cada vocal (formantes) son inversamente proporcionales a la longitud del tracto faríngeooral; es decir, mientras más largo es el tracto, menores serán las frecuencias formantes promedio.
2. Regla del primer formante (F1), constricciones orales: la frecuencia del primer formante disminuye en presencia de cualquier constricción en la mitad frontal de la cavidad bucal.

3. Regla del primer formante (F1), constricciones faríngeas: la frecuencia de F1 aumenta en presencia de constricciones de la faringe; a mayor constricción, mayor la frecuencia de F1.
4. Regla del segundo formante (F2), constricciones de la parte posterior de la lengua: la frecuencia de F2 tiende a disminuir en presencia de una constricción de la parte posterior de la lengua.
5. Regla del segundo formante (F2), constricciones de la parte frontal de la lengua: la frecuencia de F2 aumenta en presencia de constricciones de la parte frontal de la lengua.
6. Regla de la redondez de los labios: La frecuencia de todos los formantes disminuye cuando se redondean los labios.

Es importante recalcar que las reglas hacen referencia a los dos primeros formantes, que son los más significativos e importantes al diferenciar entre distintos sonidos vocalizados. En la práctica, nunca se analizan más de cuatro formantes, pues sus efectos pueden despreciarse en comparación con el ruido presente en el ambiente. Es importante mencionar que también existe una amplia variabilidad entre las características de los formantes de una vocal al comparar pronunciaciones de varios hablantes. Un ejemplo de la forma del tracto vocal para la pronunciación de vocales se muestra en la figura 5.

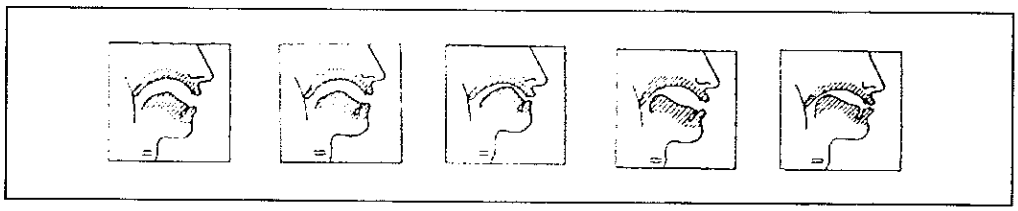


Figura No. 5 Perfil del tracto vocal. Muestra de izquierda a derecha constricciones en la pronunciación de las vocales "a", "e", "i", "o" y "u" respectivamente. Reproducido de: John Deller et al. *Discrete-Time Processing of Speech Signals*, IEEE Press. E.E.U.U., 2000. Pág. 122-123.

2) Diptongos

Constituyen otro tipo de fonema vocalizado. Un diptongo involucra un movimiento intencional de una vocal hacia otra, es decir, un movimiento voluntа-

rio de los articuladores. Existe aún cierta ambigüedad en la definición de un diptongo, pues a menudo se confunde con la pronunciación de dos vocales consecutivas. Para resolver esto, se considera que en un diptongo existen dos vocales "objetivo", de las cuales la pronunciación de la primera es más larga que la segunda, pero el tiempo de transición entre ambas vocales "objetivo" es mayor que el tiempo de pronunciación de cualquiera de las dos vocales.

3) Semivocales

Se dividen en "líquidas" y "deslizamientos"; las primeras constituyen sonidos con características espectrales parecidas a las de las vocales, pero con menor intensidad, debido a mayores constricciones a lo largo del tracto vocal. Los "deslizamientos" constituyen variaciones voluntarias en la pronunciación de ciertas vocales, especialmente en el caso de vocales terminales de una palabra.

b. Consonantes o fonemas no vocalizados

Representan sonidos cuya emisión implica un mayor grado de constricción en el tracto vocal que el propio de los fonemas vocalizados. Algunas consonantes pueden requerir de un movimiento dinámico preciso de los articuladores para su producción, mientras que otras pueden no requerir mayor movimiento de articuladores. Se identifican 4 tipos de consonantes:

1) Fricativas

Se producen al excitar el tracto vocal con un flujo de aire constante que se convierte en turbulento en algún punto de constricción. Es la ubicación del punto de constricción lo que se utiliza para clasificar las fricativas en: labiodental (dientes superiores con labio inferior), interdental (lengua detrás de dientes frontales), alveolar (lengua en contacto con el borde de las ensillas), palatal (lengua reposada en el paladar), velar (constricción con el velo del paladar) y glotal (pliegues vocales fijos y tensos).

Es claro que la clasificación anterior hace referencia a la ubicación de la constricción en el tracto. Sin embargo, la excitación del sistema puede ser tanto vocalizada (constricción en cuerdas vocales) como no vocalizada (glotis totalmente abierta, flujo de aire sin constricción a nivel de la laringe). En relación con esto se distingue entre fricativas vocalizadas y no vocalizadas. La figura 6 presenta ejemplos de constricciones en el tracto vocal en el caso de fricativas no vocalizadas.

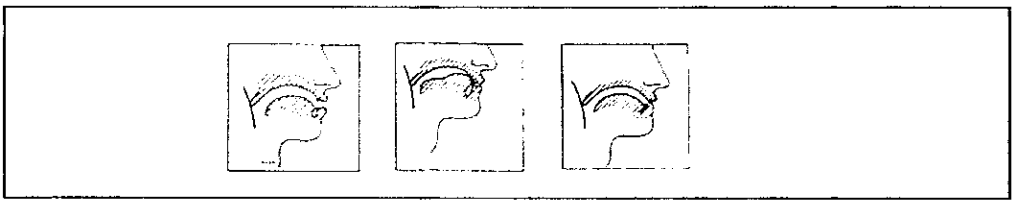


Figura No. 6 Perfil del tracto vocal. De izquierda a derecha las constricciones en la pronunciación de las consonantes "j", "s" y "f" respectivamente. Reproducido de: John Deller, *Discrete-Time Processing of Speech Signals*, IEEE Press. E.E.U.U., 2000. Pág. 133.

2) Pausas, altos o "plosivos"

Sonidos no continuos que son producidos por la acumulación de presión detrás de una constricción total (interrupción completa del flujo de aire) en algún punto a lo largo del tracto vocal, la cual es liberada en forma abrupta. Esta explosión súbita y aspiración de aire es lo que caracteriza a las consonantes "plosivas" o altos. Estas consonantes pueden ser bilabiales, alveolares o velares, según el lugar donde ocurra la constricción total (véase clasificación de fricativas). La figura 7 muestra ejemplos de constricciones en el tracto vocal en el caso de algunas consonantes plosivas.

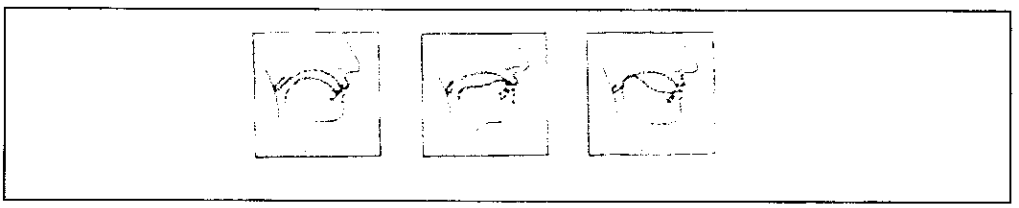


Figura No. 7 Perfil del tracto vocal. De izquierda a derecha las constricciones en la pronunciación de las consonantes "p", "t" y "k" respectivamente. Reproducido de: John Deller, *Discrete-Time Processing of Speech Signals*, IEEE Press. E.E.U.U., 2000. Pág. 135.

3) Nasales

Las consonantes nasales (como la "m" o la "n") son sonidos vocalizados producidos cuando el flujo de aire proveniente de la glotis excita una cavidad nasal abierta y una cavidad oral cerrada. Ejemplos de la forma del tracto vocal para la pronunciación de este tipo de consonantes se muestra en la figura 8.

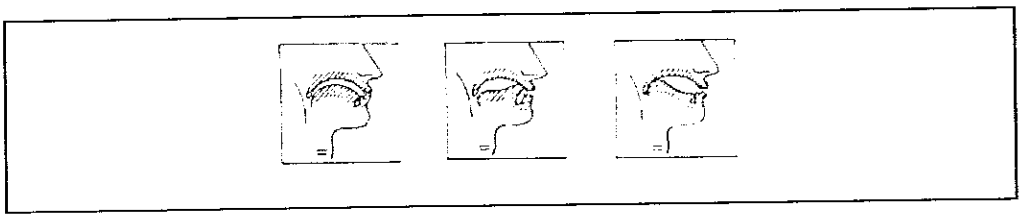


Figura No. 8 Perfil del tracto vocal. De izquierda a derecha las constricciones en la pronunciación de las consonantes "m", "n" y "g" (con vocales fuertes) respectivamente. Reproducido de: John Deller, *Discrete-Time Processing of Speech Signals*, IEEE Press. E.E.U.U., 2000. Pág. 136.

6. Características prosódicas

Al hablar de características prosódicas nos referimos a las reglas que gobiernan la pronunciación de diferentes palabras en un lenguaje. Dichas características se manifiestan en el acento tónico y la entonación de los fonemas. Los principales puntos referentes a la influencia del acento tónico y la entonación se resumen a continuación:

1. Los patrones fonéticos que corresponden al acento tónico y a la entonación se basan en una mezcla de cambios en la tensión de los pliegues vocales en la laringe, así como en intensidades máximas y mínimas de presión de aire y en la duración de vocales y consonantes.

2. El movimiento de la frecuencia fundamental se refleja en cambios en la presión de aire debajo de la glotis y en la tensión de los pliegues vocales. Si uno o ambos de estos factores aumentan, se manifiesta un incremento en la frecuencia fundamental.

7. "Coarticulación"

Este concepto resulta de vital importancia al definir los fundamentos del habla humana, pues hace referencia al cambio tanto en la articulación de los fonemas como en la acústica, que son causados por la influencia de otro sonido en la misma declaración (secuencia de sonidos). Existen dos factores primarios involucrados en los efectos de co-articulación:

1. Las formas específicas del tracto vocal que deben reproducirse.
2. El programa "motor" que se necesita para producir una secuencia de unidades del habla (vocales, consonantes, palabras, Etc.).

La coarticulación se manifiesta como una estructura acústica cambiante, causada por el movimiento de los articuladores así como por el cambio en la duración de los fonemas.

B. Reconocimiento de voz, una visión general

John Deller (2000) describe el objetivo final de la tecnología de reconocimiento de voz como la creación de máquinas capaces de recibir información hablada, procesarla, y actuar apropiadamente según esa información. Puede caracterizarse el problema de reconocimiento de voz como un eslabón más en la búsqueda de soluciones de inteligencia artificial. En este contexto es conveniente reflexionar sobre los límites tecnológicos actuales y definir un marco realista que ponga en perspectiva las palpables dificultades aún existentes para alcanzar este fin.

Deller (2000) divide los sistemas de reconocimiento de voz contemporáneos con desempeños aceptables en 3 amplias categorías:

1. Sistemas con vocabularios reducidos (de 10 a 100 palabras, aproximadamente).
2. Sistemas que detectan palabras pronunciadas aisladamente entre sí (los vocabularios en estos sistemas pueden exceder 10,000 palabras).

3. Sistemas que aceptan un flujo continuo de palabras, que se concentran en la definición de "dominios de tareas" que agrupan mensajes frecuentes en cierto ambiente de trabajo.

La mayoría de sistemas implementados en la actualidad es de vocabularios reducidos o de detección de palabras aisladas. Los sistemas pertenecientes a la tercera categoría permanecen aún en fase experimental. Es importante resaltar que no se cuenta todavía con sistemas lo suficientemente robustos para manejar el problema de señales de ruido en el ambiente. Un sistema de reconocimiento de voz responde mejor si se diseña para reconocer patrones de voz de una sola persona (la cual "entrena" al sistema). Aunque algunos sistemas existentes aprovechan las ventajas de la estructura gramatical del idioma, es sólo a nivel experimental y en un nivel muy primitivo que se han implementado habilidades "cognitivas" como el discernimiento de significados o el aprendizaje de errores.

En contraste con lo anterior, las características deseadas para un buen desempeño de sistemas de reconocimiento de voz tales como pronunciaciones claras y discretas, usuario único, ambiente relativamente silencioso, etc. no son propias de la mayoría de aplicaciones naturales que se beneficiarían al implementar este tipo de sistemas. Esto permite concluir que para obtener un máximo beneficio y aplicabilidad universal de los sistemas de reconocimiento de voz, éstos deben ser capaces de reconocer el habla continua de diferentes usuarios con pronunciación y vocabularios distintos, en ambientes con diferentes niveles de ruido. Los sistemas deberán ser además económicamente accesibles y de tamaño reducido, capaces de operar en tiempo real y de adaptarse a cambios en el idioma (reglas de sintaxis, semántica, etc.). Es claro que, al analizar el problema de reconocimiento de voz desde esta perspectiva, los desarrollos actuales están aún muy lejos de lograr los objetivos finales deseados.

Debe recalcarse que los avances más significativos en este campo de investigación se han logrado en décadas recientes. Como ejemplo de ellos pueden citarse los esfuerzos investigativos de IBM durante la década de los 80s que lograron la implementación de un sistema experimental capaz de reconocer un vocabulario de 20,000 palabras, pronuncia-

das aisladamente (Deller, 2000). Los sistemas actuales de vocabulario reducido pueden emplearse en muchas aplicaciones relativamente sencillas: para mejorar la velocidad con la que se ingresa información a una máquina, en aplicaciones donde no se cuenta con las manos (como por ejemplo, cirugía, cuarto oscuro, cabinas de aeronaves, etc.), o en aplicaciones donde el usuario debe permanecer en contacto remoto con la máquina (por vía telefónica o en ambientes peligrosos).

Los primeros sistemas de reconocimiento de voz fueron desarrollados en la década de los 70s. Se limitaron al reconocimiento de un número muy reducido de palabras en un ambiente relativamente libre de ruido y con un solo usuario. Para finales de la década de los 80s, IBM desarrolló un sistema experimental capaz de reconocer un vocabulario de 20,000 palabras (pronunciadas en forma aislada) ó 5,000 palabras pronunciadas en forma continua. Además, la idea de implementarlo en una computadora personal de reducido tamaño hizo del proyecto algo revolucionario. Es importante recalcar que los desarrollos significativos han sido gracias a los avances en técnicas de procesamiento de datos. Antes de la década de los 70s, en la era en que no se contaba con la ayuda de computadoras personales, no se hicieron avances significativos en el área de reconocimiento de voz.

El primer paso para resolver el problema de reconocimiento de voz fue dado al lograr un entendimiento profundo de su complejidad. La facilidad relativa con la cual los humanos se comunican entre sí a través del habla opaca significativamente la extrema complejidad involucrada en el proceso. Debe recordarse que la ciencia no ha podido explicar a cabalidad el funcionamiento del oído interno; específicamente lo que ocurre después del nervio auditivo. Por lo tanto, la concepción del cerebro y oído humano como un "sistema analizador de espectros" no debe ser tomada como absolutamente cierta, aunque para simplificar el problema, se asume como verdadera.

Según lo anterior, es evidente que el desarrollo de sistemas de reconocimiento de voz ha sido paralelo a los avances en el cómputo digital, fundamentado en circuitos integrados en gran escala. Desde los años 80, la combinación de velocidad de cómputo,

abundante memoria y arquitecturas especializadas para el procesamiento de señales han permitido la ejecución de algoritmos enormemente complejos.

C. El problema de reconocimiento de voz

En la sección anterior se esbozaron algunas de las principales dificultades prácticas con las que se debe lidiar en el desarrollo de un sistema de reconocimiento de voz. A continuación se trabaja el problema de reconocimiento de voz y se describen los principales factores que determinan su éxito o fracaso:

1. Reconocimiento dependiente o independiente del orador

Este factor se refiere al requerimiento de identificación de un individuo específico (independiente del orador) o a múltiples individuos (dependiente del orador) en el proceso de reconocimiento de voz. La desventaja aparente de los sistemas que dependen del orador es la necesidad de "entrenar" el sistema cada vez que lo utilice un nuevo orador. La naturaleza del sistema la dicta generalmente la aplicación a la cual se pretende adaptar.

2. Tamaño del vocabulario

Es claro que el desempeño del sistema se degrada al aumentar el tamaño del vocabulario. Se estima que la dificultad del reconocimiento se incrementa en forma logarítmica con el tamaño del vocabulario. Los requerimientos de memoria también aumentan con el tamaño del vocabulario, no tanto como consecuencia del número de palabras, sino por la complejidad de la tarea de reconocimiento. Los sistemas de reconocimiento de voz se clasifican generalmente como de vocabulario reducido (1-99 palabras), mediano (100-999 palabras) o elevado (1000 palabras o más).

Los sistemas de vocabulario reducido son los más disponibles en la actualidad y se utilizan en tareas tales como reconocimiento de números de teléfono o de tarjetas de crédito, así como en aplicaciones de selección de paquetes para envío. El enfoque que se les ha dado a los sistemas de vocabulario mediano ha sido en experimentos de laboratorio,

con fines puramente investigativos. Los sistemas de vocabulario elevado se han utilizado en el desarrollo de productos comerciales, diseñados para aplicaciones tales como correspondencia de oficina y extracción de documentos. Estos sistemas han sido del tipo de palabras aisladas (ver siguiente inciso). El tamaño del vocabulario es entonces una de varias medidas de dificultad del problema de reconocimiento de voz.

Para vocabularios reducidos se emplean a menudo estrategias de reconocimiento de palabras únicas, fundamentadas en la búsqueda exhaustiva en una base de datos con los patrones a reconocer. Este tipo de técnicas de reconocimiento no es apto cuando se incrementa el número de palabras, por lo que se recurre a modelos de subunidades (sílabas o fonemas).

3. Reconocimiento de palabras aisladas versus reconocimiento de habla continua

Se hace referencia aquí a la técnica de reconocimiento empleada, que depende del tipo de ingreso de parámetros del sistema. Se identifican 3 técnicas principales:

a. Reconocimiento de palabras aisladas (IWR por sus siglas en inglés)

También denominado "reconocimiento de pronunciación discreta". Se caracteriza porque, en la fase de reconocimiento se asume que el orador pronuncia en forma deliberadamente oraciones con pausas lo suficientemente grandes (típicamente de 200 milisegundos) entre cada palabra, de tal manera que los silencios no se confunden con elementos de la pronunciación de cada palabra (Deller, 2000). Este hecho permite simplificar grandemente la tarea de reconocimiento. Las fronteras entre cada palabra pueden localizarse de diversas formas, incluyendo algoritmos de detección de punto final. Es importante recalcar que este tipo de reconocimiento está sujeto a ciertas dificultades que resultan ser subjetivas y dependen del orador. Entre éstas podemos mencionar las pausas entre palabra y palabra, la pronunciación errónea y la inclusión de palabras no usuales en el vocabulario corriente. El procedimiento deberá por lo tanto dotarse de cierta flexibilidad para lidiar con estas dificultades.

b. Reconocimiento de habla continua (CSR por sus siglas en inglés)

En este caso el usuario (u orador) pronuncia el mensaje en forma continua, sin acentuar excesivamente las pausas entre palabras. Este tipo de sistemas debe primero lidiar con el problema de fronteras temporales desconocidas en la señal acústica. Luego, debe ser capaz de desarrollarse correctamente en la presencia de imperfecciones e irregularidades de pronunciación que acompañan al habla continua (lo que incluye el hecho de que no todas las consonantes y vocales se pronuncian de la misma manera en cada palabra). Es obvio que los algoritmos desarrollados para este tipo de reconocimiento deben ser mucho más robustos que aquellos propios del reconocimiento de palabras aisladas.

c. Reconocimiento de habla conectada (*Connected-speech recongnition*)

Esta técnica de reconocimiento es a menudo considerada como un caso específico del reconocimiento de habla continua, pues tiene como parámetro de ingreso una pronunciación continua de palabras. Su nombre (habla conectada) hace referencia a la estrategia de reconocimiento que utiliza.

En esta forma, la oración es decodificada mediante la unión de diferentes modelos construidos a partir de palabras discretas, al comparar la expresión completa con la concatenación de dichos modelos. Nuevamente se requiere de una colaboración por parte del orador para lograr mejores resultados.

4. Restricciones lingüísticas

Este factor constituye el problema más abstracto involucrado en el reconocimiento de voz. Constituye la inclusión de normas lingüísticas en el reconocedor de voz, es decir, la construcción de modelos lingüísticos incorporados en el sistema de reconocimiento. Las restricciones lingüísticas se refieren a la manera en que las unidades fundamentales del idioma se concatenan en orden y contexto para formar expresiones con significado y sentido.

El verdadero reto aquí es lograr un balance entre la necesidad de maximizar el grado de inclusión de las restricciones y el deseo de minimizar las limitaciones impuestas a la libertad de expresión del orador (pues es claro que mientras más restricciones se incluyan, más limitado será el número de mensajes admisibles por el reconocedor). El grado de limitación impuesto a la libre expresión del orador en un modelo de lenguaje se denomina "perplejidad o ha*bit*abilidad". La perplejidad representa el número de posibilidades disponibles en un punto de decisión, cuando el proceso de decodificación del mensaje se visualiza como una toma de decisiones secuenciales a lo largo de un grupo de expresiones permisibles previamente definidas. Dentro de este contexto se define un modelo que permite la identificación de cuatro tipos de componentes del código de lenguaje natural denominado Modelo de Lenguaje de Peirce. Este modelo se incluye en el Anexo 1 de este trabajo.

5. Nivel de ambigüedad y de confusión audible

El grado de similitudes entre palabras del vocabulario afectará el adecuado funcionamiento del sistema reconocedor de voz. Se identifican dos tipos de similitudes:

1. Ambigüedades. Palabras indistinguibles acústicamente (por ejemplo, asta y hasta), diferenciadas únicamente por contexto o leves diferencias prosódicas. Generalmente se resuelve este tipo de similitudes con procedimientos de alto nivel.

2. Confusión audible. En este caso se hace referencia al grado con el cual se pueden confundir palabras entre sí, dada su similitud acústica. Puede lidiarse con este tipo de similitudes en procedimientos de alto nivel, aunque en teoría podría solucionarse a nivel acústico.

6. Ruido ambiental

Este último factor enfrenta el problema de tener que implementar sistemas de reconocimiento de voz en ambientes fuera de los laboratorios experimentales, donde las condiciones de ruido ambiental pueden reducirse en un alto porcentaje. Existen varias fuentes de ruido ambiental tales como: habla de otras personas que no son el orador del

sistema, sonido de equipo, equipos de aire acondicionado o de iluminación fluorescente, etc. También puede ser que el propio orador sea la fuente de ruidos indeseables, como en el caso de respiraciones fuertes, tos, etc. El sistema debe ser robusto y poder lidiar eficientemente con el ruido ambiental. Generalmente se cuenta con cierta información *a priori* del ambiente en el cual va a operar el sistema, lo que permite implementar mejoras específicas.

Para finalizar con el problema de reconocimiento de voz, se hace hincapié en los sistemas llamados de reconocimiento del orador y de verificación del orador. Ambos sistemas emplean muchas de las técnicas anteriormente descritas, pero tienen un objetivo bastante distinto al de los sistemas de reconocimiento de voz. En el primero, se pretende identificar quién es el orador, a partir de un grupo de oradores. En el segundo, se requiere determinar si la identidad pronunciada por el orador es correcta. Ninguno de los dos casos requiere la interpretación o reconocimiento del mensaje pronunciado por el orador.

D. Principales métodos de reconocimiento de voz

Los métodos de reconocimiento de voz procesan un grupo de características propias de cada instrucción verbal para su identificación. Estas características son determinadas según la técnica de análisis empleada. De esta manera, se podría resumir el proceso de reconocimiento de voz en dos partes: análisis y reconocimiento.

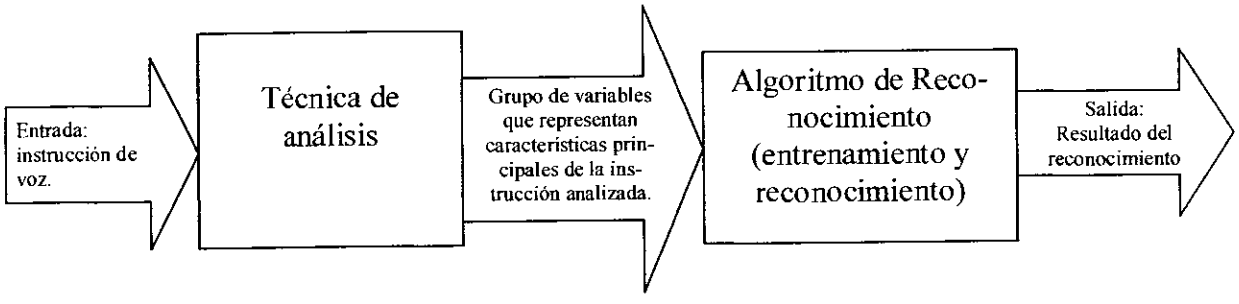


Figura No. 9 Diagrama esquemático del proceso de reconocimiento de voz (se asume la aplicación de técnica adecuada para detección de inicio y final de palabra, en el caso de algoritmos de habla continua).

1. Técnicas de análisis

Al hablar de técnicas de análisis se hace referencia al proceso de caracterización del problema, a la manera en que se extraen las principales características de una instrucción verbal, las que luego serán procesadas por uno de los diferentes métodos de reconocimiento de voz. Las principales técnicas de análisis de instrucciones verbales son:

a. Análisis espectral de tiempo corto

Involucra la obtención de "marcos de voz", fragmentos espectrales que corresponden a intervalos cortos de tiempo de la muestra analizada. Estos marcos caracterizan en forma dinámica la forma de onda de la instrucción para su posterior reconocimiento.

b. Análisis de predicción lineal

Parte del principio de modelar una señal de voz humana discreta en el tiempo, como una sucesión de filtros digitales. El análisis implica la identificación de los parámetros asociados con la función de transferencia "todo polo" (que resume el efecto de todos los filtros digitales), así sirve como una aproximación del funcionamiento del sistema de producción del habla.

c. Análisis "Cespectral*" (*Cepstral*)

Centrado alrededor del análisis de fonemas vocalizados, permite la linealización de una señal de voz; la cual es modelada como la convolución entre una secuencia de excitación y la respuesta al impulso unitario del "filtro" del tracto vocal. Sin linealizar, no es posible determinar el efecto que un filtro lineal tendría sobre la señal de voz (que es el resultado de una convolución), es decir, los filtros lineales no son aplicables. Sin embargo, después del proceso de linealización, la aplicación de estos filtros es válida, ya que los resultados del mismo son predecibles. Así se pueden eliminar compo-

* La palabra Cespectral es una traducción del inglés, *cepstral*. Forma parte de la jerga utilizada en la discusión de esta técnica de análisis y se deriva del hecho de que se está realizando una transformación de una señal que ya está en el "dominio de la frecuencia" hacia el "dominio de la frecuencia". El resultado es una transformación hacia el "dominio de la quefrecuencia (*quefrecuency*)", cuyo espectro es denominado "cespectro".

nentes frecuenciales que no se desean tomar en cuenta en el análisis. Lo anterior justifica que se utilice esta técnica para fonemas vocalizados, pues al eliminar componentes de ruido de alta frecuencia, se atenúa el efecto de ciertas consonantes como por ejemplo, las fricativas, que pertenecen a los fonemas no vocalizados.

2. Métodos de Reconocimiento

a. Métodos de ajuste dinámico en tiempo (*Dynamic Time Warping*)

Ideado originalmente para reconocimiento de palabras aisladas, este método se fundamenta principalmente en una forma de comparación de patrones (*templates*), formados por características importantes de la instrucción. Utiliza técnicas estadísticas generales para el reconocimiento de patrones en un grupo de datos. El método alinea en tiempo las muestras que se van a comparar, compensando. Así compensa variaciones en la duración de la pronunciación de la palabra y de fonemas de duración variable que forman parte de la palabra y no ocurren como una pronunciación aislada.

b. El Modelo Markov (*The Hidden Markov Model*)

El modelo Markov puede ser considerado como una "máquina" abstracta, utilizada para modelar una pronunciación (Deller, 2000). Esta "máquina" es capaz de generar grupos de observaciones (las que se obtuvieron luego de aplicar una de las técnicas de análisis de datos descrita con anterioridad). Una "unidad HMM" produce con mayor probabilidad grupos de observaciones que se notarían al pronunciar la palabra asociada a dicha unidad (y para la cual la unidad ha sido entrenada). El entrenamiento de las unidades HMM incorpora en las mismas el "esquema estadístico" propio de la palabra que se le asocia. En la fase de reconocimiento, se asume que el grupo de observaciones perteneciente a la nueva instrucción de entrada fue producido por una unidad HMM desconocida. Luego se analizan las probabilidades que este grupo de observaciones se haya producido por una de las unidades HMM previamente entrenadas. La unidad con la mayor probabilidad identificará la instrucción ingresada* .

* Además del Modelo Markov, los algoritmos de redes neurales se consideran también como técnicas de procesos estocásticos que, aunque parecidos a la lógica del modelo Markov, involucran conceptos distintos.

CAPITULO III
MARCO METODOLÓGICO

A. Hipótesis

Con base en los principios expuestos en la justificación de este trabajo se propone la siguiente hipótesis:

La información brindada por la caracterización de una instrucción verbal en términos de las amplitudes de su espectro frecuencial permite lograr un reconocimiento de dichas instrucciones, mediante la aplicación de un análisis de correlación.

B. Población y muestra

Tal y como se especificó en la sección de alcances y límites de este trabajo, para efectos del cumplimiento de los objetivos planteados, se analizaron los espectros frecuenciales de 10 palabras cortas[1] del idioma inglés. La motivación para escoger este idioma fue el hecho de que en él, la mayoría de imperativos son de corta duración, contrario a lo que ocurre en el español. Este hecho resulta importante dada la limitación del número de muestras que la unidad de apoyo en hardware (encargada del cómputo de la transformada rápida de Fourier) es capaz de manejar.

Tabla No. 1
Instrucciones verbales seleccionadas

PALABRA	PRONUNCIACION[2]
"on"	'on
"off"	'of
"time"	'tIm
"up"	'&p
"down"	'daun
"left"	'left
"right"	'rIt
"start"	'stärt
"stop"	'stäp
"get"	'get

[1] En este contexto, al hablar de palabras cortas se hace referencia a palabras con una duración media de aproximadamente 0.74 segundos.
[2] Según Guía de Pronunciación del Diccionario *Merriam Webster*, presentada en la sección de anexos. (http://www.m-w.com)

C. Detalle del análisis:

Se pretende el estudio de espectros frecuenciales para determinar su aptitud para el desarrollo de técnicas de reconocimiento. Para obtener los espectros frecuenciales (amplitud vs. tiempo), se requiere de la aplicación del operador lineal definido como la transformada discreta de Fourier (dado que una muestra de una instrucción constituye una sucesión discreta de diferentes valores de amplitud). Existe un algoritmo mucho más práctico para el cálculo de la transformada discreta de Fourier, conocido como la transformada rápida de Fourier o FFT por sus siglas en inglés. La aplicación de dicha herramienta matemática la realiza el apoyo en Hardware desarrollado como proyecto paralelo a éste por el Sr. Pedro M. Viscovich. La idea de este enfoque es lograr la mayor rapidez en la aplicación de la herramienta matemática, lo que permite que la computadora personal se dedique únicamente al muestreo de las instrucciones, al despliegue de los resultados y, eventualmente, a la implementación de los procedimientos de reconocimiento en régimen frecuencial. A continuación se presenta un diagrama de bloques que ilustra cómo se une la herramienta implementada en hardware con el programa en software desarrollado para el muestreo y captura de la instrucción verbal.

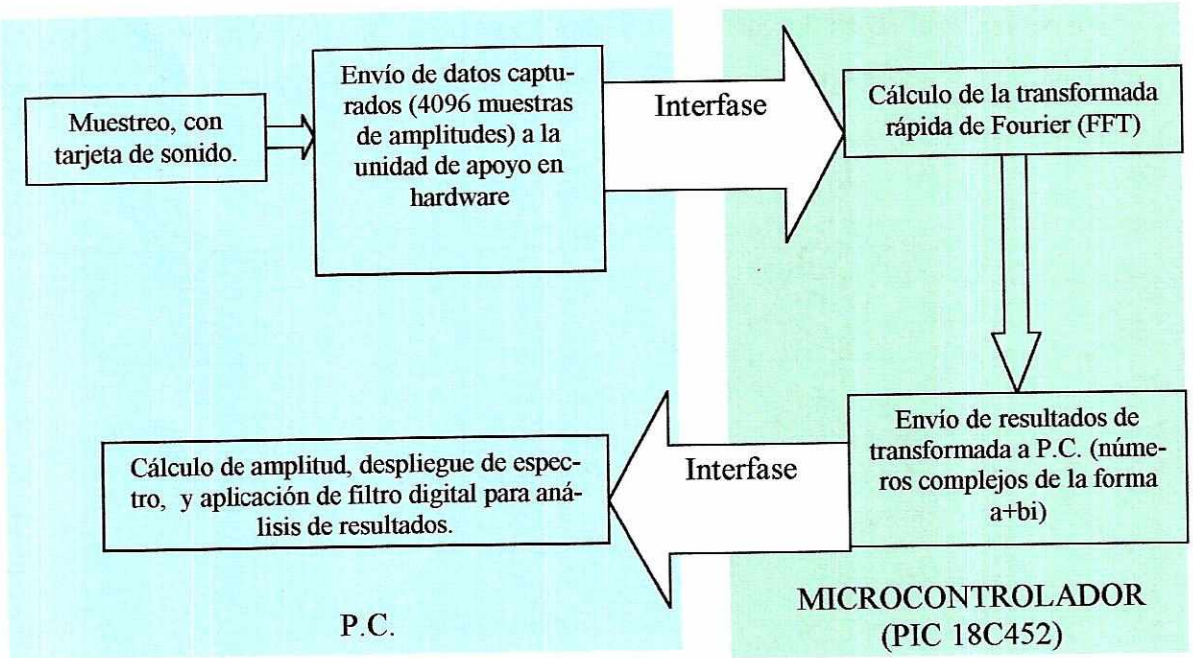


Figura No. 10 Diagrama de funcionamiento, muestreo con *buffer* circular e interfase con apoyo en hardware para implementación de FFT.

Para la solución del problema planteado con anterioridad, se dividió el mismo en 2 fases secuenciales:

1. Muestreo.
2. Análisis de datos.

1. Muestreo

Esta fase inicial tiene como objetivo el muestreo satisfactorio de una señal de audio, con la utilización de la tarjeta de sonido de una computadora personal y el envío a la unidad de apoyo en hardware de la información digital que la representa por medio de una interfase. La elección de la misma estuvo basada en consideraciones del proyecto de apoyo en *hardware*, desarrollado en forma paralela a éste. Dada la limitación técnica de hacer una interfase más rápida, adecuada para aplicaciones de reconocimiento de voz, se optó por una interfase serial descrita por el protocolo RS-232. Esto se justifica por la fase experimental del proyecto, lo que da lugar a la aparición de limitaciones de implementación durante su desarrollo. Cada uno de estos objetivos se logró con el diseño y realización de una aplicación desarrollada en lenguaje de alto nivel (Borland Delphi 6.0).

El funcionamiento de la aplicación para la captura de instrucciones verbales se muestra en el diagrama de flujo de la figura No.11.

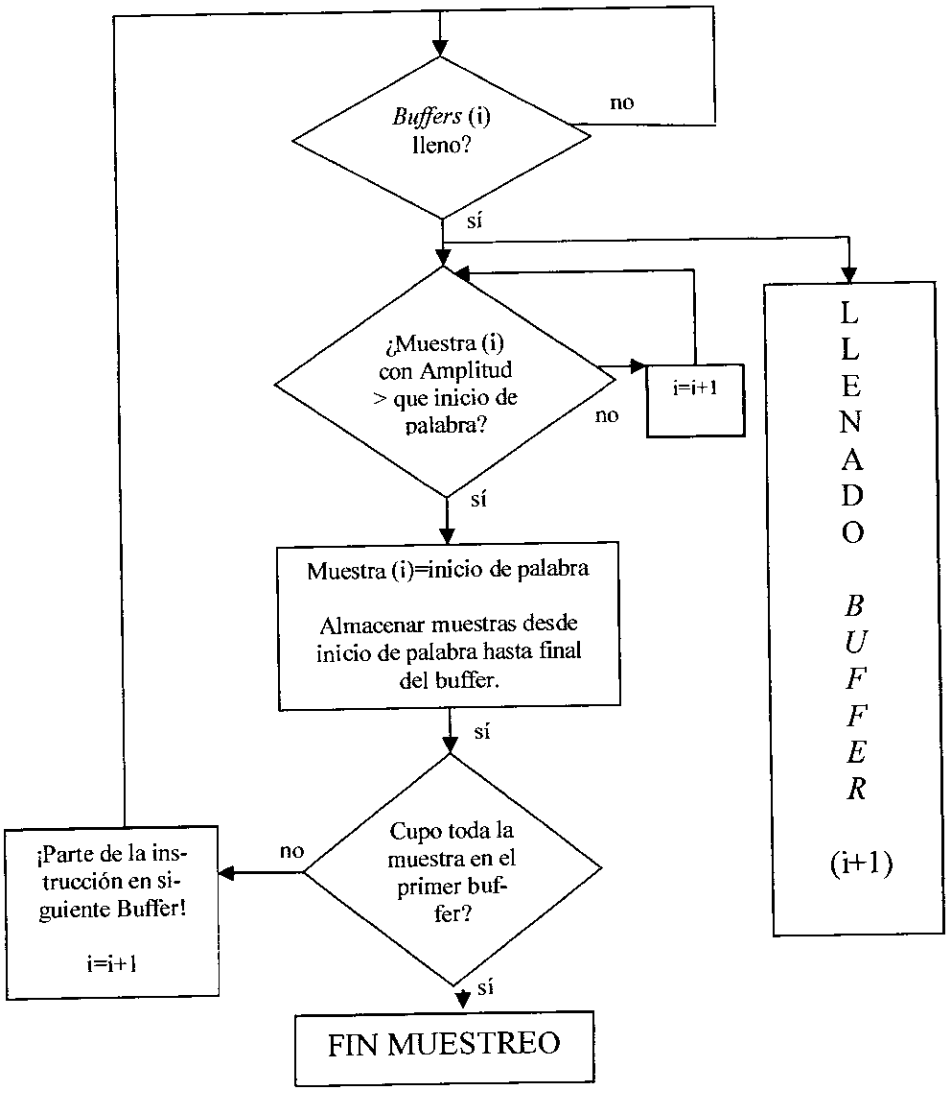


Figura No. 11 Diagrama de flujo, muestreo de instrucciones verbales utilizando un *buffer* circular.

Inicialmente la herramienta captura una instrucción verbal por medio de un micrófono conectado en forma directa a la tarjeta de sonido del ordenador. Como puede inferirse del tipo de aplicaciones para las cuales se pretende aplicar este tipo de reconocimiento, no es considerado viable tener que obligar al usuario a oprimir una tecla para iniciar la

captura de una instrucción verbal. Por el contrario, se requiere que la aplicación esté recibiendo constantemente sonidos del exterior y que determine el inicio de una palabra con un límite de amplitud debajo del cual, toda señal es considerada como "ruido ambiental". Este principio es la base del funcionamiento de la aplicación. Para lograr implementar este funcionamiento, se diseñó un *buffer* circular, formado a su vez por dos *buffers* de igual tamaño. Una vez lleno uno de éstos, los datos almacenados son analizados para determinar si hay o no un inicio de palabra (si una de las amplitudes capturadas sobrepasa el límite de ruido preestablecido). Sin embargo, durante este proceso de análisis no se interrumpe la captura de información, pues se procede (en forma paralela) a llenar el segundo buffer. Toda la rutina de almacenamiento se ejecuta con la suficiente rapidez para evitar llegar a tener los dos *buffers* llenos en forma simultánea (lo que podría acarrear no disponer de espacio para continuar el muestreo).

El muestreo no comienza en el punto donde se sobrepasa el límite de amplitud preestablecido. Al hacer esto se incurriría en pérdida de datos que pertenecen a la palabra, pero que no son lo suficientemente intensos como para sobrepasar el límite de ruido preestablecido. Por ello, el inicio de la instrucción se toma 50 muestras antes del punto donde se excede el límite de ruido preestablecido. Esta regla es el resultado de un análisis preliminar de las instrucciones muestreadas. Es claro que las 8192 muestras de la instrucción pueden estar en un buffer o repartidas entre ambos. Por esta razón la muestra estará compuesta, con una alta probabilidad, por elementos de ambos *buffers*.

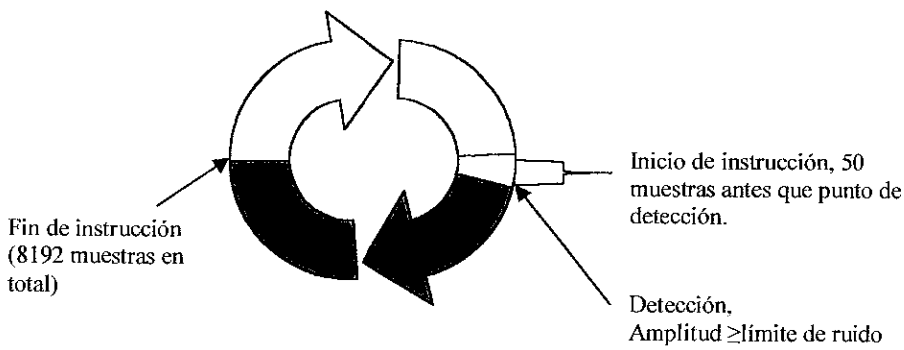


Figura No. 12 Distribución de una instrucción verbal entre ambos *buffers*.

Los parámetros de muestreo son transferidos a la tarjeta de sonido de la computadora, desde la misma rutina de software. El procedimiento de muestreo en este dispositivo está fundamentado en herramientas y procedimientos de software multi-tareas. Cuando un sonido alimenta a la tarjeta de sonido por medio de un micrófono, con el afán de capturar dicho sonido y almacenarlo en un dispositivo de memoria (típicamente una unidad de disco duro), interactúan 3 de sus principales constituyentes:

1. Un convertidor análogo digital (ADC).
2. Un procesador de señales digitales (DSP).
3. Un bus de datos, que permite la comunicación con el procesador central del ordenador.

La aplicación diseñada permite el muestreo de una señal presente en el micrófono, por medio de un *buffer* circular. Independiente del tipo de *buffer* que se emplee, así como de la manera en que efectivamente se captura la instrucción verbal en la computadora, la adquisición de un sonido por medio de una tarjeta de sonido obedece el siguiente proceso fundamental:

a. Inicialización de la tarjeta de sonido

Las tarjetas de sonido están a menudo equipadas con dos canales de entrada. Debe seleccionarse por software el canal que se utilizará. En esta etapa también se especifican los parámetros para el muestreo de la señal. Los valores que pueden asignarse a cada parámetro dependen de cada fabricante de tarjetas, aunque todas obedecen al mismo estándar fundamental.

Tabla No. 2

Parámetros de configuración

PARÁMETRO	VALORES PERMISIBLES
Número de canales de entrada	1 ó 2
Número de muestras por segundo (Frecuencia de muestreo)	8.0 kHz, 11.025 kHz, 22.05 kHz ó 44.1 kHz
Número de *bits* por muestra	8 ó 16

b. Captura de la instrucción verbal

Luego de la configuración de la tarjeta de sonido, se configura la misma para la recepción de señales analógicas. La señal alimentada, que varía no sólo en amplitud sino también en frecuencia, es procesada por la tarjeta de sonido de la siguiente manera:

1. La señal alimenta al convertidor análogo digital, que lo procesa en tiempo real y genera una secuencia de *bits*.
2. La señal digital resultante alimenta al procesador de señales digitales (DSP), el que ha sido programado con un grupo de instrucciones que le permiten manejar, en este caso, todo el proceso de captura, de tal forma que el microprocesador central no interviene en el proceso. Algunas tarjetas de sonido incorporan rutinas de compresión de datos en el DSP.

Los datos recibidos y digitalizados son almacenados en un *buffer* de memoria, cuyo tamaño se especifica al configurar la tarjeta de sonido. Una vez lleno el *buffer*, se realiza una lectura secuencial de los valores allí almacenados

El procesador de señales digitales transfiere luego la información al microprocesador por medio del bus de datos, de donde la información es enviada a la unidad donde se desea almacenar.

Fue necesario realizar un ajuste a las muestras capturadas, debido a limitaciones de hardware en el proyecto desarrollado paralelamente a éste. Al muestrear una analógica continua en el tiempo, se define una frecuencia crítica de muestreo:

$$Fs=2Fc$$

Donde Fc representa el ancho de banda *B* (límite en frecuencia) de la señal, de tal forma que todos sus componentes frecuenciales por arriba de esta frecuencia son cero. La frecuencia crítica Fs se conoce como la frecuencia de Nyquist, y a las consecuencias de esta definición se les denomina teorema de Nyquist. Su definición caracteriza la mínima frecuencia de muestreo que permite la recuperación completa de una señal analógica previamente digitalizada. Una frecuencia de muestreo menor que esta, ocasionará un fenómeno denominado *Aliasing*, producto de la ambigüedad asociada con el dominio en frecuencia de señales discretas en el tiempo. "Cuando se muestrea una señal en continua en el tiempo a una frecuencia de Fs muestras por segundo, y sea *k* un número entero positivo o negativo, no es posible distinguir entre los valores muestreados de una onda sinusoidal con frecuencia Fo y una onda sinusoidal de Fo+kFs" (Lyons, 1997). Este resultado puede visualizarse analizando la siguiente figura, nótese que los puntos muestreados corresponden a todas las funciones sinusoidales con frecuencia Fo+kFs.

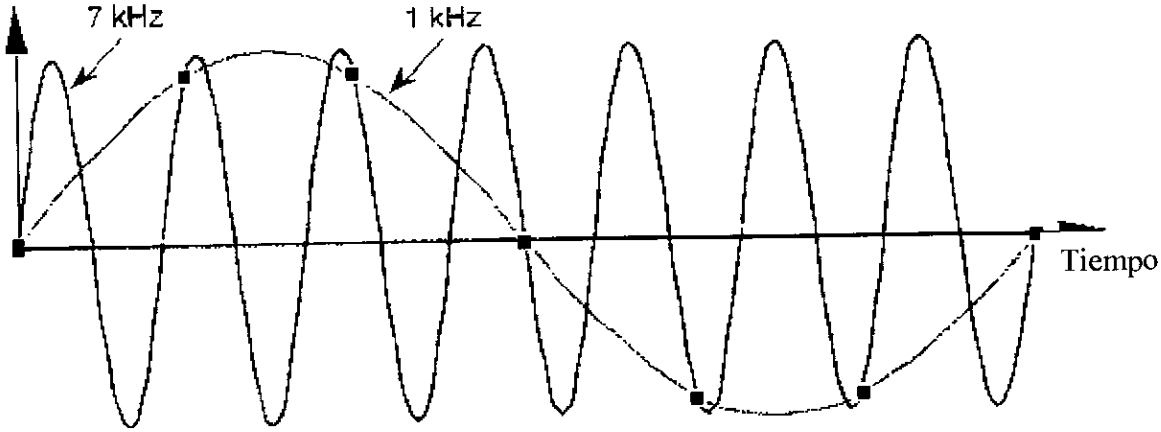


Figura No. 13 Ejemplo de ambigüedad en el muestreo de dos ondas sinusoidales, una con frecuencia Fo=1 kHz. y otra con frecuencia Fo+kFs, donde Fs=7Khz, y k=1. Reproducido de: Lyons, R. *Understanding Digital Signal Processing*. Addison Wesley. E.E.U.U., 1997. Pág. 27.

En el espectro frecuencial, este fenómeno se manifiesta tal y como lo presenta la figura 15 C. Nótese que el fenómeno de Aliasing se da únicamente al trabajar con señales discretas en el tiempo y no con señales continuas.

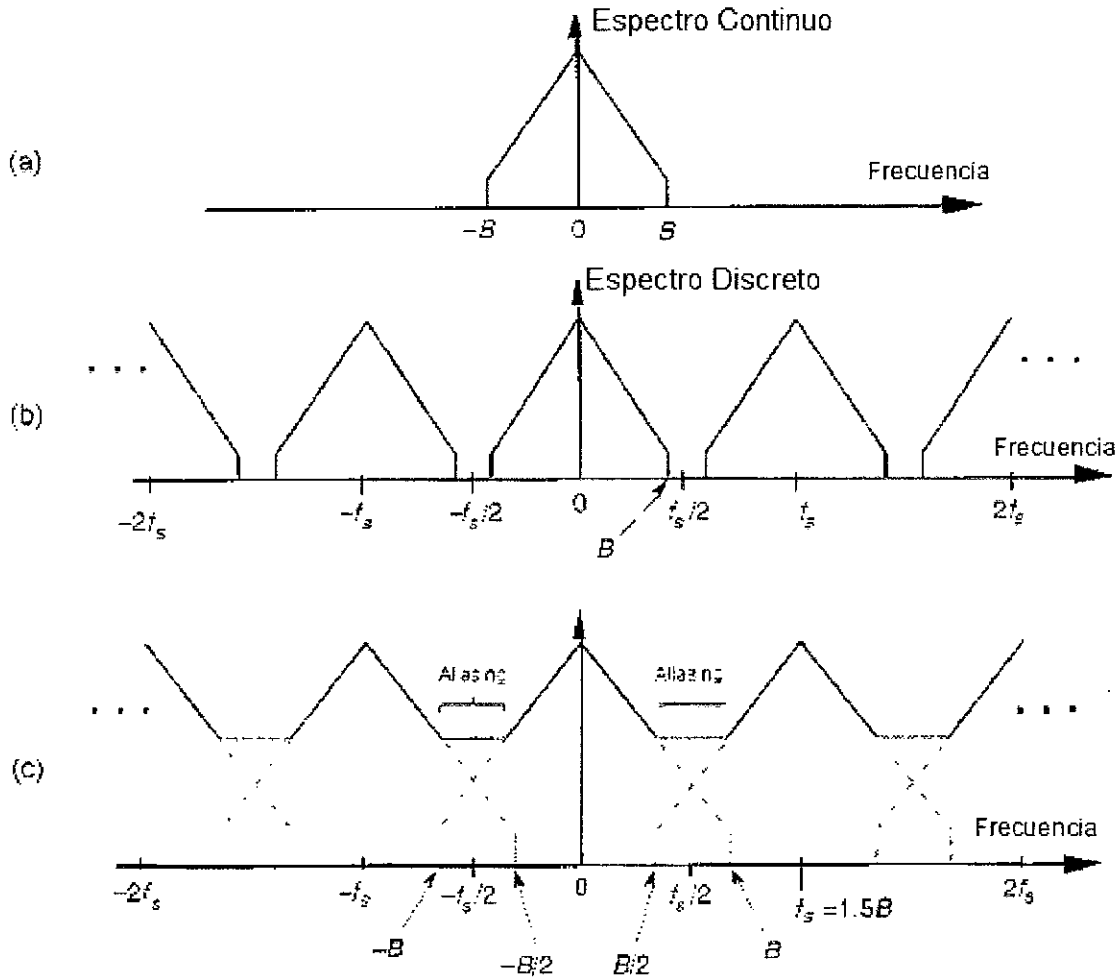


Figura No. 14 Aliasing en régimen frecuencial. a. Espectro Continuo. b. Espectro discreto, sin aliasing. c. Espectro discreto, con aliasing, causado por una frecuencia de muestreo menor que el ancho de banda B de la señal. Reproducido de: Lyons, R. *Understanding Digital Signal Processing*. Addison Wesley. E.E.U.U., 1997. Pág. 30.

Se considera que todas las señales audibles de voz tienen un ancho de banda de 5kHz. (Fallside, 1985). Esto significa, según el criterio de Nyquist, que la mínima frecuencia de muestreo para poder recuperar la señal completamente será de 10 kHz. La frecuencia de muestreo utilizada en la tarjeta de sonido fue de 11 kHz, con un *buffer* circular

formado por dos secciones de 8192 *bytes* cada una, así se garantiza que no se pierde la información adquirida por la tarjeta de sonido mientras se procesa el primer grupo de datos. La muestra total en este punto está constituida por 8192 *bytes* de longitud. El apoyo en hardware con que se realiza la transformada de Fourier para obtener los espectros de la instrucción verbal tiene como límite de procesamiento 4096 muestras. Por ello, se realiza un ajuste de la muestra antes de enviarla al microcontrolador para la ejecución del algoritmo de la transformada de Fourier. El ajuste realizado consiste en tomar únicamente la mitad de las muestras (un total de 4096), lo que equivale a realizar el muestreo a 5.5kHz. Este valor de frecuencia resultó ser suficiente para evitar *aliasing*, como puede observarse en las gráficas espectrales de cada instrucción, presentadas en el siguiente capítulo.

Luego de este ajuste, es necesario sustraer el valor numérico 128 de cada una de las amplitudes que forma la muestra. Lo anterior se hace necesario, ya que las amplitudes relativas (que pueden ser positivas o negativas) son representadas con un *byte* de información, de tal forma que una amplitud de cero equivale a 128, entonces es posible la representación tanto de valores positivos y negativos, sin necesidad de tener un *bit* de información para especificar el signo de la amplitud.

Finalmente, la aplicación envía los datos por medio de una interfase serial al procedimiento de apoyo en hardware, el que retorna 4096 números complejos, donde las posiciones pares del arreglo recibido coincidirán con las partes reales de cada número y las posiciones impares con las partes imaginarias.

2. Aplicación de la transformada rápida de Fourier (FFT)

La aplicación de la FFT se realizó en el procedimiento de apoyo en *hardware*, desarrollado en forma paralela a este proyecto. El punto central de este trabajo no es la aplicación de la transformada, sino el análisis de los resultados obtenidos a partir de la misma.

3. Análisis de Datos

La parte de análisis de datos fue implementada en una hoja electrónica dinámica, que permite realizar tanto la aplicación del filtro digital como el análisis estadístico sobre las muestras estudiadas.

El primer paso consiste en tomar los valores devueltos por el apoyo en hardware y calcular la norma de cada número complejo (el análisis se limita al espectro amplitud vs. frecuencia).

Los espectros frecuenciales obtenidos poseen una naturaleza ruidosa que dificulta su análisis y comparación. Para resolver este problema se recurre a las técnicas de filtrado que permiten la eliminación de la mayoría de ruido de alta frecuencia. La aplicación de un filtro digital en este punto, sobre muestras en el régimen frecuencial, escapa un poco del enfoque tradicional bajo el cual se aplican estos filtros. Debe recordarse que cuando se pretende eliminar el ruido de una señal, el filtro (típicamente un pasa bajos) es aplicado antes de realizar la transformada de Fourier, es decir, en régimen de tiempo. De esta manera, al aplicar la transformada, las componentes de frecuencias altas se ven atenuadas en el espectro frecuencial; el efecto en régimen del tiempo es el "suavizar" la curva de amplitud vs. tiempo. Es esta última característica de los filtros digitales la que nos interesa explotar en este caso; es decir, se pretende suavizar el espectro de frecuencias obtenido, lo que tendrá como resultado una caracterización más clara que facilita el análisis discriminante entre instrucciones. El filtro aplicado a las muestras fue un filtro digital IIR de primer orden, en cascada, con uno de segundo orden. A continuación se especifican las funciones de transferencia de cada uno de los filtros utilizados, las cuales parten de la función de transferencia para el filtro analógico correspondiente.

IIR, primer orden:

$$A(p) = \frac{A_0}{1 + a_1 p} \Rightarrow A(z) = \alpha_0 \frac{1 + Z^{-1}}{1 + \beta_1 Z^{-1}};\ \text{con:}$$

$$\alpha_0 = \alpha_1 = \frac{A_0}{1 + a_1 l} \quad \text{y} \quad \beta_1 = \frac{1 - a_1 l}{1 + a_1 l}$$

IIR, segundo orden:

$$A(p) = \frac{A_0}{1 + a_1 p + b_1 p^2} \Rightarrow A(z) = \alpha_0 \frac{1 + 2Z^{-1} + Z^{-2}}{1 + \beta_1 Z^{-1} + \beta_2 Z^{-2}};\ \text{con:}$$

$$\alpha_0 = \frac{A_0}{1 + a_1 l + b_1 l^2};\quad \beta_1 = \frac{2(1 - b_1 l^2)}{1 + a_1 l + b_1 l^2} \quad \text{y} \quad \beta_1 = \frac{1 - a_1 l + b_1 l^2}{1 + a_1 l + b_1 l^2}$$

La importancia de aplicar estos dos filtros en cascada radica en que se obtiene una caída muy pronunciada de amplitud (60 decibeles por década), a partir de la frecuencia de corte. Al recordar la simetría de los resultados proporcionados por la transformada de Fourier, basta tomar únicamente la mitad de las muestras para el análisis, es decir, 2048. Para clarificar el efecto del filtro digital en el espectro de frecuencias, considérese la gráfica No. 3 de la siguiente sección, donde claramente se observa cómo el filtro "suaviza" el espectro obtenido.

En este punto se deben poner en perspectiva los dos problemas involucrados en el reconocimiento de voz. De esta manera puede tenerse una mejor idea del objetivo que se pretende alcanzar una vez se analicen los resultados obtenidos. Se opta por dividir el problema de reconocer instrucciones verbales cortas en dos etapas sucesivas y complementarias:

1. El entrenamiento o aprendizaje de las instrucciones que se desea reconocer.
2. El reconocimiento de una nueva instrucción ingresado al asumir que se ha dado el proceso de aprendizaje correspondiente.

Estas dos etapas dividen el análisis estadístico a realizar de aquí en adelante. Para la primera etapa, debe determinarse un valor en base a las muestras de una misma instrucción con que se entrenó el sistema. Luego, debe escogerse una medida que permita determinar el grado de relación que manifiesten muestras de una misma instrucción con las unidades típicas de cada uno; misma medida que será empleada para discriminar entre instrucciones durante la fase de reconocimiento.

Para realizar el análisis, resulta conveniente considerar la información proporcionada por cada muestra como un vector en el plano R^{2048}; donde cada componente equivale al valor en magnitud de cada una de las frecuencias del espectro. Por razones de claridad, dichos vectores serán representados como líneas rectas en el plano R^2. Esta equivalencia facilita la visualización de los conceptos que se describirán.

a. Fase 1. Entrenamiento

En esta fase, el usuario deberá proporcionar al sistema un grupo de muestras para cada instrucción que desee reconocer. En este proyecto se utilizaron 5 muestras para entrenar al sistema. Al considerar cada muestra como un vector en R^{2048}, representados en R^2 por razones de claridad, se tendrá un conjunto (en este caso de cinco elementos) de vectores relativamente cercanos entre sí. La primera parte para resolver el problema de reconocimiento consiste en encontrar un vector típico que represente a cada grupo de vectores (a cada instrucción). La manera más sencilla es obtener un vector promedio de todas las muestras con las que se entrenó el sistema. Este vector se obtiene al calcular la media aritmética por componente; es decir:

$$\vec{A} = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ \cdots \\ a_n \end{bmatrix} \quad ; \quad \vec{B} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ \cdots \\ b_n \end{bmatrix} \Rightarrow \bar{X} = \begin{bmatrix} (a_1 + b_1)/2 \\ (a_2 + b_2)/2 \\ \cdots \\ \cdots \\ (a_n + b_n)/2 \end{bmatrix}$$

Donde A y B representan dos muestras de una misma instrucción y X al vector promedio correspondiente. De esta forma, cada instrucción puede ser representada por un solo vector. La siguiente figura muestra una perspectiva más clara de esta idea.

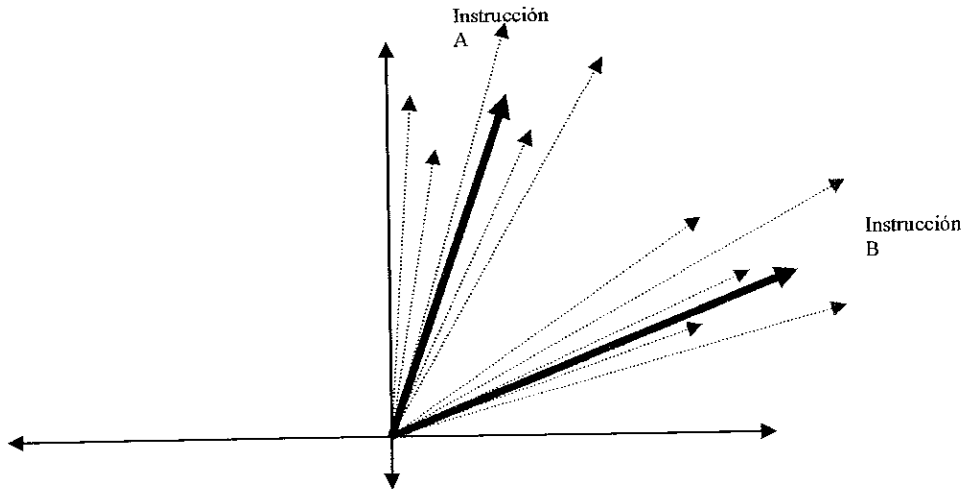


Figura No. 15 Representación de muestras de dos instrucciones verbales distintas, como vectores en el espacio. Los vectores resaltados constituyen el vector típico de cada grupo de muestras.

El proceso de entrenamiento constituye entonces la determinación de los vectores promedio de cada grupo de muestras. Es importante resaltar que los vectores mostrados en la figura anterior no poseen la misma magnitud, esto se debe a variaciones en la intensidad de pronunciación. La dispersión de las muestras se debe a que las componentes frecuenciales no coinciden exactamente, dado el funcionamiento del aparato productor del habla, descrito en el marco teórico.

b. Fase 2. Reconocimiento

En esta fase se pretende determinar la viabilidad de los procedimientos de reconocimiento de voz con base en espectros frecuenciales. Para los efectos, se toman como base los vectores típicos calculados en la fase anterior y una muestra adicional de cada instrucción, la cual representa la "nueva instrucción" recién ingresado al sistema para ser reconocido. La base de esta fase es la determinación de una regla matemática que permita la discriminación entre instrucciones, es decir, que permita asociar correctamente la nueva instrucción con uno de los ingresados en la fase de entrenamiento. Para determi-

nar esta regla matemática de discriminación, se optó por un análisis de correlación entre la nueva muestra y los vectores promedio de la fase de entrenamiento.

El análisis de correlaciones, como se utilizó en este proyecto, está enmarcado en las técnicas de análisis multivariado de datos. Este tipo de análisis engloba los procesos estadísticos que se ocupan de las relaciones entre grupos de variables dependientes y los individuos que las manifiestan (Kendall, 1980). Entre los objetivos de este tipo de análisis está el simplificar la complejidad del problema, ya que se analiza un grupo de n objetos, cada uno con observaciones de p variables aleatorias (en este caso, se analizan n muestras de instrucciones diferentes, cada una con p elementos del espectro). Existe interés en las técnicas de análisis multivariado de datos que permitan determinar si un objeto cae o no dentro de un grupo de observaciones. Existe una cantidad considerable de técnicas de clasificación, sin embargo, al inicio de la concepción del proyecto se consideró fundamentar el reconocimiento en conceptos geométricos a partir del espectro de frecuencias. El análisis de correlaciones es en realidad una relación geométrica entre vectores en un plano; además, resulta ser independiente de la magnitud de los vectores. Esto es importante pues de todas las variables involucradas en un espectro de una instrucción verbal, la de mayor variabilidad es la amplitud con que se emite la instrucción (esto será evidente al analizar las gráficas presentadas en la sección de resultados) y es una de las principales razones de distorsión de las instrucciones verbales. Por ello resulta conveniente adoptar una medida de discriminación que sea totalmente independiente de esta cantidad. Antes de describir detalladamente el algoritmo de correlación, se incluyen las definiciones de algunos conceptos importantes, siempre dentro del contexto de análisis multivariado de datos (Kendall, 1980):

1) Media

Si se tiene un vector n dimensional, la media de sus valores se define como:

$$\overline{x_i} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$$

2) Matriz de covarianza

A menudo resulta conveniente la representación de un grupo de datos en términos de la distancia de cada uno con la media respectiva. A partir de un grupo de p vectores n dimensionales:

$$X_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \dots \\ \dots \\ x_{n1} \end{bmatrix}; \; X_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ \dots \\ \dots \\ x_{n2} \end{bmatrix}; \; \dots\dots; \; X_p = \begin{bmatrix} x_{1p} \\ x_{2p} \\ \dots \\ \dots \\ x_{np} \end{bmatrix}$$

Se define la matriz de covarianza o matriz de dispersión como:

$$C = \begin{bmatrix} 1 & c_{12} & .. & c_{1p} \\ c_{12} & 1 & .. & c_{2p} \\ .. & .. & .. & .. \\ c_{1p} & c_{2p} & .. & 1 \end{bmatrix}; \text{ donde } c_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})$$

A c_{jk} se le llama covarianza de X_j y X_k.

3) Matriz de correlación

Se define el análisis de correlación como una medida del grado de cercanía entre dos variables (Yamame, 1964). Este es el enfoque univariado tradicional. Según Kendall (1980), esta definición puede expanderse para el caso multivariado al definir la matriz de correlación de un grupo de p vectores n dimensionales en términos de su respectiva matriz de covarianza.

Sean nuevamente:

$$X_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \dots \\ \dots \\ x_{n1} \end{bmatrix}; \; X_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ \dots \\ \dots \\ x_{n2} \end{bmatrix}; \; \dots\dots; \; X_p = \begin{bmatrix} x_{1p} \\ x_{2p} \\ \dots \\ \dots \\ x_{np} \end{bmatrix}$$

Si cada componente x_{ij} de un vector j es primero dividida entre la raíz cuadrada de la varianza del vector (definida como: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_j})^2$), la matriz de covarianza se convierte en la matriz de correlación:

$$r = \begin{bmatrix} 1 & r_{12} & .. & r_{1p} \\ r_{12} & 1 & .. & r_{2p} \\ .. & .. & .. & .. \\ r_{1p} & r_{2p} & .. & 1 \end{bmatrix}$$

Como punto final, se presenta una observación sobre la forma como se interpreta un típico grupo de datos multivariados (Kendall, 1980). Esta observación resulta interesante pues en ella se fundamenta el enfoque geométrico que en este caso se le da al análisis de correlaciones. Un típico grupo de datos multivariados se arregla en una matriz p x n:

$$\begin{matrix} x_{11} & x_{12} & .. & x_{1p} \\ x_{21} & x_{22} & .. & x_{2p} \\ .. & .. & .. & .. \\ x_{n1} & x_{n2} & .. & x_{np} \end{matrix}$$

Comúnmente, se expresan los datos con p incrementado en forma horizontal y n en forma vertical (como se representaron los vectores en las dos definiciones anteriores). El grupo anterior de datos puede considerarse desde dos puntos de vista, de acuerdo con la forma en que se lean los datos de la matriz. Si por ejemplo se comparan dos de las columnas de la matriz, se estará analizando la relación entre dos observaciones, cada una con n variables aleatorias. Si por el contrario, se comparan dos filas, se estará estudiando la relación entre un par de variables aleatorias del grupo.

Esta dualidad de interpretación de la matriz de datos da lugar a dos métodos completamente diferentes de representación geométrica de los mismos. El primero es una generalización natural del diagrama de dispersión. Puede pensarse en p ejes ortogonales que definen un espacio p-dimensional, donde cada miembro del grupo corresponde a un

punto en dicho espacio. Las observaciones constituyen entonces una nube de puntos en p dimensiones. El punto de interés en este caso yace en el patrón que exhibe esa nube de puntos.

La segunda representación geométrica resulta menos familiar, pero es la que se adoptará en este análisis. Consiste en considerar los datos no como n puntos en p dimensiones, sino como p puntos en n dimensiones. De esta manera, se toman n ejes ortogonales que corresponden a cada miembro del grupo de datos (las n variables aleatorias de cada observación). Se obtiene entonces un vector por cada grupo de n variables aleatorias. De esta manera, se tendrá un conjunto de vectores n dimensionales, lo que coincide con la interpretación geométrica que se mencionó con anterioridad. Si se calcula la distancia desde el origen hasta el punto (la norma del vector), tenemos

$$OP_j^2 = \sum_{i=1}^{N} (x_{ij} - \overline{x_j})^2$$

Es decir, n veces la varianza de X_j.

Si luego se estandarizan los vectores de tal forma que sean unitarios (lo que se logra al dividirlos por sus respectivas varianzas), sus extremidades yacerán en una hiperesfera de radio unitario. Además, el coseno del ángulo entre ambos vectores coincide con la definición de correlación dada con anterioridad. Esto se hace evidente al plantear la ecuación para el coseno del ángulo entre dos vectores X_j y X_k, a partir de la definición del producto escalar entre ambos; así

$$\cos(\theta_{jk}) = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})}{\left[\sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2 \sum_{k=1}^{n} (x_{ik} - \overline{x_k})^2 \right]^{1/2}}$$

Entonces, el grado de cercanía entre dos de las muestras será medido en términos del ángulo entre ambos vectores. Es importante recalcar que aunque en el desarrollo anterior se normalizaron los vectores, el valor del coseno del ángulo entre ambos es totalmente independiente de la magnitud de los vectores; lo que resulta favorable en nuestra situación, pues no se desea fundamentar la herramienta de discriminación en la amplitud (que

varía de forma significativa). Es claro que una correlación cercana a uno indica que los vectores son muy próximos entre sí, mientras que una correlación muy cercana a cero indica una significativa separación entre los mismos.

Una vez obtenidos los parámetros de la matriz de correlación, se calcularán los ángulos correspondientes. Esto se hace con el fin de tener una mejor idea de la separación entre los datos. La función coseno es no lineal y por su naturaleza, dos valores muy próximos son a menudo no concluyentes en cuanto a la proximidad de los vectores. Por ello, se opta por la comparación entre ángulos, pues éstos constituyen una escala lineal de valores.

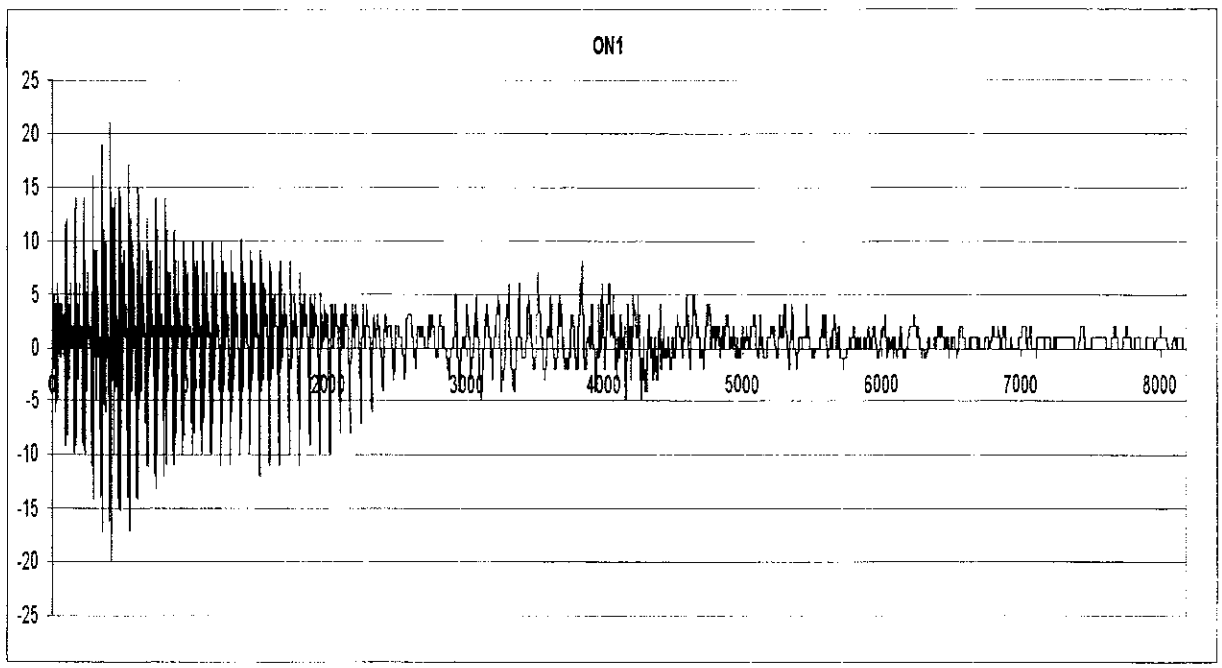
El proceso de discriminación obedece luego a un simple proceso de comparación entre el ángulo de la nueva muestra y cada uno de los vectores que caracterizan a las instrucciones (los vectores típicos determinados en la fase de entrenamiento). La nueva muestra deberá entonces corresponder a la instrucción cuyo vector representativo esté más próximo.

CAPITULO IV

PRESENTACIÓN E INTERPRETACIÓN DE RESULTADOS

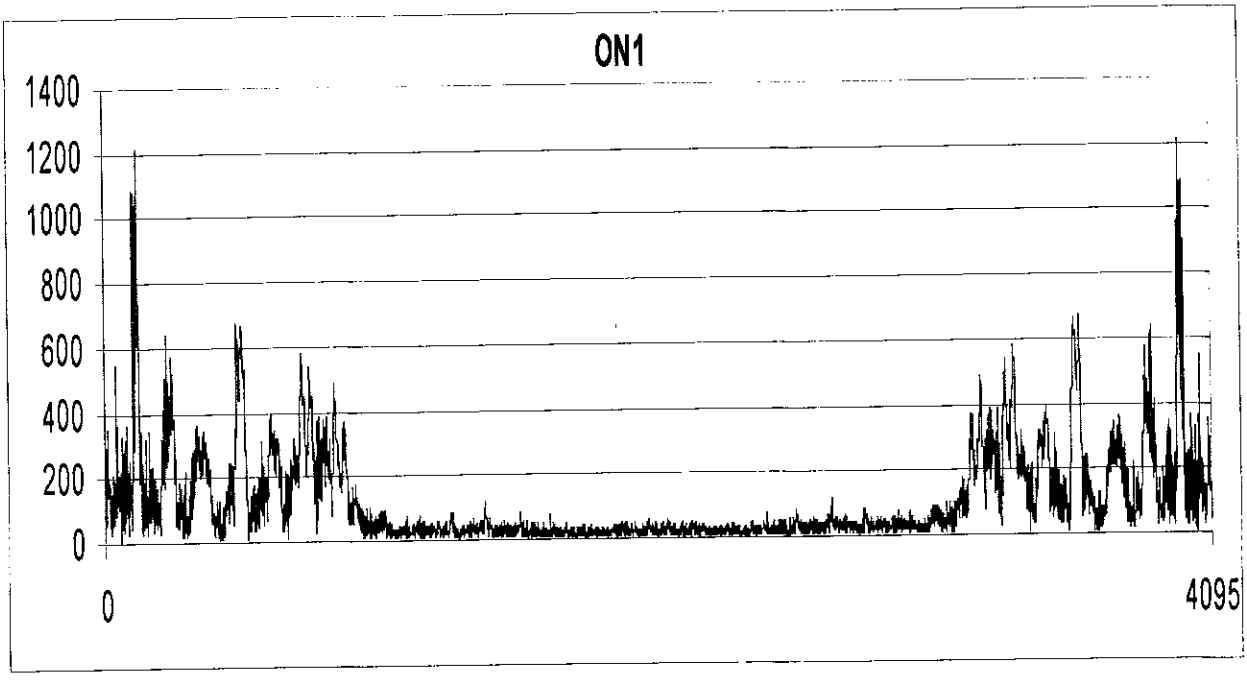
Se comenzará con la descripción en forma detallada de los resultados obtenidos para una de las muestras y la ejemplificación de lo expuesto en la sección anterior en cuanto a las características ruidosas del espectro obtenido, el efecto del filtro digital y la justificación de la frecuencia de muestreo de 5.5 Khz. sin *aliasing*. Luego, se presentarán los resultados obtenidos para cada instrucción verbal, con su respectiva interpretación y se dejará para el final el análisis comparativo entre las muestras y las observaciones relacionadas con el cumplimiento de la hipótesis formulada en el capítulo anterior.

Para los efectos de la primera parte de la descripción de los resultados, se tomará a título de ejemplo la primera observación de la instrucción "ON". La condición de inicio de palabra fue: amplitud >= 2. A continuación se muestra la gráfica de amplitud vs. tiempo obtenida de la lectura de la tarjeta de sonido, con la herramienta de software descrita con anterioridad.



Gráfica No. 1 Amplitud vs. Tiempo.
Observación 1, instrucción "ON".
Total de muestras: 8192.

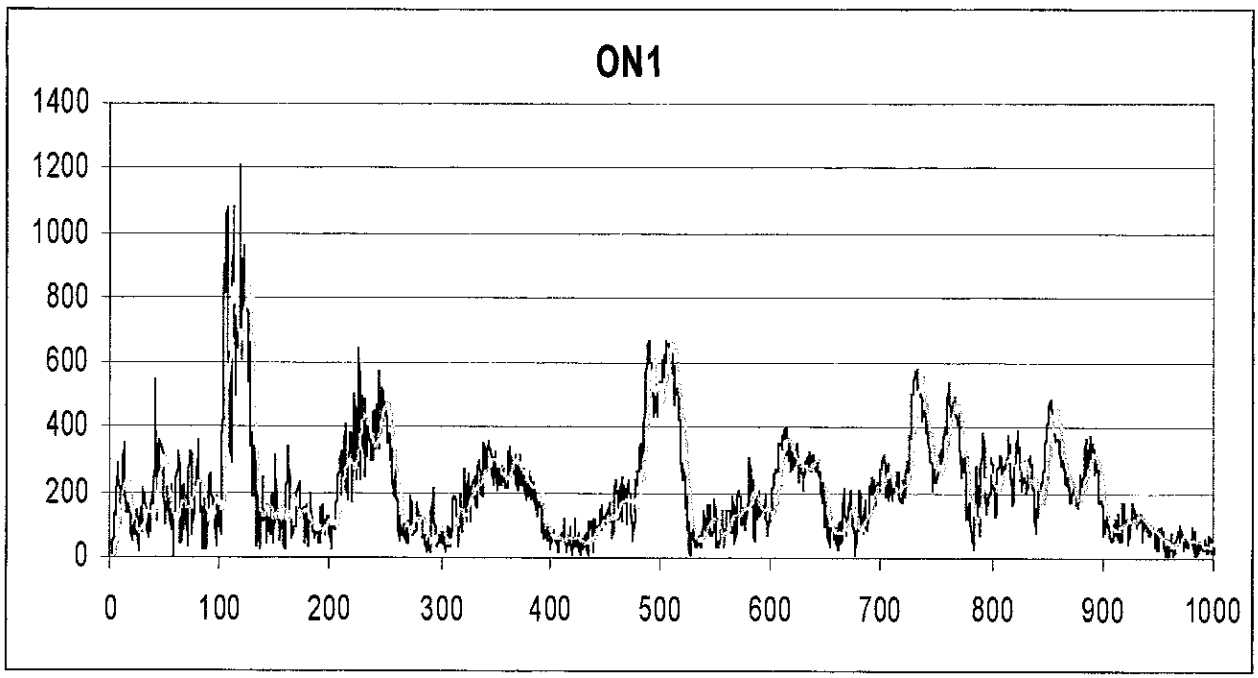
Al aplicar la transformada rápida de Fourier y calcular las amplitudes correspondientes se obtiene el espectro de color azul mostrado en la gráfica No. 2. Se eliminó la componente DC (correspondiente a la muestra cero), como es común en el análisis de frecuencias. Es evidente la característica ruidosa que presenta este espectro, y lo complicado que resulta analizarlo para fines de reconocimiento. Es importante recalcar que este espectro constituye el resultado de aplicar la transformada rápida de Fourier a la mitad de los datos mostrados en la gráfica No.1 (4096 muestras en total), lo que equivale a realizar un muestreo a 5.5 Khz. El espectro muestra que no existe *aliasing* en este caso, por lo que se considera como suficiente el muestreo a esta frecuencia.



Gráfica No. 2 Amplitud vs. Frecuencia.
Observación 1, instrucción "ON".
Total de muestras: 4096.

A este espectro se le aplica entonces la combinación de filtros digitales, con lo que se obtiene la representación "suave" del espectro, mostrada en la gráfica No.3. Para efectos de este primer estudio, se aplicó el filtro con una frecuencia de corte de 1kHz. Al realizar el análisis comparativo se discutirá más sobre la influencia de la selección de la frecuencia de corte en los resultados obtenidos. Por motivos de claridad, se muestra única-

mente la parte del espectro comprendida entre cero y 1000 Hz. La gráfica mostrada en color azul constituye la porción del espectro original (gráfica No. 2) que se está analizando, se incluye aquí para mostrar en una mejor manera el efecto que tiene la aplicación del filtro digital.



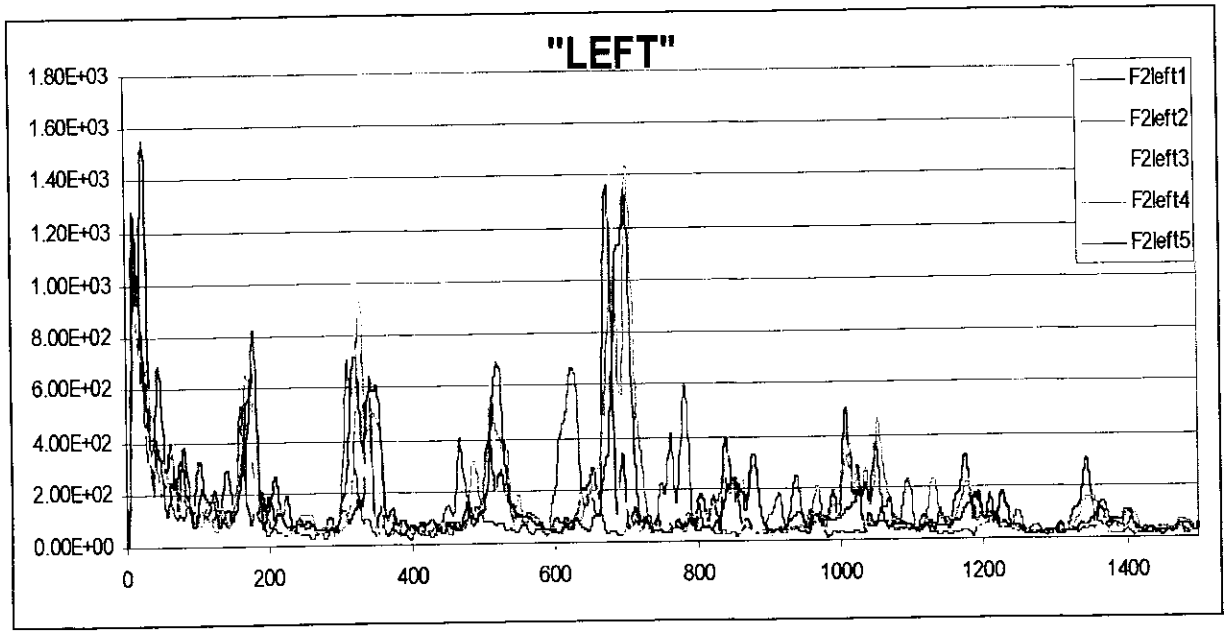
Gráfica No. 3 Espectro original■ y espectro filtrado□ .
Observación 1, instrucción "ON"; frecuencia de corte en filtro = 400 Hz.
Total de muestras: 4096.

Es en base al espectro obtenido después de la aplicación del filtro digital que se pretende estudiar la viabilidad de un reconocimiento de voz. Se desarrolló un compromiso entre la frecuencia de corte del filtro y los valores de correlación obtenidos, pues como es evidente, mientras menor sea la frecuencia de corte de filtro, se estará sacrificando una mayor resolución de los datos, lo que ocasiona que aumenten las correlaciones entre muestras de una misma instrucción y entre los vectores típicos de los mismos. Se determinó entonces que la frecuencia con la cual se logra "suavizar" el espectro de la mejor manera, sin perder componentes significativos del espectro fue de 400 Hz.

Hasta aquí, todo lo que se ha hecho es acomodar los datos arrojados por el apoyo en hardware para que se acoplen a los requisitos del proyecto. Es hasta ahora que se inicia el análisis de viabilidad de reconocimiento, fundamentado en la división original del problema en dos fases, tal y como se indicó en el marco metodológico. Al seguir la estructura del marco metodológico, se divide la presentación de resultados en las mismas dos fases en las que se dividió el problema de reconocimiento. Se debe recordar que dada la simetría del espectro frecuencial, basta tomar únicamente las primeras 2048 amplitudes. Debido a la cantidad de información, los resultados se presentan en forma gráfica, ya que esto facilita su comparación.

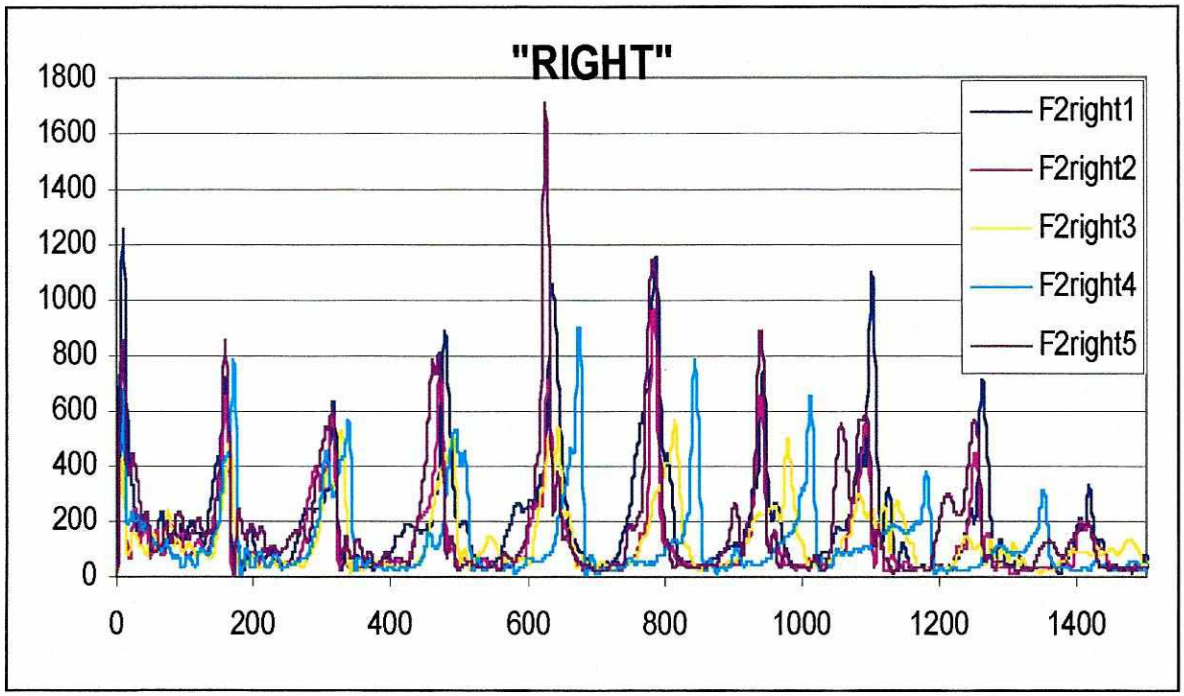
A. Entrenamiento (primera fase):

1. Presentación de muestras de entrenamiento para cada instrucción:



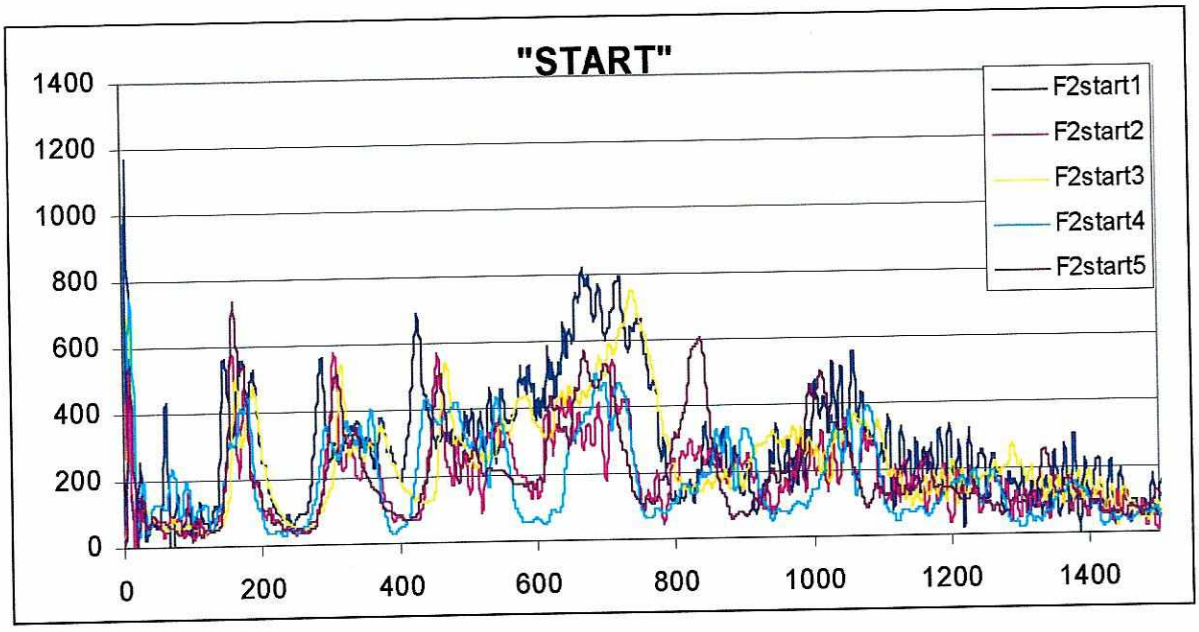
Gráfica No. 4 Espectro filtrado de muestras para entrenamiento, instrucción *"left"*.

Puede observarse que todas las muestras poseen aproximadamente los mismos componentes principales de frecuencias, aunque no coinciden en forma exacta. Se manifiesta un corrimiento de los componentes frecuenciales, principalmente en la segunda muestra (color lila). Es importante notar que el corrimiento no es uniforme a lo largo del espectro. Esto se evidencia más en la segunda muestra, donde el corrimiento se hace más notorio a partir de 300 Hz. Las componentes de alta frecuencia se atribuyen al uso del fricativo no vocalizado "f" al final de la palabra.



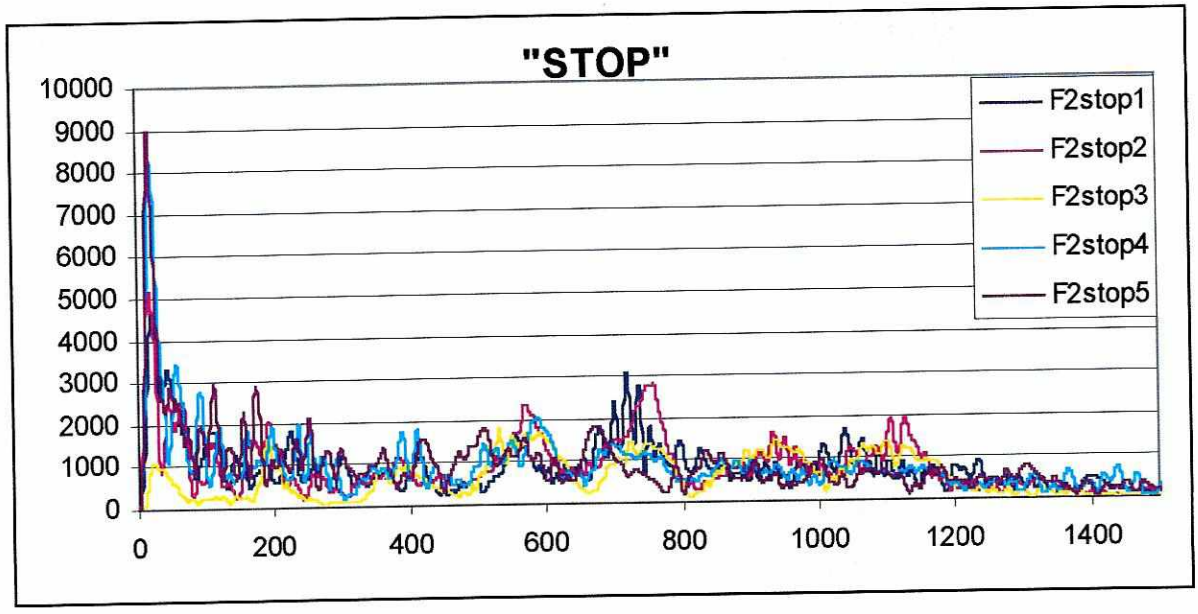
Gráfica No. 5 Espectro filtrado de muestras para entrenamiento, instrucción *"right"*.

Nuevamente se manifiesta un corrimiento de frecuencias, aunque en este caso es menor que en la palabra anterior. El flujo turbulento de aire producido por la espiración al pronunciar la letra "t" final tiene características de ruido que aparecen como componentes frecuenciales elevados.



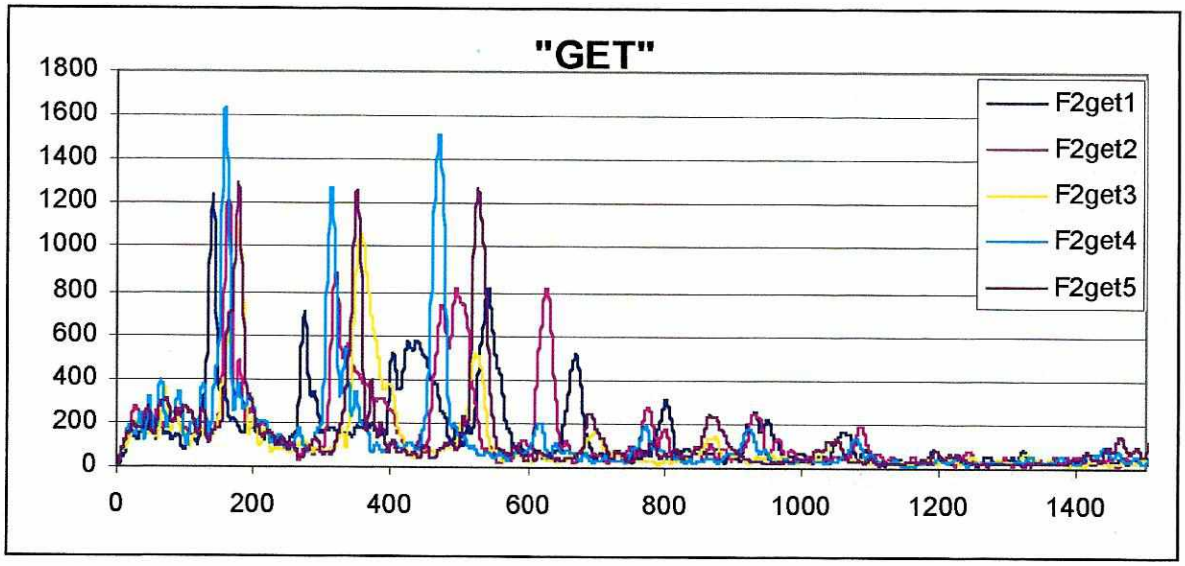
Gráfica No. 6 Espectro filtrado de muestras para entrenamiento, instrucción *"start"*.

En esta instrucción, la fuerte pronunciación inicial de la consonante "s" ocasiona la presencia de componentes "ruidosos" de alta frecuencia. Se observan también las características típicas de consonantes plosivas, al ver los altos componentes de baja frecuencia.



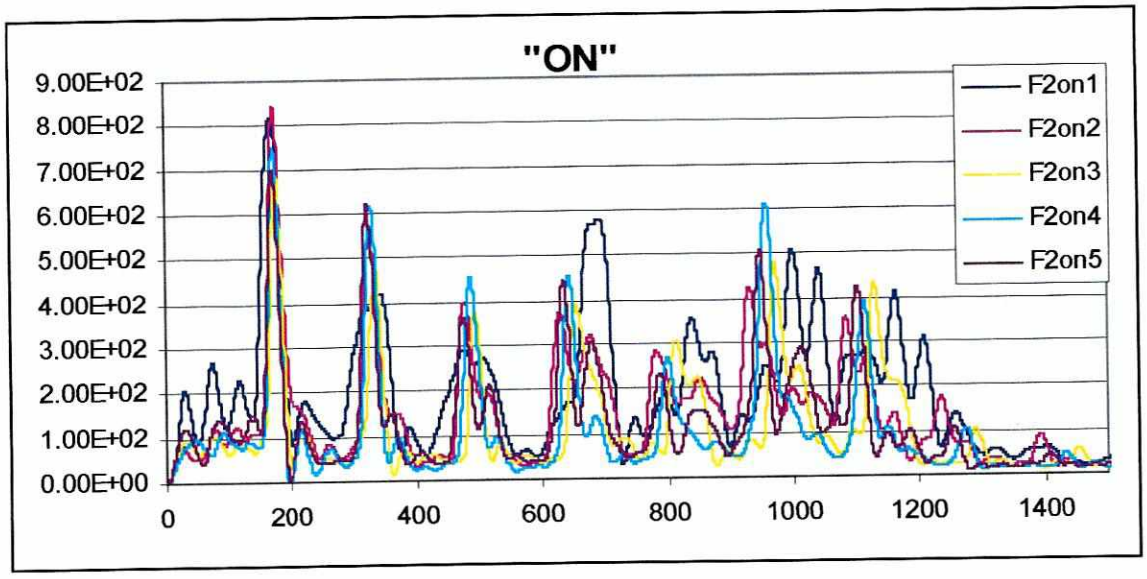
Gráfica No. 7 Espectro filtrado de muestras para entrenamiento, instrucción *"stop"*.

Esta es una de las palabras que muestra mayor variación en su espectro, al comparar las diferentes muestras obtenidas. Los significativos componentes de baja frecuencia son característicos de palabras con consonantes plosivas ubicadas en medio de la palabra. Los plosivos al inicio, no ocasionan estos componentes como puede inferirse del análisis de las muestras de "*time*", que se caracteriza por la consonante plosiva "t" del inicio. En este caso, existen dos consonantes plosivas, la "t" pronunciada después de la "s", y la "p" final. El problema con este tipo de consonantes es la marcada pausa que se hace antes de pronunciarlas, así como el flujo turbulento y repentino que sigue a esta pausa.



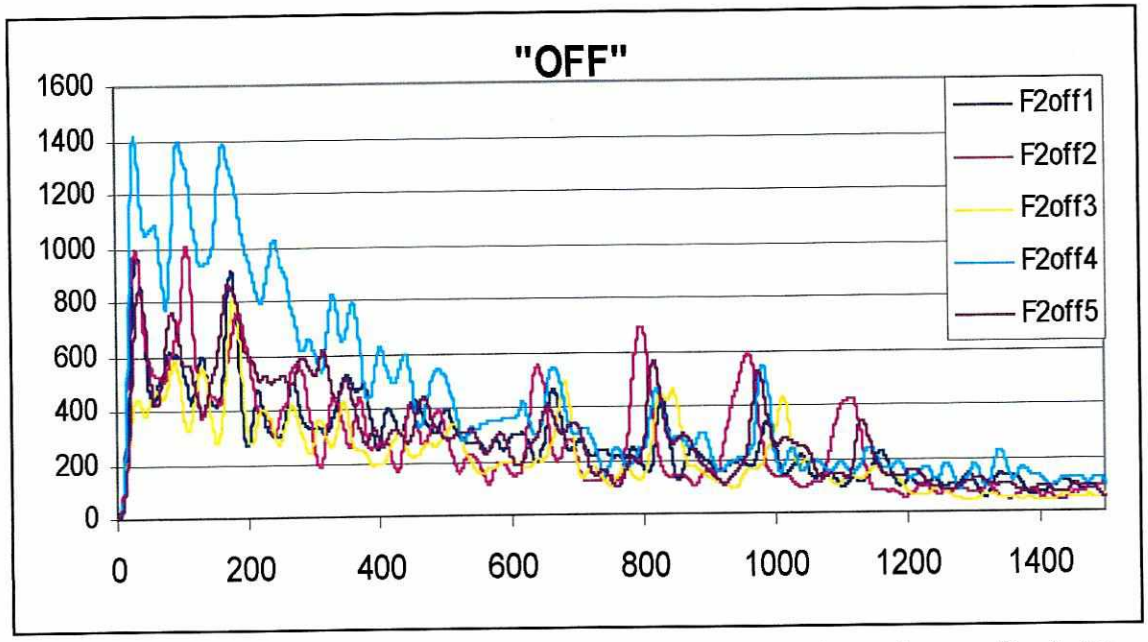
Gráfica No. 8 Espectro filtrado de muestras para entrenamiento, instrucción *"get"*.

De nuevo es evidente el corrimiento de las frecuencias, aunque esta instrucción tiene la característica de tener componentes significativos de baja frecuencia, es decir, bastaría analizar el espectro con los primeros 600 Hz.



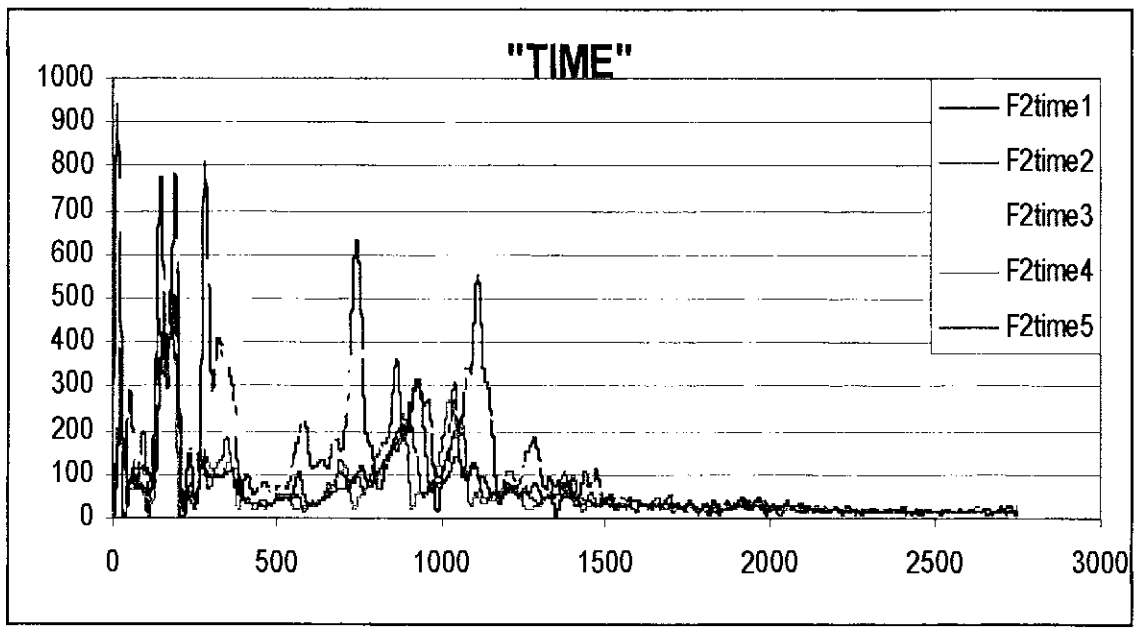
Gráfica No. 9 Espectro filtrado de muestras para entrenamiento, instrucción "*on*".

Los espectros de las muestras de la instrucción "on" son bastante similares. En este caso, no existe presencia ni de consonantes fricativas ni de plosivos. La instrucción constituye una combinación de una vocal fuerte con una vocal nasal. La transición entre ambas es lo suficientemente suave como para obtener un espectro bien definido, por lo menos dentro de los componentes bajos (hasta 700 Hz.).



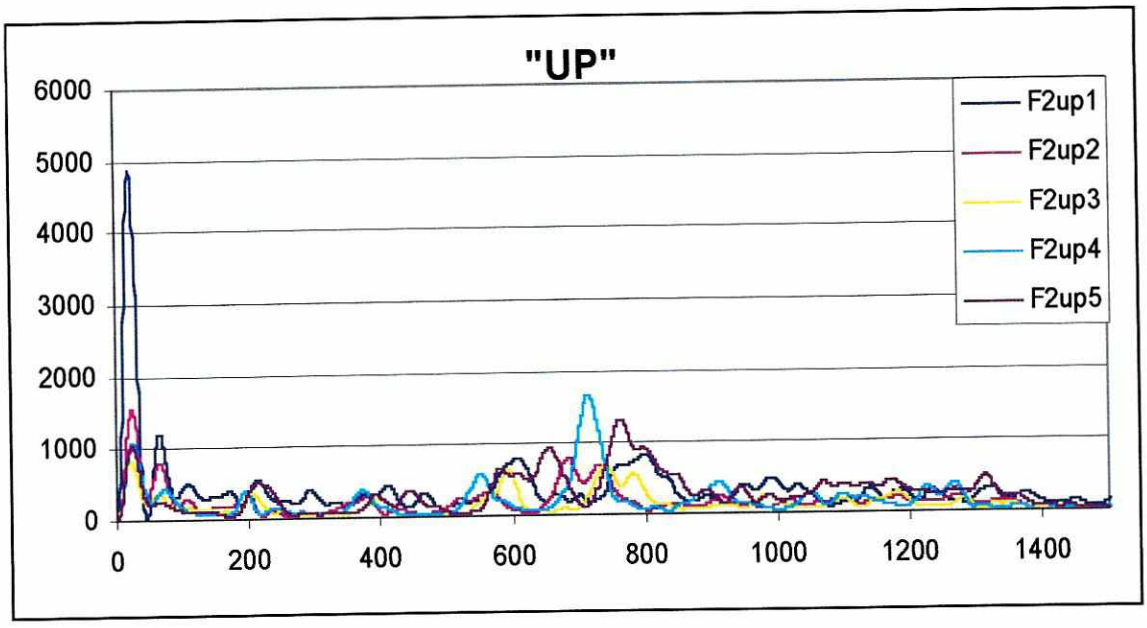
Gráfica No. 10 Espectro filtrado de muestras para entrenamiento, instrucción "*off*".

Nuevamente, la presencia de fricativos no vocalizados (fuerte pronunciación de la letra "f" al final), causa que el espectro aparezca ruidoso aún después de filtrado. Sin embargo, existen definidos componentes de baja frecuencia entre los 50 y 200 Hz, que se atribuyen en este caso a la vocal "o". Se observan también componentes altos de frecuencia entre los 800 y 1200 Hz, mismos que caracterizan a la letra "f" y coinciden con sus contribuciones en la instrucción *"lefi"*, donde aparecieron pero no de forma tan clara y delimitada como en este caso.



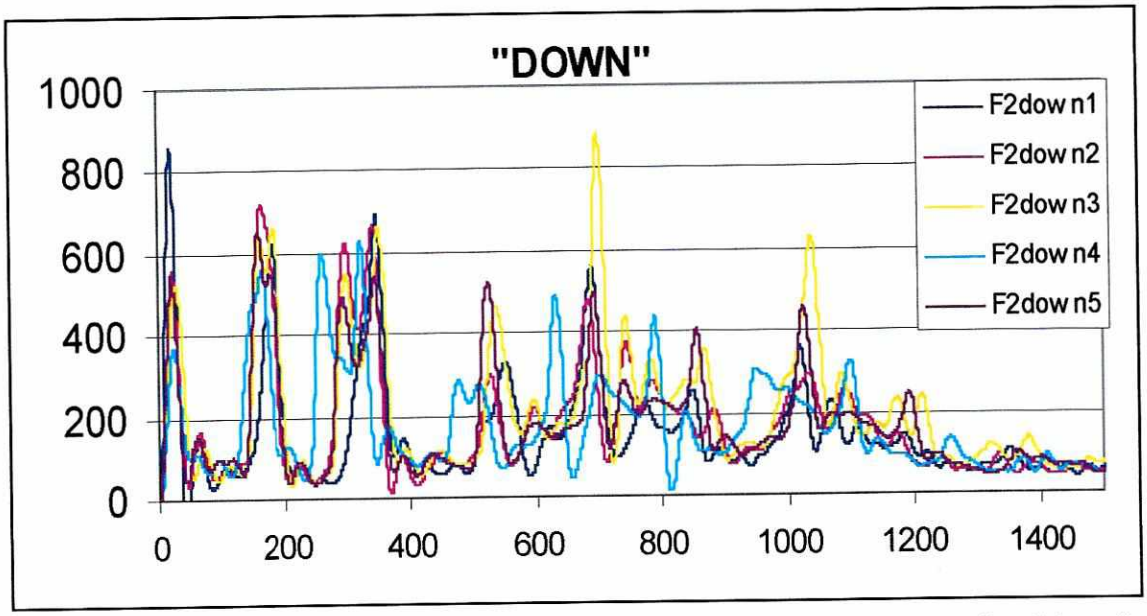
Gráfica No. 11 Espectro filtrado de muestras para entrenamiento, instrucción *"time"*.

Se observa que los componentes de baja frecuencia coinciden mejor que los de alta frecuencia. Debe recordarse que la espiración que acompaña a la instrucción verbal es de hecho un componente de alta frecuencia, que aparece en esta instrucción entre los 1000 y 1500 Hz.



Gráfica No. 12 Espectro filtrado de muestras para entrenamiento, instrucción *"up"*.

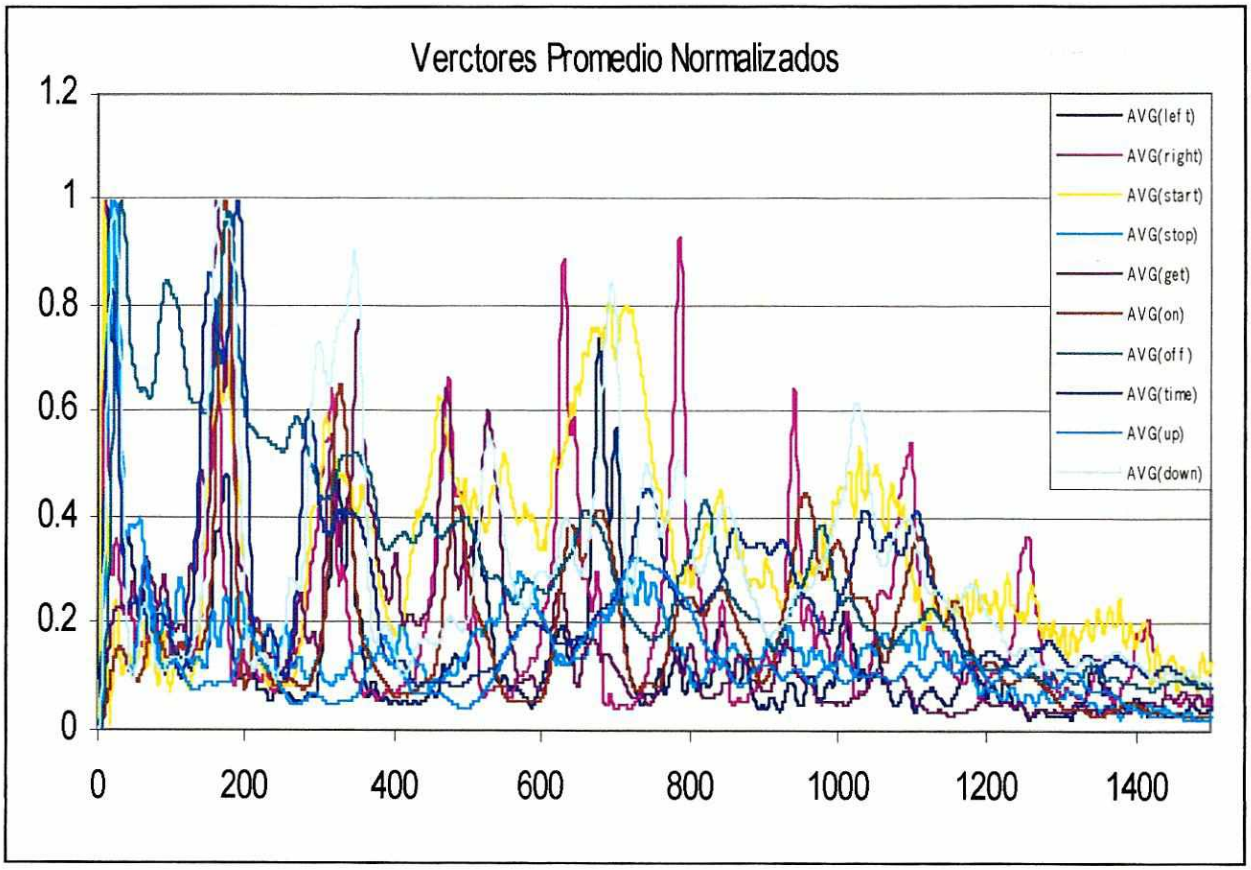
Estos espectros resultan interesantes dado que la presencia de la consonante plosiva al final de la palabra domina completamente el espectro (componente de baja frecuencia). La amplitud tan elevada del plosivo no permite que se aprecien adecuadamente las frecuencias formantes de la vocal "u", mismas que aparecen en el rango aproximado de 50 a 300 Hz.



Gráfica No. 13 Espectro filtrado de muestras para entrenamiento, instrucción *"down"*.

Esta instrucción es bastante parecida a la instrucción "on", aunque muestra un mayor corrimiento de frecuencias. Nótese nuevamente la claridad de los patrones frecuenciales al compararlo con espectros de palabras que contienen consonantes fricativas o plosivas.

La siguiente gráfica muestra los espectros típicos de cada instrucción. La nomenclatura AVG (*instrucción*) se deriva del inglés *average* que significa promedio. Debe recordarse que el espectro típico representa a cada una de las instrucciones al realizar las comparaciones propias de la fase de reconocimiento. Los espectros están normalizados para facilitar su comparación visual.



para facilitar su comparación visual.

Gráfica No. 14 Espectros típicos de cada instrucción, equivalentes a vectores promedio representativos de cada grupo de muestras de entrenamiento.

B. Reconocimiento (segunda fase)

Una vez determinados los vectores promedio que representan a cada muestra, se procedió al análisis de correlaciones entre éstos. Se presentan los datos de correlación y los ángulos correspondientes. Aquí debe recalcarse que dado que se analiza un espectro de amplitudes, todos los componentes de los vectores de cada muestra son positivos, lo que elimina las consideraciones que deben tomarse al calcular funciones trigonométricas inversas de vectores que estén fuera de los cuadrantes uno y cuatro (en el caso de R^2).

Tabla No. 3 A.

Correlaciones entre vectores promedio.

	AVG ("left")	*AVG ("right")*	*AVG ("start")*	*AVG ("stop")*	*AVG ("get")*	*AVG ("on")*	*AVG ("off")*	*AVG ("time")*	*AVG ("up")*	*AVG ("down")*
AVG("left")	1									
AVG("right")	0.432402	1								
AVG("start")	0.569209	0.564366	1							
AVG("stop")	0.695682	0.406875	0.531928	1						
AVG(get)	0.498341	0.428522	0.491826	0.382448	1					
AVG(on)	0.501256	0.668601	0.664392	0.378753	0.546237	1				
AVG(off)	0.57627	0.489487	0.480283	0.680024	0.70754	0.616028	1			
AVG("time")	0.518843	0.483039	0.631561	0.623909	0.54163	0.621848	0.670863	1		
AVG("up")	0.595684	0.377898	0.552101	0.842717	0.198213	0.290857	0.488027	0.554092	1	
AVG("down")	0.669344	0.557366	0.756089	0.559636	0.57124	0.771089	0.621147	0.795531	0.526353	1

Tabla No. 3 B.

Ángulos (en grados sexagesimales) entre vectores promedio (en R^{2048})

	AVG ("left")	*AVG ("right")*	*AVG ("start")*	*AVG ("stop")*	*AVG ("get")*	*AVG ("on")*	*AVG ("off")*	*AVG ("time")*	*AVG ("up")*	*AVG ("down")*
AVG("left")	0.0									
AVG("right")	64.4	0.0								
AVG("start")	55.3	55.6	0.0							
AVG("stop")	45.9	66.0	57.9	0.0						
AVG(get)	60.1	64.6	60.5	67.5	0.0					
AVG(on)	59.9	48.0	48.4	67.7	56.9	0.0				
AVG(off)	54.8	60.7	61.3	47.2	45.0	52.0	0.0			
AVG("time")	56.7	61.1	50.8	51.4	57.2	51.5	47.9	0.0		
AVG("up")	53.4	67.8	56.5	32.6	78.6	73.1	60.8	56.4	0.0	
AVG("down")	48.0	56.1	40.9	56.0	55.2	39.5	51.6	37.3	58.2	0.0

Al observar los valores de las tablas 3 A y B, se hace evidente la importancia de convertir los valores de correlación a una escala lineal (ángulos en grados sexagesimales), pues valores aparentemente muy próximos en la tabla 3 A equivalen a separaciones mayores en una escala lineal. En la tabla No.3 B, se observa que los vectores promedio que están más cercanos corresponden a la instrucción *"up"* y a la instrucción *"stop"*, separados por un ángulo de 32.6 grados. El promedio de separación de los valores mostrados en la tabla No. 3 B es de 55.6 grados. Los vectores más separados son *get* y *"up"* con 78.6 grados. Debe recordarse que el ángulo entre los vectores constituye la regla de discriminación para la clasificación de las instrucciones. Al ingresar una nueva muestra, el sistema ya entrenado se limita al cálculo de ángulos entre la muestra y los vectores promedio determinados y permite su correcta clasificación. Antes de pasar al detalle de los procedimientos de discriminación conviene realizar un análisis de la dispersión de las muestras de entrenamiento, alrededor del vector promedio que las representa. Nuevamente, la dispersión se medirá en términos del ángulo entre los vectores.

Las Tablas 4 a 13 presentan los análisis de correlación entre las muestras de entrenamiento y los vectores promedio calculados. Se incluyen tres tablas, la primera contiene los resultados de la matriz de correlación, la segunda representa los ángulos correspondientes y la última presenta un resumen de los ángulos obtenidos para cada instrucción.

Tabla No.4 A.

Análisis de correlaciones instrucción *"left"*.

	F2left1	*F2left2*	*F2left3*	*F2left4*	*F2left5*	*AVG("left")*
F2left1	1					
F2left2	0.479514	1				
F2left3	0.741228	0.384631	1			
F2left4	0.651849	0.320371	0.724744	1		
F2left5	0.609254	0.394679	0.654986	0.889177	1	
AVG("left")	0.8522	0.586301	0.853719	0.899411	0.886815	1

Tabla No.4 B.

Análisis de correlaciones instrucción *"left"*, en grados.

	F2left1	*F2left2*	*F2left3*	*F2left4*	*F2left5*	*AVG("left")*
F2left1	0					
F2left2	61.34635	0				
F2left3	42.16389	67.37917	0			
F2left4	49.31887	71.31461	43.55243	0		
F2left5	52.46443	66.75406	49.08139	27.22993	0	
AVG("left")	31.54822	54.10508	31.38145	25.91921	27.52426	0

Al analizar los valores de la tabla No. 4 B, se observa una significativa variabilidad entre los ángulos que separan a cada una de las diferentes muestras del vector promedio. Esta variabilidad podría justificarse dada la variación observada entre los espectros de la gráfica No. 4. La siguiente tabla resume las principales características de los ángulos propios entre las muestras y el vector promedio. Entre éstas se indica el promedio de los ángulos obtenidos, que da una idea de la dispersión de las muestras alrededor del vector promedio.

Tabla No. 4 C.

Resumen de ángulos entre muestras de entrenamiento y vector promedio instrucción *"left"*.

Valor máximo (grados)	# de muestra más alejada	Valor mínimo (grados)	# de muestra más cercana	Promedio de ángulos
54.10508	2	25.91921	4	34.09564

Tabla No.5 A.

Análisis de correlaciones instrucción *"right"*.

	F2right1	*F2right2*	*F2right3*	*F2right4*	*F2right5*	*AVG("right")*
F2right1	1					
F2right2	0.784735	1				
F2right3	0.658882	0.620316	1			
F2right4	0.254838	0.220436	0.433951	1		
F2right5	0.704436	0.883716	0.55299	0.168689	1	
AVG("right")	0.885283	0.907046	0.793273	0.473373	0.871261	1

Tabla No.5 B.

Análisis de correlaciones instrucción *"right"*, en grados.

	F2right1	*F2right2*	*F2right3*	*F2right4*	*F2right5*	*AVG("right")*
F2right1	0					
F2right2	38.30381	0				
F2right3	48.78537	51.66077	0			
F2right4	75.236	77.26537	64.28144	0		
F2right5	45.21601	27.90613	56.42763	80.28841	0	
AVG("right")	27.71369	24.89977	37.50756	61.74653	29.39453	0

Tabla No. 5 C.

Resumen de ángulos entre muestras de entrenamiento y vector promedio instrucción *"right"*.

Valor máximo (grados)	# de muestra más alejada	Valor mínimo (grados)	# de muestra más cercana	Promedio de ángulos
61.74653	4	24.89977	2	36.25242

Tabla No.6 A.

Análisis de correlaciones instrucción *"start"*.

	F2start1	*F2start2*	*F2start3*	*F2start4*	*F2start5*	*AVG("start")*
F2start1	1					
F2start2	0.764935	1				
F2start3	0.829741	0.764306	1			
F2start4	0.695271	0.794826	0.683625	1		
F2start5	0.701751	0.83837	0.69787	0.713414	1	
AVG("start")	0.911379	0.920686	0.899747	0.855696	0.876139	1

Tabla No.6 B.

Análisis de correlaciones instrucción *"start"*, en grados.

	F2start1	*F2start2*	*F2start3*	*F2start4*	*F2start5*	*AVG("start")*
F2start1	0					
F2start2	40.0988	0				
F2start3	33.9279	40.15471	0			
F2start4	45.95119	37.36121	46.87245	0		
F2start5	45.43232	33.03155	45.7436	44.48663	0	
AVG("start")	24.30342	22.97343	25.87517	31.1633	28.81992	0

Tabla No. 6 C.

Resumen de ángulos entre muestras de entrenamiento y vector promedio instrucción *"start"*.

Valor máximo (grados)	# de muestra más alejada	Valor mínimo (grados)	# de muestra más cercana	Promedio de ángulos
28.81992	5	22.97343	2	26.62705

Tabla No.7 A.

Análisis de correlaciones instrucción *"stop"*.

	F2stop1	*F2stop2*	*F2stop3*	*F2stop4*	*F2stop5*	*AVG("stop")*
F2stop1	1					
F2stop2	0.795424	1				
F2stop3	0.548598	0.744399	1			
F2stop4	0.785544	0.784978	0.488675	1		
F2stop5	0.759464	0.707005	0.33292	0.870052	1	
AVG("stop")	0.900432	0.922771	0.672085	0.929199	0.880958	1

Tabla No.7 B.

Análisis de correlaciones instrucción *"stop"*, en grados.

	F2stop1	*F2stop2*	*F2stop3*	*F2stop4*	*F2stop5*	*AVG("stop")*
F2stop1	0					
F2stop2	37.30464	0				
F2stop3	56.72915	41.89254	0			
F2stop4	38.22903	38.28133	60.74648	0		
F2stop5	40.583	45.00822	70.55387	29.53533	0	
AVG("stop")	25.7851	22.66539	47.77181	21.68963	28.24183	0

Tabla No. 7 C.

Resumen de ángulos entre muestras de entrenamiento y vector promedio instrucción *"stop"*.

Valor máximo (grados)	# de muestra más alejada	Valor mínimo (grados)	# de muestra más cercana	Promedio de ángulos
47.77181	3	21.68963	4	29.230752

Tabla No.8 A.

Análisis de correlaciones instrucción *get*.

	F2get1	*F2get2*	*F2get3*	*F2get4*	*F2get5*	*AVG(get)*
F2get1	1					
F2get2	0.239338	1				
F2get3	0.309754	0.496013	1			
F2get4	0.336978	0.667889	0.32068	1		
F2get5	0.327168	0.420754	0.825186	0.273966	1	
AVG(get)	0.58159	0.781848	0.793768	0.735695	0.768806	1

Tabla No.8 B.

Análisis de correlaciones instrucción *get*, en grados.

	F2get1	*F2get2*	*F2get3*	*F2get4*	*F2get5*	*AVG(get)*
F2get1	0					
F2get2	76.15255	0				
F2get3	71.95559	60.26344	0			
F2get4	70.30715	48.09567	71.29596	0		
F2get5	70.90303	65.1178	34.39267	74.09962	0	
AVG(get)	54.43755	38.56988	37.46093	42.634	39.75321	0

Tabla No. 8 C.

Resumen de ángulos entre muestras de entrenamiento y vector promedio instrucción *get*.

Valor máximo (grados)	# de muestra más alejada	Valor mínimo (grados)	# de muestra más cercana	Promedio de ángulos
54.43755	1	37.46093	3	42.571114

Tabla No.9 A.

Análisis de correlaciones instrucción *on*.

	F2on1	*F2on2*	*F2on3*	*F2on4*	*F2on5*	*AVG(on)*
F2on1	1					
F2on2	0.793835	1				
F2on3	0.713422	0.711834	1			
F2on4	0.610861	0.789941	0.743158	1		
F2on5	0.743333	0.880335	0.591922	0.82573	1	
AVG(on)	0.884618	0.939554	0.837901	0.878935	0.906615	1

Tabla No.9 B.

Análisis de correlaciones instrucción *on*, en grados.

	F2on1	*F2on2*	*F2on3*	*F2on4*	*F2on5*	*AVG(on)*
F2on1	0					
F2on2	37.45467	0				
F2on3	44.48595	44.61567	0			
F2on4	52.34819	37.82003	41.99883	0		
F2on5	41.98391	28.3172	53.70649	34.33746	0	
AVG(on)	27.79548	20.02323	33.08087	28.4858	24.9583	0

Tabla No. 9 C.

Resumen de ángulos entre muestras de entrenamiento y vector promedio instrucción *on*.

Valor máximo (grados)	# de muestra más alejada	Valor mínimo (grados)	# de muestra más cercana	Promedio de ángulos
33.08087	3	20.02323	2	26.868736

Tabla No.10 A.

Análisis de correlaciones instrucción *off*.

	F2off1	*F2off2*	*F2off3*	*F2off4*	*F2off5*	*AVG(off)*
F2off1	1					
F2off2	0.817339	1				
F2off3	0.8964	0.758983	1			
F2off4	0.915358	0.845903	0.867009	1		
F2off5	0.905646	0.841206	0.901068	0.923381	1	
AVG(off)	0.955033	0.904647	0.924228	0.972879	0.965497	1

Tabla No.10 B.

Análisis de correlaciones instrucción *off*, en grados.

	F2off1	*F2off2*	*F2off3*	*F2off4*	*F2off5*	*AVG(off)*
F2off1	0					
F2off2	35.18073	0				
F2off3	26.31116	40.62539	0			
F2off4	23.74339	32.23121	29.88707	0		
F2off5	25.08956	32.73229	25.70119	22.57454	0	
AVG(off)	17.24744	25.22423	22.44784	13.37445	15.0947	0

Tabla No. 10 C.

Resumen de ángulos entre muestras de entrenamiento y vector promedio instrucción *off*.

Valor máximo (grados)	# de muestra más alejada	Valor mínimo (grados)	# de muestra más cercana	Promedio de ángulos
25.22423	2	13.37455	4	18.677732

Tabla No.11 A.

Análisis de correlaciones instrucción *"time"*.

	F2time1	*F2time2*	*F2time3*	*F2time4*	*F2time5*	*AVG("time")*
F2time1	1					
F2time2	0.66795	1				
F2time3	0.778088	0.530087	1			
F2time4	0.588249	0.873135	0.612493	1		
F2time5	0.685771	0.803614	0.674545	0.791357	1	
AVG("time")	0.906138	0.839624	0.87114	0.82984	0.873199	1

Tabla No.11 B.

Análisis de correlaciones instrucción *"time"*, en grados.

	F2time1	*F2time2*	*F2time3*	*F2time4*	*F2time5*	*AVG("time")*
F2time1	0					
F2time2	48.09094	0				
F2time3	38.91412	57.98865	0			
F2time4	53.96717	29.17497	52.23005	0		
F2time5	46.7037	36.52335	47.58119	37.68747	0	
AVG("time")	25.02299	32.89951	29.40859	33.91769	29.16744	0

Tabla No. 11 C.

Resumen de ángulos entre muestras de entrenamiento y vector promedio instrucción *"time"*.

Valor máximo (grados)	# de muestra más alejada	Valor mínimo (grados)	# de muestra más cercana	Promedio de ángulos
33.91769	4	25.02299	1	30.083244

Tabla No.12 A.

Análisis de correlaciones instrucción *"up"*.

	F2up1	*F2up2*	*F2up3*	*F2up4*	*F2up5*	*AVG("up")*
F2up1	1					
F2up2	0.713002	1				
F2up3	0.683039	0.689465	1			
F2up4	0.391444	0.757631	0.541831	1		
F2up5	0.537374	0.498537	0.710202	0.295419	1	
AVG("up")	0.877365	0.883823	0.860772	0.684757	0.740127	1

Tabla No.12 B.

Análisis de correlaciones instrucción *"up"*, en grados.

	F2up1	*F2up2*	*F2up3*	*F2up4*	*F2up5*	*AVG("up")*
F2up1	0					
F2up2	44.52034	0				
F2up3	46.91838	46.41219	0			
F2up4	66.95561	40.74419	57.19162	0		
F2up5	57.49493	60.09672	44.74861	72.81731	0	
AVG("up")	28.67388	27.89299	30.59658	46.78346	42.25777	0

Tabla No. 12 C.

Resumen de ángulos entre muestras de entrenamiento y vector promedio instrucción *"up"*.

Valor máximo (grados)	# de muestra más alejada	Valor mínimo (grados)	# de muestra más cercana	Promedio de ángulos
46.78346	4	27.89299	2	35.240936

Tabla No.13 A.

Análisis de correlaciones instrucción *"down"*, en grados.

	F2down1	*F2down2*	*F2down3*	*F2down4*	*F2down5*	*AVG("down")*
F2down1	1					
F2down2	0.795058	1				
F2down3	0.799983	0.843031	1			
F2down4	0.502283	0.6858	0.582683	1		
F2down5	0.815634	0.921632	0.886939	0.659021	1	
AVG("down")	0.875503	0.950232	0.927577	0.756374	0.959585	1

Tabla No.13 B.

Análisis de correlaciones instrucción *"down"*, en grados.

	F2down1	*F2down2*	*F2down3*	*F2down4*	*F2down5*	*AVG("down")*
F2down1	0					
F2down2	37.33928	0				
F2down3	36.87151	32.53845	0			
F2down4	59.84883	46.70146	54.3605	0		
F2down5	35.3499	22.83416	27.50895	48.77471	0	
AVG("down")	28.89549	18.15221	21.93976	40.85441	16.34489	0

Tabla No. 13 C.

Resumen de ángulos entre muestras de entrenamiento y vector promedio instrucción *"down"*.

Valor máximo (grados)	# de muestra más alejada	Valor mínimo (grados)	# de muestra más cercana	Promedio de ángulos
40.85441	4	16.34489	5	25.237352

De las trece tablas anteriores, se concluye que la instrucción cuyas muestras están más dispersas es *get*, con un ángulo promedio entre muestras y vector promedio de 42.6 grados. Por otro lado, las muestras de la instrucción *off* son las menos dispersas, con un ángulo promedio de 18.7 grados. Al colocar las instrucciones en orden descendente de dispersión de muestras de entrenamiento, se obtiene la siguiente tabla:

Tabla No.14

Instrucciones ordenadas según dispersión de muestra

Instrucción	Promedio	Valor máximo (grados)	Valor mínimo (grados)
get	42.57111	54.44	37.46
"right"	36.25242	61.75	24.90
"up"	35.24094	46.78	27.89
"left"	34.09564	54.11	25.92
"time"	30.08324	33.92	25.02
"stop"	29.23075	47.77	21.70
On	26.86874	33.08	20.02
"start"	26.62705	28.82	22.97
"down"	25.23735	40.85	16.34
Off	18.67773	25.22	13.37

Si se analiza la pronunciación de cada una de estas instrucciones, puede explicarse el por qué de su desigual dispersión. Nótese que las instrucciones menos elaboradas, es decir, los que involucran la pronunciación de un menor número de fonemas, son los grupos más compactos de muestras. Esto debido a que se están limitando drásticamente las frecuencias que se emiten, al grado de solamente pronunciar dos fonemas distintivos en la palabra *off*, por ejemplo.

A medida que el número de fonemas pronunciados aumenta, las muestras tienden a estar más separadas entre sí. Existe cierta contradicción con la instrucción *on* y lo expuesto anteriormente en cuanto al número de fonemas pronunciados. Al ver esta instrucción, puede pensarse que la instrucción *on* debería tener una distribución menos dispersa (como el caso de *off*). La razón por la que esto no es así radica en el uso de una consonante nasal (la letra "n"), que es de hecho un sonido vocalizado (ver marco teórico). Las consonantes nasales son fonemas de elevada energía, por lo que muestran componentes frecuenciales significativos que son susceptibles a variación al pronunciar la instrucción. Las consonantes fricativas, por el contrario, poseen características de ruido, y no permiten al orador un amplio grado de variación al pronunciarlas. De esta manera, una persona puede pronunciar una "n" en formas significativamente distintas, por lo que varía significativamente la agudeza del sonido. Contrariamente, las letras "s" o "f" no son tan fácilmente variables en cuanto a pronunciación.

La instrucción con mayor variabilidad es *get*, lo cual es de esperarse luego de analizar los espectros obtenidos (gráfica No. 8), donde claramente hay poca correspondencia entre los componentes frecuenciales de cada una de las muestras con que se entrena el sistema. Lo anterior se debe posiblemente a la combinación de dos consonantes plosivas en una palabra tan corta. El nivel de variabilidad aquí es bastante diferente, pues la vocal plosiva afectará la forma en que se pronuncie eventualmente la vocal fuerte "e". Esto explicaría por qué se observan 4 componentes de frecuencia distintivos en la gráfica 8, pero desplazados en forma irregular en cada una de las muestras tomadas.

Resulta interesante que las instrucciones con consonantes plosivas después de una vocal obtienen ángulos promedio de dispersión bastante parecidos (las instrucciones *"left", "right"* y *"up"*). Esto confirma que las consonantes plosivas varían sus características según su posición dentro de la palabra.

Para completar el análisis y determinar si el reconocimiento de voz basado en los espectros frecuenciales de las instrucciones es viable (dentro del marco establecido), se tomaron primero los vectores promedio determinados en la sección de entrenamiento. Posteriormente, se tomó un nuevo grupo de instrucciones (una muestra de cada uno), y se sometieron al proceso de discriminación basado en la matriz de correlación. Las tablas que se muestran a continuación resumen la información de correlaciones entre las nuevas muestras y los vectores promedio. Nótese que la tabla en sí no es una matriz de correlación (no es una matriz simétrica), sino que está formada por las relaciones de cada uno de los nuevos instrucciones con los promedios previamente definidos. Recuérdese que las nuevas muestras que se están evaluando fueron previamente tratadas en igual forma que las muestras con que se entrena el sistema (mismo filtro digital, Etc.). Nuevamente, se incluye la información en términos de correlaciones y de ángulos.

Tabla No. 15 A.

Análisis de correlaciones, nuevas muestras vs. vectores promedio.

	AVG ("left")	*AVG ("right")*	*AVG ("start")*	*AVG ("stop")*	*AVG (get)*	*AVG (on)*	*AVG (off)*	*AVG ("time")*	*AVG ("up")*	*AVG ("down")*
F2left6	0.502253	0.494356	0.504277	0.654416	0.573799	0.537226	0.819780	0.626120	0.503265	0.609467
F2right6	0.246212	0.532057	0.616900	0.349311	0.382585	0.561921	0.366959	0.474298	0.351654	0.562000
F2start6	0.440914	0.588202	0.653393	0.340423	0.488752	0.727537	0.487174	0.405950	0.317175	0.592920
F2stop6	0.607785	0.300591	0.401100	0.643285	0.341443	0.490746	0.732829	0.508276	0.538675	0.538520
F2get6	0.199153	0.416484	0.390624	0.172270	0.538511	0.453339	0.421429	0.316339	0.052293	0.412845
F2on6	0.279210	0.447608	0.518010	0.265128	0.385605	0.631662	0.352908	0.587924	0.236670	0.617601
F2off6	0.212101	0.641831	0.480286	0.273326	0.157444	0.665717	0.292646	0.386188	0.364375	0.467747
F2time6	0.341394	0.140415	0.480317	0.296016	0.396098	0.353326	0.334302	0.597686	0.231958	0.520139
F2up6	0.426367	0.321242	0.666429	0.448787	0.509964	0.430819	0.513578	0.438466	0.398457	0.548646
F2down6	0.29253	0.423687	0.516929	0.335237	0.484375	0.489464	0.4706254	0.536308	0.243744	0.626047

Tabla No. 15 B.

Análisis de correlaciones, nuevas muestras vs. vectores promedio, en grados.

	AVG ("left")	AVG ("right")	AVG ("start")	AVG ("stop")	AVG (get)	AVG (on)	AVG (off)	AVG ("time")	AVG ("up")	AVG ("down")
F2left6	59.85083	60.37271	59.71665	49.12461	54.98444	57.50499	34.93720	51.23557	59.78376	52.44906
F2right6	75.74654	57.85545	51.90986	69.55482	67.50608	55.81127	68.47181	61.68637	69.41150	55.80578
F2start6	63.83778	53.97050	49.20212	70.09738	60.74139	43.31965	60.84501	66.04935	71.50783	53.63552
F2stop6	52.57049	72.50687	66.35304	49.96279	70.03519	60.61037	42.87594	59.45092	57.40651	57.41706
F2get6	78.51256	65.38722	67.00666	80.08018	57.41770	63.04186	65.07515	71.55831	87.00246	65.61635
F2on6	73.78696	63.40967	58.80113	74.62546	67.31870	50.82717	69.33470	53.99014	76.30991	51.85883
F2off6	77.75450	50.07153	61.29593	74.13772	80.94143	48.26264	72.98355	67.28251	68.63086	62.11188
F2time6	70.03816	81.92812	61.29391	72.78152	66.66552	69.30910	70.46991	53.29565	76.58762	58.65843
F2up6	64.76279	71.26197	48.20799	63.33410	59.33855	64.48048	59.09752	63.99398	66.51823	56.72584
F2down6	72.990536	64.93245	58.873489	70.413043	61.028497	60.694668	61.925102	57.567337	75.892353	51.240899

En las casillas color celeste se resaltan los valores de correlación entre las nuevas muestras de cada instrucción y su vector promedio correspondiente. En las casillas marcadas de color naranja, se resaltan los casos en los que una instrucción verbal está más correlacionada (es decir, el vector se encuentra más próximo) con un vector promedio de una instrucción distinta, que con su vector característico. En rojo se resalta la mayor de estas correlaciones (menor ángulo), que indica la pertenencia que el algoritmo hubiera asignado a la instrucción, la cual resulta incorrecta. La diferencia se hace aún más notoria en la tabla 15 B, donde se comparan los valores angulares de separación entre los vectores. La herramienta funciona para discriminar únicamente las instrucciones *get, on, "time"* y *"down"*. Resulta interesante el hecho de que la instrucción con mayor dispersión de muestras según la tabla 14 es reconocido correctamente, mientras que la instrucción cuyas muestras se encuentran menos dispersas es uno de los que el algoritmo falla en interpretar. En el caso ideal de reconocimiento de todas las instrucciones, las tablas 15 "A" y "B" exhibirían únicamente las casillas de color celeste. La comparación de la información contenida en las tablas 14 y 15 B demuestra el hecho de que una distribución poco dispersa de muestras no garantiza el adecuado reconocimiento de dichas instrucciones verbales. Se manifiesta un bajo porcentaje de correcto reconocimiento (40%).

CAPITULO V
CONCLUSIONES Y RECOMENDACIONES

A. Conclusiones

1. La utilización de un *buffer* circular durante la captura de la instrucción y la definición de principio de palabra en base a un nivel de amplitud determinado contrarrestan el efecto de diferencia en inicio de pronunciación.

2. La restricción de vocabulario implementada restringe el nivel de variación en la duración de cada instrucción.

3. Los resultados analizados son independientes de la amplitud relativa de emisión de la instrucción, dada la naturaleza del análisis de correlaciones.

4. La aplicación de un filtro digital al espectro frecuencial permite la observación clara y objetiva de sus principales características.

5. Una baja dispersión de muestras de una misma instrucción no implica que éstos sean fácilmente reconocibles.

6. Las instrucciones estudiadas que incluyen en su pronunciación consonantes plosivas emitidas después de una vocal manifiestan una elevada dispersión de las muestras.

7. La dispersión de muestras de una misma instrucción es proporcional al número de fonemas distintos que lo forman.

8. Las consonantes fricativas (con características de ruido) limitan la dispersión de muestras de una misma instrucción, debido a que no permiten significativa variabilidad en su pronunciación.

9. La hipótesis inicial propuesta se cumplió en condiciones experimentales. El reconocimiento del 40% de las instrucciones estudiadas se logró por medio de un análisis de correlación.

10. El porcentaje obtenido es insuficiente para asegurar la viabilidad de la aplicación de esta herramienta para implementar este procedimiento en situaciones reales.

B. Recomendaciones

1. Es necesario profundizar en la aplicación de otras herramientas matemáticas aplicables en la investigación del reconocimiento de voz basado en espectros frecuenciales, para explorar la posibilidad de lograr mayores porcentajes de reconocimiento que los obtenidos con el análisis de correlación.

2. Con la utilización de las definiciones del marco metodológico de esta investigación, las futuras investigaciones pueden orientarse durante la fase de aprendizaje al desarrollo de una técnica basada en algoritmos de redes neuronales de clasificación, de modo que permitan el ajuste flexible de los vectores representativos de cada grupo y que hagan que el algoritmo "aprenda" progresivamente de las variaciones que una misma instrucción exhibe al ser repetida varias veces.

3. Para continuar con el desarrollo de la investigación en este campo, con base en los resultados de este estudio, y con la forma de enfocar el problema de reconocimiento en dos fases complementarias (aprendizaje y reconocimiento), se sugiere en la segunda fase, el diseño de una medida de discriminación flexible, que permita la mejor definición de pertenencia de una instrucción a cierto grupo previamente delimitado, para lo cual se recomienda la aplicación de algoritmos de redes neuronales de competencia.

4. Para investigaciones posteriores, se recomienda explorar la aplicación de técnicas combinadas de reconocimiento, como las descritas en el marco teórico de este trabajo.

CAPITULO VI

BIBLIOGRAFÍA

1. Chávez, J.J. Elaboración de Proyectos de Investigación. 2ª. Edición. XL Publicaciones. Guatemala, 1994. Pp. 75.

2. Dallas, J. Métodos Multivariados Aplicados al Análisis de Datos. Internacional Thomson Editores, S.A. de C.V. México, 2000. Pp.566.

3. Deller, J., J.H. Hansen, J. Proakis. *DiscreteTime Processing of Speech Signals*. IEEE Press. E.E.U.U., 2000. Pp. 908.

4. Fallside, F., W. Woods (Editores). *Computer Speech Processing*. Prentice Hall. E.E.U.U., 1985. Pp. 485.

5. Fausett, Laurene. *Fundamentals of Neural Networks*. Prentice Hall. E.E.U.U., 1994. Pp.461.

6. Kendall, M. *Multivariate Analysis*. 2a. Edición. Charles Griffin & Company LTD. Gran Bretaña, 1980. Pp. 180.

7. Lyons, R. *Understanding Digital Signal Processing*. Addison Wesley. E.E.U.U., 1997. Pp. 517.

8. Rouviere, H. Compendio de Anatomía y Disección. Tercera Edición Española. Salvat Editores, S.A. España, 1971. Pp. 857.

9. Soliman, S., M. Srinath. *Continuous and Discrete Signals and Systems*. 2a. Edición. Prentice Hall. E. E.U.U., 1998. Pp. 525.

10. Yamane, Taro. *Statistics, an Introductory Analysis*. Harper & Row Publishers. E.E.U.U., 1964. Pp. 734.

CAPITULO VII

ANEXOS

A. Modelo de lenguaje de Peirce según John Deller (2000):

Este modelo identifica cuatro tipos de componentes del código de lenguaje natural: componentes simbólicos, gramaticales, semánticos y pragmáticos.

1. Componente simbólico (símbolos):

Constituyen las unidades fundamentales del idioma, y que por ende componen todos los mensajes. En la forma hablada del idioma, los componentes simbólicos pueden ser las palabras, mientras que en la forma escrita pueden ser las letras del abecedario.

2. Componente gramatical:

Sc ocupa de la manera en que se relacionan los símbolos entre sí. Se incluyen restricciones de léxico y de sintaxis. Las primeras gobiernan la manera en que los fonemas forman palabras, mientras que las segundas se ocupan de cómo estas palabras forman oraciones concretas.

3. Componente semántico:

Técnicamente, la gramática (o componente gramatical) de un idioma es completamente ar*bit*raria, en el sentido que cualquier regla de combinación de símbolos es permisible. Es decir, una oración puede estar gramaticalmente correcta, pero carecer totalmente de sentido práctico. El componente semántico del idioma permite resolver esta situación, regula la forma en la que se combinan los símbolos para formar comunicaciones con significado lógico.

4. Componentes pragmáticos:

Este componente se ocupa de la relación que guardan los símbolos con sus usuarios y el entorno de la comunicación. Permiten identificar expresiones fonéticamente

iguales con situaciones diferentes, según el contexto en el cual se emplean. Si una fuente de conocimiento pragmático se incorporara en un sistema de reconocimiento de voz, éste debería ser capaz de discernir entre diferentes significados de cadenas de símbolos y decodificarlos correctamente.

Las restricciones lingüísticas se incorporan a un sistema de reconocimiento de voz a través de "fuentes de conocimiento", cada una de las cuales se asocia con un componente del modelo de lenguaje de Peirce (ver figura No.10). Las "fuentes de conocimiento" constituyen bancos implícitos o explícitos de conocimientos lingüísticos, los cuales se incorporan al sistema. El primer sistema en incorporar exitosamente restricciones lingüísticas fue el sistema HARPY desarrollado en Carnegie-Mellon University como parte del proyecto de entendimiento de habla de ARPA (1980).

El diagrama de bloques mostrado en la figura No.10 abarca la mayoría de sistemas de reconocimiento de voz. Sin embargo, el campo de reconocimiento de voz es muy variado, por lo que es posible que ciertos sistemas escapen a este modelo. La figura también evidencia la clasificación de reconocedores de voz existentes en base al tipo de procesamiento empleado (esta clasificación es independiente de la mencionada al inicio de esta discusión):

1. Modo *bottom-up*: Utiliza el procesamiento acústico para hipotetizar diferentes fonemas, palabras, etc. las que son procesadas "hacia arriba" en el modelo presentado en la figura 10. Esta categoría incluye la mayoría de sistemas, como el HARPY.
2. Modo *top-down*: Este tipo de procesamiento comienza con una oración hipotética propuesta en el nivel superior de procesamiento (parte superior del diagrama de bloques). Esta hipótesis es luego escudriñada en cada uno de los niveles inferiores y determina si representa en un grado aceptable la onda acústica recibida. Cada nivel solicita información al inmediato inferior para evaluar el grado de similitud en su nivel, hasta llegar al procesador acústico que determina si los símbolos hipotéticos

son consistentes con los resultados del análisis acústico de la señal de ingreso. Es claro que este tipo de procesamiento es más complicado y requiere mucho más cómputo que el anterior.

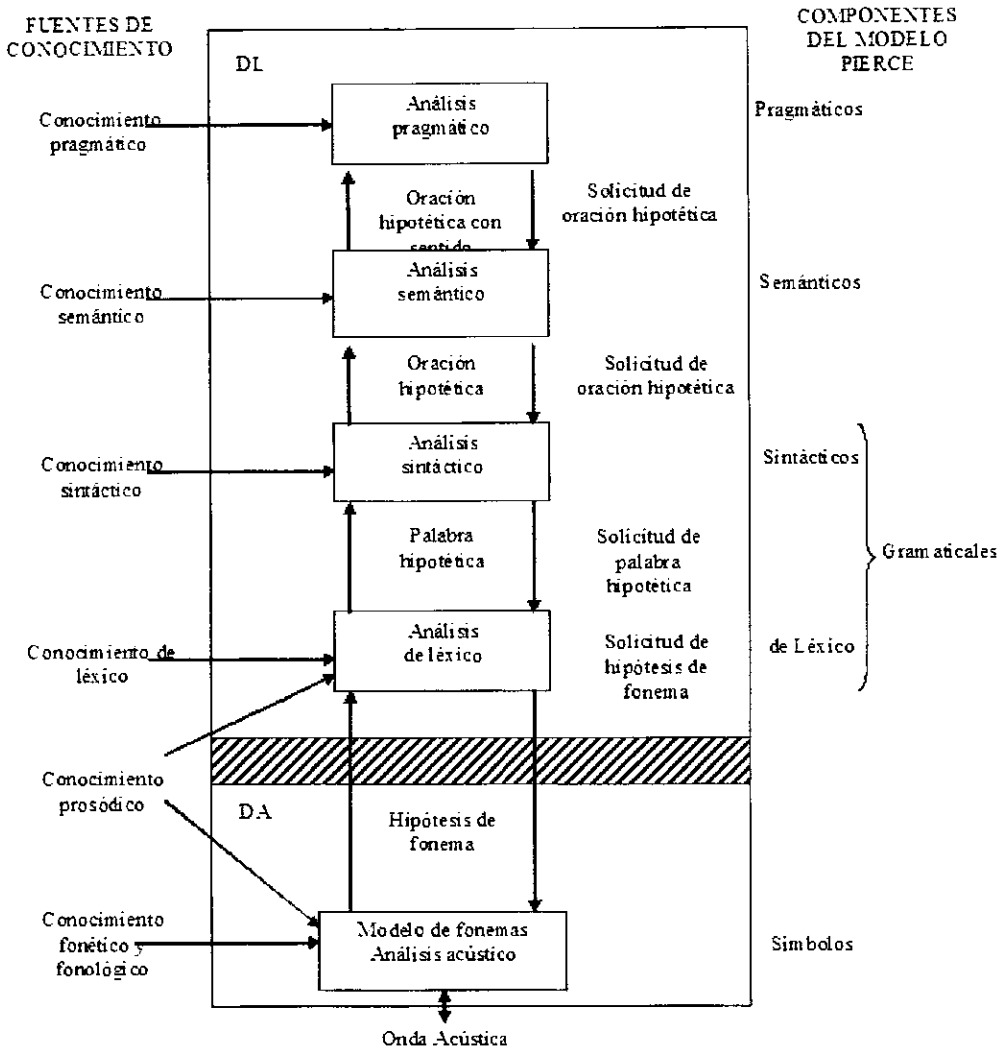


Figura No. 10 Diagrama de bloques de un sistema de reconocimiento de voz general que muestra los procesadores acústicos (DA) y lingüísticos (DL). Reproducido de: John Deller, *Discrete-Time Processing of Speech Signals*, IEEE Press. E.E.U.U., 2000. Pág. 618.

B. Código de la aplicación (Borland Delphi 6.0):

```
unit Unit1;

interface

uses
  Windows, Messages, SysUtils, {Variants,} Classes, Graphics, Controls, Forms,
  Dialogs, StdCtrls, mmsystem, ComCtrls, TeEngine, Series, ExtCtrls,
  TeeProcs, Chart, math, Buttons, VaClasses, VaComm;

type
  TForm1 = class(TForm)
    Chart2: TChart;
    Series2: TFastLineSeries;
    GroupBox1: TGroupBox;
    Label6: TLabel;
    GroupBox2: TGroupBox;
    Label3: TLabel;
    ComboBox1: TComboBox;
    Label2: TLabel;
    ComboBox2: TComboBox;
    Label5: TLabel;
    Edit5: TEdit;
    GroupBox3: TGroupBox;
    Button1: TButton;
    Button2: TButton;
    Label1: TLabel;
    Edit1: TEdit;
    GroupBox4: TGroupBox;
    BitBtn1: TBitBtn;
    BitBtn2: TBitBtn;
    VaComm1: TVaComm;
    GroupBox5: TGroupBox;
    BitBtn3: TBitBtn;
    BitBtn4: TBitBtn;
    ListBox1: TListBox;
    GroupBox6: TGroupBox;
    GroupBox7: TGroupBox;
    Label4: TLabel;
    Label7: TLabel;
    ListBox2: TListBox;
    procedure Button1Click(Sender: TObject);
    procedure Button2Click(Sender: TObject);
    procedure BitBtn1Click(Sender: TObject);
    procedure BitBtn2Click(Sender: TObject);
    procedure BitBtn3Click(Sender: TObject);
    procedure BitBtn4Click(Sender: TObject);
    procedure vacomm1rxchar(sender:tobject; count:integer);


  private
    { Private declarations }
    procedure BufferLleno;
    procedure MMInDone(var msg: Tmessage); message MM_WIM_DATA;
  public
    { Public declarations }
  end;

var
  Form1: TForm1;
    resultReal : array[0..4096] of double;
    resultImag : array[0..4096] of double;
```

```
    matrizotac : integer;
    matrizota : array[0..54783]of array[0..1]of double;
    SAMPLE:array[0..8191]of double;     //ARREGLO QUE ALMACENA MUESTRAS (AMPLITUD)
    tempvar1:integer;
    BUFFERSIZE:integer;
    lflag:integer;
    lflag1:integer;
    lflag2:integer;
    inicio:integer;
    MCount:integer;
    buffercount:integer;
    temp1:array[0..4095] of integer;   //ALMACENAMIENTO DE DATOS ENVIADOS A PUERTO SERIAL


implementation

{$R *.dfm}

//INICIALIZACION DE VARIABLES
var
    MWaveInHandle : PHWaveIn;
    MWaveFormat : PWaveFormatEX;
    MWaveBuffer : PWaveHdr;
    MWaveBuffer2: PWaveHdr;

    MBuffer : PChar;
    MBuffer2: Pchar;

    parar : boolean;
    i:integer;


procedure TForm1.Button1Click(Sender: TObject);
var MError : integer;
    msg : PChar;
begin
    BUFFERSIZE:= strtoint(combobox1.text);
    listbox2.Clear;
    matrizotac:=0;//mio
    lflag:=0;
    lflag1:=0;                          //VARIABLES DE CONTROL PARA CICLO DE BUFFER CIRCULAR
    lflag2:=0;
    MCount:=0;
    new (MWaveFormat);
    new (MWaveBuffer);
    new (MWaveInHandle);
    new (MWaveBuffer2);
    label6.Caption:='';

    GetMem(MBuffer, BUFFERSIZE);        //UBICACION DE MEMORIA DEL BUFFER1
    getmem(mbuffer2,BUFFERSIZE);        //UBICACION DE MEMORIA DEL BUFFER2

    MWaveFormat.wFormatTag := 1;        //CONFIGURACION DE TARJETA DE SONIDO
    MWaveFormat.nChannels := 1;
    MWaveFormat.nSamplesPerSec := strtoint(combobox2.text);//11000;
    MWaveFormat.wBitsPerSample := 8;
    MWaveFormat.nBlockAlign := 8;
    MWaveFormat.nAvgBytesPerSec := strtoint(combobox2.text);//11000;
    MWaveFormat.cbSize := 0;

    MError  :=  waveInOpen(MWaveInHandle,  0,  MWaveFormat,  form1.Handle,  0,  CALL-
BACK_WINDOW);
    if MError <> 0 then
```

```
    begin
        GetMem(msg, 100);
        waveInGetErrorText(MError, msg, 100);
        ShowMessage (string(msg));
        FreeMem(msg);
    end;

    MWaveBuffer.lpData := MBuffer;      //ASIGNACION DE BUFFERS A TARJETA DE SONIDO
    MWaveBuffer.dwBufferLength := BUFFERSIZE;
    MWaveBuffer.dwFlags := 0;
    MWaveBuffer.dwLoops := 0;
    MWaveBuffer2.lpData := MBuffer2;
    MWaveBuffer2.dwBufferLength := BUFFERSIZE;
    MWaveBuffer2.dwFlags := 0;
    MWaveBuffer2.dwLoops := 0;

    //PREPARACION DE LOS BUFFERS
    MError       :=       waveInPrepareHeader(Integer(MWaveInHandle^),       MWaveBuffer,
sizeof(MWaveBuffer^));
    if MError <> 0 then
    begin
        GetMem(msg, 100);
        waveInGetErrorText(MError, msg, 100);
        ShowMessage (string(msg));
        FreeMem(msg);
    end;

    MError        :=        waveInAddBuffer(integer(MWaveInHandle^),        MWaveBuffer,
sizeof(MWaveBuffer^));
    if MError <> 0 then
    begin
        GetMem(msg, 100);
        waveInGetErrorText(MError, msg, 100);
        ShowMessage (string(msg));
        FreeMem(msg);
    end;
    MError       :=       waveInPrepareHeader(Integer(MWaveInHandle^),       MWaveBuffer2,
sizeof(MWaveBuffer2^));
    if MError <> 0 then
    begin
        GetMem(msg, 100);
        waveInGetErrorText(MError, msg, 100);
        ShowMessage (string(msg));
        FreeMem(msg);
    end;
    MError        :=        waveInAddBuffer(integer(MWaveInHandle^),        MWaveBuffer2,
sizeof(MWaveBuffer2^));
        if MError <> 0 then
        begin
            GetMem(msg, 100);
            waveInGetErrorText(MError, msg, 100);
            ShowMessage ('interno: ' + string(msg));
            FreeMem(msg);
        end;

end;


procedure TForm1.Button2Click(Sender: TObject);
begin
    waveInStart (Integer(MWaveInHandle^));
end;
```

```
procedure TForm1.BufferLleno;
begin
end;

//FUNCION QUE SE EJECUTA AL TERMINAR DE LLENAR UNO DE LOS BUFFERS.
procedure TForm1.MMInDone(var msg: Tmessage);
var MError : integer;
 msg2 : PChar;
 Header : PWaveHdr;
 k:integer;
// myvar:Pchar;
begin
if lflag<>2 then
begin
                k:=0;
                i:=0;
                Header := PWaveHdr(msg.LParam);//CASTING, CAMBIO DEL TIPO DE VARIABLE DE
ENTERO A UN PUNTERO DE TIPO PWAVEHDR
                                                //PARA TENER ACCESO A TODA LA INFO QUE
PROPORCIONA
                listbox2.Items.Add(inttostr((integer(Header.lpdata))));

                if Mcount=0 then
                begin
                while (i<BUFFERSIZE) do
                        begin
                        if byte(header.lpData[i])-128>=strtoint(edit5.Text) then
                                begin
                                for k:=i to 8191 do
                                        begin
                                        SAMPLE[k-i]:=byte(header.lpdata[k])-128;
                                        Mcount:=Mcount+1;        //INDICADOR DE NUMERO DE
POSICIONES ALMACENADAS.
                                        end;
                                        i:=Buffersize;          //SALIDA DEL WHILE
                                end;
                        i:=i+1;

                        end;
                //SE PREPARA SIGUIENTE BUFFER A SER LLENADO
                 MError      :=      waveInAddBuffer(integer(MWaveInHandle^),      header,
sizeof(MWaveBuffer^));
                 if MError <> 0 then
                 begin
                 GetMem(msg2, 100);
                 waveInGetErrorText(MError, msg2, 100);
                 ShowMessage ('interno: ' + string(msg2));
                 FreeMem(msg2);

                 end;

                end
                else
                begin
                lflag:=2;
                for i:=Mcount to 8191 do
                        begin
                        SAMPLE[i]:=byte(header.lpdata[i-Mcount])-128;
                        end;
                        //SI YA TERMINO DE MUESTREAR....
                label6.Caption:='FIN CAPTURA';
                waveInStop (Integer(MWaveInHandle^));
```

```
                lflag1:=2;                                          //PARA QUE IGNORE SI-
GUIENTE EJECUCION DE LA FUNCION
                series2.clear;
                for i:= 0 to BUFFERSIZE - 1 do                      //ESCRITURA A ARCHIVO.
                        begin
                        series2.AddXY(i,SAMPLE[i]);
                        end;
                //TERMINA DE USAR EL DISPOSITIVO, LIBERA MEMORIA
                waveInClose(Integer(MWaveInHandle^));
                waveInUnprepareHeader       (Integer(MWaveInHandle^),       MWaveBuffer,
sizeof(MwaveBuffer^));
                FreeMem(MBuffer);
                waveInUnprepareHeader       (Integer(MWaveInHandle^),       MWaveBuffer2,
sizeof(MwaveBuffer2^));
                FreeMem(MBuffer2);
                Dispose (MWaveFormat);
                Dispose (MWaveBuffer);
                Dispose (MWaveInHandle);
                Dispose (MWaveBuffer2);
                end;

     end;

end;


procedure TForm1.BitBtn1Click(Sender: TObject);
var
data:textfile;
i_cont:integer;

begin
    assignfile(data,edit1.Text);
    rewrite(data);

        for i_cont:= 0 to strtoint(combobox1.text) - 1 do//ESCRITURA A ARCHIVO.
                begin
                write(data,SAMPLE[i_cont]);
                write(data,chr(13));
                write(data,chr(10));
                end;
        closefile(data);
end;

procedure TForm1.BitBtn2Click(Sender: TObject);
var
data:textfile;
i_cont:integer;

begin
    series2.clear;
    assignfile(data,edit1.Text);
    reset(data);

        for i_cont:= 0 to strtoint(combobox1.text) - 1 do//ESCRITURA A ARCHIVO.
                begin
                read(data,SAMPLE[i_cont]);
                series2.AddXY(i_cont,SAMPLE[i_cont]);
                end;
        closefile(data);

end;
```

```
//DE AQUI PARA ABAJO, TODO LO QUE TIENE QUE VER CON SERIAL
procedure TForm1.VaComm1RxChar(Sender: TObject; Count: Integer);
begin
vacomm1.readbuf(temp1[i_cont],1);   //RECEPCION DE DATOS VIA SERIAL
i_cont:=i_cont+1;
vacomm1.ReadChar(temp1);
temp2:=temp1;
end;


procedure TForm1.BitBtn3Click(Sender: TObject);
var
        i:integer;
        arreglo1:array[0..4095] of integer;
        test:integer;
begin

        for i:=0 to 4095 do
        begin
        test:=vacomm1.WriteBuf(SAMPLE[i],1);//ENVIO DE DATOS VIA SERIAL
        if test=0 then
                label1.caption:='error';
        end;

end;


end.
```

C. Código de *macros* utilizados para el análisis de datos (Microsoft Excel):

```
Private Sub CommandButton1_Click()
componer
End Sub

Private Sub CommandButton2_Click()
fourier ("com5")  'Lo escrito en comillas, depende de la hoja electrónica donde se
End Sub           'encuentra el comando analizado.

Private Sub CommandButton3_Click()
dcoff
End Sub

'-------------------------------------------------
Public Sub componer()

cond = ActiveSheet.Cells(15, 8).Value



For j = 11 To 15
'ini = ActiveSheet.Cells(j - 3, 8).Value
i = 2
While i < 16385
If ActiveSheet.Cells(i, j - 9) >= cond Then
    ini = i
    i = 16385 'para que se salga del ciclo
End If
i = i + 1
Wend

For i = 1 To 8192

ActiveSheet.Cells(i + 1, j) = ActiveSheet.Cells(i + ini - 51, j - 9) 'asumo quel a pala-
bra comienza 30 muestras antes.
Next i
Next j


For j = 17 To 21
k = j - 6
For i = 1 To 4096
ActiveSheet.Cells(i + 1, j) = ActiveSheet.Cells(i * 2 + 1, k)
Next i
'For i = 1 To 4096
'ActiveSheet.Cells(i + 1, j + 6).Clear
'Next i

Next j

End Sub

Public Sub dcoff()
For j = 17 To 21
For i = 1 To 4096
temp = ActiveSheet.Cells(i + 1, j)
ActiveSheet.Cells(i + 1, j) = temp + ActiveSheet.Cells(j - 9, 10)

Next i
```

```
Next j
End Sub

Public Sub fourier(temp1)
For j = 17 To 21
For i = 1 To 4096
ActiveSheet.Cells(i + 1, j + 6).Clear
Next i
Next j


For i = 1 To 5
    temp = temp1 & i
    Application.Run "ATPVBAEN.XLA!Fourier", ActiveSheet.Range(temp), ActiveSheet.Cells(2,
22 + i), False, False
Next
End Sub
```