Diseño de un sistema de reconocimiento del habla para controlar dispositivos eléctricos

■ Ing. D. K. Salcedo
dayanasa@gmail.com

■ Ing. A. P. Teixeira
apteixeira@gmail.com

■ Tutor Ing. María Stefanelli
mstefan@usb.ve.com

Resumen

En este trabajo se plantea el reto de desarrollar un sistema que, usando el reconocimiento del habla como herramienta, permita el control de dispositivos eléctricos de una forma sencilla y sin ambigüedades. El desarrollo de este proyecto se basa en un estudio teórico sobre la naturaleza, parametrización, características y procesamiento de señales de voz, además de un estudio práctico que contempla simulaciones en MATLAB de distintas señales de voz, para analizar sus semejanzas y diferencias de acuerdo a la persona que las genera. Se realizó la evaluación teórica y práctica del desempeño de las técnicas de parametrización de la señal de voz *Mel Frequency Cepstrum Coefficients* (MFCC), *Linear Frequency Cepstrum Coefficients* (LFCC) y *Linear Predictive Cepstrum Coefficients* (LPCC) dentro del esquema del reconocimiento del habla, la programación de un microcontrolador como elemento de procesamiento dentro del hardware encargado de controlar los dispositivos eléctricos, para interpretar los comandos de la aplicación y convertirlos en acciones de encendido y apagado de los mismos; además se recolectaron un total de 300 señales de voz entre 20 participantes, 200 de ellas para la creación de una base de datos de entrenamiento y 100 de ellas para la base de datos de pruebas del sistema. Entre los resultados más resaltantes se destacan la escogencia del MFCC como técnica de parametrización por su alto porcentaje de acierto en el reconocimiento del habla, además del "Sistema de reconocimiento del habla para controlar dispositivos eléctricos" completamente funcional.

Palabras clave: procesamiento digital de la voz, parametrización de la voz, reconocimiento del habla

Abstract *The development of this project is based on a theoretical study about the nature, parametric representation, characteristics and signal voice processing, in addition to a practical study that it contemplates simulations in MATLAB with different signals voices in order to analyze its similarities and differences according to the person who generates them. The theoretical and practical evaluation of the performance of the parametric representation techniques was made, including in this study the Mel Frequency Cepstrum Coefficients (MFCC), Linear Frequency Cepstrum Coefficients (LFCC) y Linear Predictive Cepstrum Coefficients (LPCC) within the scheme of the speech recognition. In addition, this study contemplates the programming of a microcontroller as a processing element within the hardware, in charge of controlling the electrical devices in order to interpret the commands of the application and to turn them into actions (On/Off). A total of 300 signals of voice between 20 participants were collected, 200 of them for the creation of a training data base and 100 of them for the testing data base of the system. Between the most important results that stand out, is the chose of MFCC as the parametric representation technique used in the application, because of its high percentage of success in the speech recognition.*

Introducción

El reconocimiento del habla tiene como objetivo permitir la comunicación hablada entre el ser humano y el computador, para ejecutar acciones determinadas mediante comandos de voz. El principal reto que se plantea en un sistema de reconocimiento del habla es el de interpretar un conjunto de informaciones que proceden de diversas fuentes de conocimiento (acústica, fonética, fonológica, léxica, sintáctica, semántica y pragmática), en presencia de ambigüedades e incertidumbres para obtener una interpretación aceptable del mensaje acústico recibido. La finalidad de uso del reconocimiento del habla en el sistema desarrollado en este trabajo es la del control de dispositivos eléctricos utilizando comandos de voz.

El principal objetivo de este proyecto es el diseño de un sistema que, utilizando el reconocimiento del habla como herramienta, permita controlar dispositivos eléctricos usualmente encontrados en un ambiente doméstico y hacer de éste un espacio cómodo, flexible y moderno. Además, este estudio pretende traer beneficios adicionales para las personas discapacitadas, las cuales requieren de una ayuda extra para realizar tareas cotidianas como encender una luz, mover una cama clínica o apagar la televisión.

Para la implementación del sistema que se muestra en la figura 1 se desarrolló una aplicación encargada del reconocimiento del habla, basada en los resultados obtenidos del estudio de la voz, sus características y técnicas de parametrización. Además, se implementó el hardware encargado del encendido y apagado de los dispositivos eléctricos conectados a él, de acuerdo al comando de voz procesado por el sistema de reconocimiento del habla. Los dispositivos que el sistema puede manejar son aquellos cuyo encendido y apagado no depende de un circuito electrónico y que al activarse un relé que permite o impida el paso de corriente hacia ellos, se enciendan o apaguen. Algunos ejemplos de este tipo de dispositivos son: televisores y lámparas con perilla de encendido/apagado, cargador del celular y cargador del laptop.

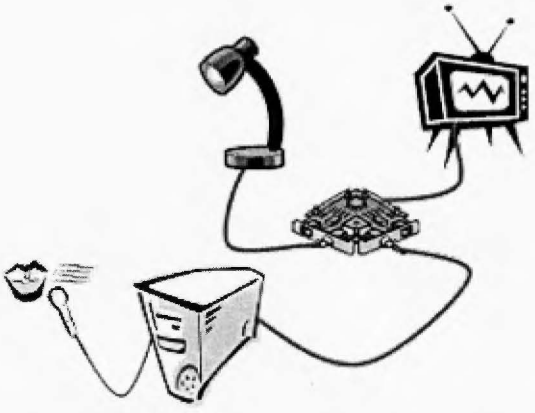


Figura 1. Componentes del Sistema de Reconocimiento del habla para el control de dispositivos eléctricos

La aplicación encargada de realizar el reconocimiento del habla está diseñada para el idioma español y para reconocer la palabra clave que constituye el comando de voz, no quién la dice (hombre, mujer, niño o niña). El sistema descrito maneja cinco comandos de voz correspondientes a las vocales a, e, i, o, u. Además, contempla una única implementación de hardware, el cual se conecta al

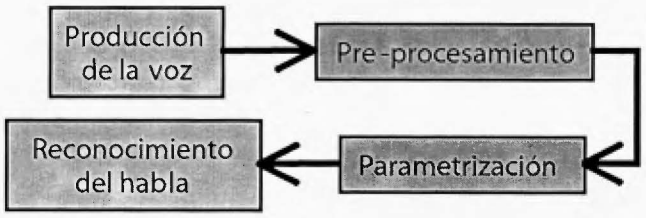
computador mediante el puerto serial, utilizando la interfaz EIA232 y es el encargado del encendido y apagado de los dispositivos eléctricos.

En este artículo se presenta, de una forma estructurada por secciones, el desarrollo del "Sistema de reconocimiento del habla para controlar dispositivos eléctricos". En la sección 1 se hace un resumen del proyecto, la sección 2 habla sobre el planteamiento del proyecto, sus objetivos, alcance y limitaciones. Por su parte, en la sección 3 se hace un breve resumen de la investigación teórica realizada para el desarrollo del proyecto; seguido por la sección 4 en la que se describen las fases en la metodología seguidas para la implementación del sistema. En la sección 5 se muestran los resultados obtenidos de cada una de las fases descritas en la sección 4 y, por último, la sección 6 se refiere a las conclusiones y recomendaciones que se derivan del diseño e implementación del sistema en cuestión.

Marco referencial

Para la realización de cualquier tarea que involucre el procesamiento de la señal de voz es fundamental el conocimiento de ciertos aspectos que ayudan a crear las bases necesarias para la comprensión de las características de la voz y del tratamiento que se le debe dar de acuerdo al fin deseado. Para la aplicación particular de reconocimiento del habla, el esquema de procesamiento que se sigue es similar al que se muestra en la figura 2.

Figura 2. Esquema de procesamiento de la voz para el reconocimiento del habla



Reyes y Herrera (2005) indican que la voz se obtiene por la acción conjunta de varias regiones de órganos: el *tracto pulmonar*, el cual controla la amplitud de los sonidos, *la laringe*, en la que se sitúan las cuerdas vocales, las cuales funcionan como un control de entrada a los sonidos y cuya tensión afecta la frecuencia de la señal de voz, por último se encuentra *el tracto vocal*, que se encarga de la articulación de la voz. Dependiendo del comportamiento de la cavidad bucal y de la configuración de los órganos, la señal que se genera se clasifica en dos tipos:

- *Señal sonora:* se produce por la vibración de las cuerdas vocales, las cuales se abren y cierran modificando el área de la tráquea y produciendo un tren de impulsos cuasi-periódicos, con un período o frecuencia fundamental llamado *pitch.*
- *Señal no sonora*: en su generación el aire fluye libremente hasta alcanzar el tracto vocal al permanecer abiertas las cuerdas vocales, por lo que la señal generada se constituye de una contribución desordenada de componentes frecuenciales y se caracteriza por presentar aleatoriedad similar a la del ruido blanco.

Para las señales sonoras, el tracto vocal actúa como una cavidad resonante, estando centradas las frecuencias de resonancia generalmente en 500 Hz y sus armónicos pares. Esta resonancia produce grandes picos en el espectro resultante, a los cuales se les llama *formantes*. Por su parte, para las señales no sonoras, el tracto vocal presenta una estructura ruidosa y aleatoria tanto en el dominio del tiempo como en el de la frecuencia, por lo que no se tienen formantes.

La Transformada de Fourier es una herramienta muy importante para el análisis de las señales de voz en el dominio de la frecuencia, ya que permite obtener información de las mismas que no es evidente en el dominio del tiempo. Una de las características de la Transformada de Fourier es que permite extraer el contenido frecuencial de la señal, lo cual es muy importante para establecer el muestreo adecuado para la misma, el cual deber ser por lo menos 2 veces mayor que la frecuencia máxima de la señal.

Pre-procesamiento

La etapa de pre-procesamiento de la señal de voz corresponde a los pasos previos a la parametrización, necesarios para resaltar las características más importantes de la señal de voz y luego poder

analizarlas con la técnica de parametrización escogida para tal fin. La figura 10 muestra un diagrama de bloques del pre-procesamiento aplicado a una señal de voz antes del paso correspondiente a la parametrización.

Figura 3. Esquema de pre-procesamiento de la señal de voz

- *Preénfasis:* después de que la señal de entrada se tiene digitalizada, la etapa de *preénfasis* se realiza para hacer el procesamiento de la señal menos susceptible a truncamientos, aplanarla espectralmente y para compensar la caída de 6 dB que experimenta la señal al pasar a través del tracto vocal *(Kornhauser, 1999).*
- *Endpoint detection:* Una vez que la señal de voz se tiene grabada en el computador, es importante determinar el comienzo y final de la parte útil de la señal, para lo cual se utiliza la técnica denominada *endpoint detection*, que mediante un algoritmo busca los puntos de la señal donde se concentra la energía y extrae sólo ese fragmento (Slavinsky, 1999).
- *Segmentación y aplicación de la ventana:* después de que se tiene la información útil de la señal de voz, se hace necesario dividir la señal en tramas de N muestras, donde N es un valor que se escoge tomando en cuenta que la señal de voz es estacionaria a "trozos" (condición necesaria para poder realizar el análisis de Fourier en tiempo corto). El intervalo de tiempo en el que la señal se considera estacionaria comúnmente se establece en un valor entre 20 y 40 ms. En la figura 4 se muestra un ejemplo de segmentación. Por lo general se utilizan tramas de 256 muestras por ser un valor que establece un equilibrio entre la resolución en tiempo y en frecuencia para una señal de voz.

La siguiente fase de la etapa de pre-procesamiento corresponde a la escogencia del tipo de ventana que se le aplica a cada trama. Dicha escogencia es muy importante para analizar el efecto de la misma sobre la resolución espectral de la señal. Otro aspecto que es importante analizar es el solapamiento entre tramas consecutivas, el cual para la figura 4 es de 156 muestras y se realiza con dos propósitos principales: el primero consiste en que las mismas guarden relación entre sí para que el análisis de cada trama no sea aislado y el segundo consiste en compensar la caída que tienen comúnmente las ventanas en los bordes de la trama.

Técnicas de parametrización de la señal de voz

La selección de la mejor representación paramétrica de la señal de voz es una tarea importante en el diseño de cualquier sistema de reconocimiento del habla. Los objetivos principales de esta representación son los de comprimir los datos correspondientes a la señal de voz, eliminando información no pertinente al análisis fonético de la información y mejorar esas características de la señal de voz que contribuyen significativamente con la detección de las diferencias fonéticas y que no son apreciables mediante un simple análisis en tiempo o en frecuencia, sino usando un análisis más exhaustivo como el Cepstrum.

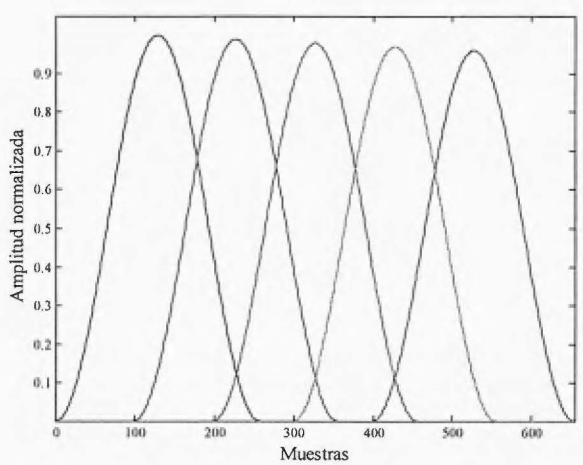


Figura 4. Segmentación de la señal de voz en tramas solapadas entre ellas

Cepstrum

Es una herramienta muy utilizada para la representación paramétrica de las señales de voz y se define como la Transformada de Fourier del espectro logarítmico de la señal, por lo que existe un cepstrum complejo y un cepstrum real dependiendo de si la función logarítmica está definida para valores reales o complejos. La diferencia entre uno y otro radica en el hecho de que el cepstrum complejo permite reconstruir la señal y el real no, ya que se pierde la información correspondiente a la fase. El cepstrum real se utiliza en muchas aplicaciones y como no depende de la fase de la transformada de la señal es mucho más fácil de calcular que el cepstrum complejo, aunque la señal original no pueda ser recuperada (Oppenheim & Schafer, 1989).

El análisis cepstral es comúnmente utilizado para obtener información de la señal de voz que permita parametrizarla para luego ser usada en la fase de reconocimiento. Las representaciones paramétricas más utilizadas para trabajar con señales de voz y que están basadas en el análisis cepstral de la misma, pueden ser divididas en dos grupos: aquellas basadas en la predicción lineal del espectro y aquellas basadas en el espectro de Fourier.

Una de las técnicas más usadas en el procesamiento de señales de voz es el análisis de predicción lineal. Esta técnica ha probado ser muy eficiente debido a la posibilidad de parametrizar la señal con un número pequeño de patrones con los cuales es posible, entre otras cosas, reconstruirla adecuadamente. Mediante esta técnica se puede representar a la señal vocal usando parámetros que varían en el tiempo y que están relacionados con la función de transferencia del tracto vocal y las características de la fuente sonora.

Predicción lineal de los coeficientes cepstrales: LPCC

La técnica LPCC permite estimar los coeficientes cepstrales mediante el uso del algoritmo LPC, el cual establece un modelo que permite predecir la próxima muestra de la señal empleando un filtro de puros polos, cuyos coeficientes varían en el tiempo. Dicho filtro se basa en el modelo de producción de la voz y su idea fundamental es que la voz puede modelarse a través de una combinación lineal de *p* muestras anteriores más una señal de excitación (periódica o ruido blanco dependiendo de la naturaleza de la señal).

Los coeficientes de predicción se usan en el proceso de parametrización para calcular los coeficientes cepstrales. La determinación de los coeficientes de predicción se realiza minimizando el error de predicción de orden *p* que se comete cuando se intenta realizar la aproximación de la señal. Estos coeficientes modelan el tracto vocal y pueden ser utilizados para el cálculo de los coeficientes cepstrales o LPCC's mediante la ecuación 1

$$c_i = a_i + \sum_{k=1}^{i-1} \left(\frac{k-i}{i} \right) \cdot c_{i-k} \cdot a_k$$

De esta manera se puede observar cómo la predicción lineal de los coeficientes del filtro que modela el tracto vocal puede ser utilizada para estimar los coeficientes cepstrales utilizados para el reconocimiento del habla.

Por su parte, para las técnicas de parametrización basadas en el espectro de Fourier, las características espectrales de la señal de voz se derivan del análisis de Fourier de tiempo corto. En la siguiente sección se muestran dos aplicaciones la técnica basada en el espectro de Fourier: MFCC y LFCC.

Coeficientes cepstrales de frecuencia Mel: MFCC

Los MFCC (Mel-Frequency Cepstrum Coefficients) son de gran utilidad en la extracción de los parámetros de la señal de voz, ya que están basados en la variación conocida de los anchos de banda de las frecuencias críticas del oído (Do, M). Los filtros que se le aplican a la señal en la técnica MFCC están espaciados linealmente para frecuencias menores a 1000 Hz y logarítmicamente para frecuencias mayores de 1000 Hz, con el fin de capturar las características fonéticamente importantes del habla. A esta escala se le denomina "Escala Mel" y su fórmula matemática se describe en la ecuación 2

$$Mel(f) = 2595 \cdot \log\left(1 + \frac{f}{700}\right)$$

Los pasos necesarios para el cálculo de los MFCC se muestran en la figura 5:

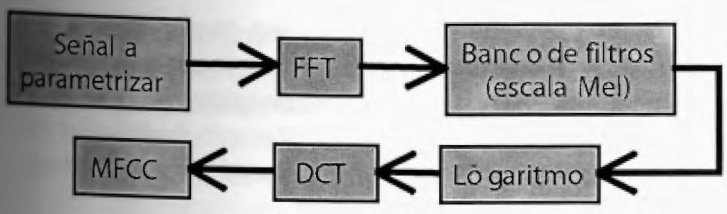


Figura 5. Diagrama de bloques para el cálculo de los MFCC's

El banco de filtros linealmente espaciado en la escala Mel puede tener distintas formas de ventana, tales como: Hamming, Hanning, Kaizer o triangular, siendo este último el más utilizado.

Mediante el proceso descrito anteriormente, para cada trama de voz de duración aproximada igual a 30 ms con solapamiento, se calcula un conjunto de coeficientes cepstrales. Este es el resultado de la transformada Discreta de Coseno de la Densidad Espectral de Potencia expresada en la escala Mel. A este conjunto de coeficientes se le denomina *Vector Acústico*, por lo que cada entrada es transformada mediante este proceso en una secuencia de vectores acústicos que representan las características más importantes de la voz, necesarias para el proceso de reconocimiento del habla.

Coeficientes cepstrales de frecuencia lineal: LFCC

El esquema bajo el cual funciona esta técnica es similar al mostrado en la figura 5, la única diferencia es que los filtros que conforman el banco de filtros que se le aplica a la señal, se encuentran sobre una escala lineal, es decir: $f(Hz) = f_{Mel}$.

Reconocimiento del habla

Las técnicas de parametrización explicadas en la sección 5 tienen como finalidad generar una serie de coeficientes que representan las características de la señal de voz, que pueden ser usadas en la fase de reconocimiento del habla y que no se obtienen mediante un análisis temporal o frecuencial. El tamaño de la matriz obtenida del proceso de parametrización depende directamente de la longitud (variable) de la señal de voz, por lo que se hace necesaria la estandarización de la matriz que contiene los coeficientes cepstrales calculados, para que el tamaño de las matrices usadas para el reconocimiento del habla sea el mismo. Este proceso se denomina *cuantización vectorial* y el paso siguiente corresponde al cálculo de la diferencia entre la señal de voz del hablante y las señales que se encuentran en la base de datos de entrenamiento del sistema mediante el cálculo de la *distancia euclidiana* en varias dimensiones.

Cuantización vectorial

La esencia de la cuantización vectorial, aplicada al caso particular del reconocimiento del habla, es la de obtener a partir de una matriz cualquiera de coeficientes cepstrales, una matriz de tamaño fijo que se parezca lo más posible a la original. Para ello, el espacio generado por los coeficientes cepstrales es dividido en un conjunto de regiones convexas mutuamente excluyentes y para cada una se calcula el *centroide,* que en dos dimensiones se representa como un punto que se encuentra a la menor distancia de todos los puntos (coeficientes cepstrales) que pertenecen a esa región. El conjunto de centroides obtenidos de la partición del espacio generado por los coeficientes cepstrales se denomina *codebook.* Por lo tanto, para caracterizar cada palabra se calcula su *codebook* correspondiente (Faúndez y Rodríguez, 2005).

Distancia euclidiana

Este método se emplea para calcular la diferencia existente entre el *codebook* obtenido de la palabra dicha por el hablante y el resto de *codebooks* almacenados en la base de datos de entrenamiento del sistema. El resultado final de dicha comparación es un valor numérico que representa la distancia entre dos matrices de iguales dimensiones y dependiendo de dicho valor se establece semejanza o diferencia entre las palabras

Metodología

Las fases en la metodología que se siguieron para el desarrollo e implementación del sistema de reconocimiento del habla para controlar dispositivos eléctricos se describen a continuación:

Recopilación de información y realización de pruebas con señales de voz

La búsqueda y recopilación de información realizada se basó en una consulta bibliográfica en sitios de Internet, así como en artículos de la IEEE *Transactions on Acoustics, Speech, and Signal Processing* y bibliografía especializada en el procesamiento de señales y sistemas como el libro Signals and Systems de Alan V. Oppenheim, entre otros.

Diseño de la aplicación encargada del reconocimiento del habla

El código desarrollado para la programación de la aplicación encargada del reconocimiento del habla está conformado por un bloque principal denominado "POcoNE", desde el cual se hacen llamadas a varios bloques secundarios conformados por funciones que realizan las tareas de pre-procesamiento, parametrización y reconocimiento de la señal de voz. La primera acción que se ejecuta en el bloque principal de la aplicación es el entrenamiento del sistema, leyendo todas las señales de voz guardadas en la base de datos de entrenamiento del sistema y aplicando los algoritmos correspondientes al pre-procesamiento y parametrización.

Pre-procesamiento

El diseño de las funciones que realizan las tareas de pre-procesamiento para cada una de las etapas que se muestran en la figura 3, se describen a continuación

- *Preénfasis:* se realizó usando una función denominada *prenfasis,* en la que se le aplica un filtro pasaalto a la señal de voz con la finalidad de prepararla para la detección de la información útil mediante el *endpoint detection.*
- *Endpoint detection:* la finalidad de esta etapa es la de eliminar los silencios usualmente presentes al comienzo y final de la señal de voz que graba el usuario. Se realizó mediante la función *energia* que divide la señal de voz en tramas (sin solapamiento), calcula la energía de cada una de las mismas y guarda esos valores en un vector *E*. Luego se establece un valor de comparación por encima del cual se considera que la energía es suficiente como para tomar la decisión de dónde empieza y termina la parte de la señal grabada que se considera información útil, extrayendo ese intervalo de la señal de voz original (sin preénfasis), para guardarlo como información útil.
- *Segmentación:* se realizó mediante la función *framing* que divide la señal en tramas de 256 muestras, que representan 25,6 mseg de voz, para considerar la señal de voz como estacionaria en ese intervalo de tiempo y crea una matriz *"M"* que tiene en cada una de sus columnas las muestras correspondientes a cada trama en la que se dividió la señal de voz. Para la segmentación se utilizó un solapamiento entre tramas consecutivas de 156 muestras.
- *Aplicación de la ventana Hamming:* una vez obtenida la matriz *M* de la etapa de segmentación, mediante la función *windowing* se creó una matriz *h* que en su diagonal contiene los valores de una ventana Hamming y que multiplicada por *M*, devuelve una matriz *"M2"* que corresponde a la ventana aplicada a cada una de las tramas contenidas en *M*.

Una vez que ya se tienen las tramas correspondientes a la señal de voz con la ventana aplicada a cada una de ellas, la parametrización es el paso siguiente en el esquema de reconocimiento del habla

Parametrización

El bloque correspondiente a la parametrización de la señal de voz, se basa en el desarrollo de tres algoritmos básicos que representan cada una de las técnicas de parametrización estudiadas, ya que esta fase corresponde a las pruebas que se realizaron con cada una de ellas y no a la escogencia de un método en particular. La explicación del funcionamiento de cada uno de los algoritmos desarrollados se muestra a continuación

- ✓ Coeficientes cesptrales de frecuencia Mel y lineal

- *Cálculo de la FFT:* mediante la función *fftframe* se calcula la Transformada de Fourier a

cada columna de la matriz *M2* (que corresponde a la señal de salida de la etapa de pre-procesamiento), obteniéndose la matriz *"fframe"* que corresponde a la FFT de cada trama. Esto se hace con el fin de aprovechar las características estacionarias de la voz en cortos intervalos de tiempo.

- *Creación del banco de filtros:* la creación del banco de filtros se realiza mediante la función *filterbank*, la cual tiene la opción de crear un banco de filtros separados linealmente en la escala Mel (MFCC) o separados linealmente en frecuencia (LFCC). Para el primer caso, se crea una escala de frecuencias Mel de acuerdo a la ecuación 2 y calculan cada una de las frecuencias en las cuales se centrarán los filtros del banco. Para el segundo caso, las frecuencias centrales de los filtros se calculan mediante una correspondencia directa y no logarítmica. Para ambos casos se utilizó un banco de 20 filtros, por ser un número que establece un equilibrio entre la rapidez del algoritmo y la calidad de la parametrización. Luego de que la escala de frecuencias está establecida, se crea el "banco de filtros triàngulares de escala Mel" o "banco de filtros triangulares de escala lineal" según sea el caso. La función descrita devuelve una matriz *"fb"* de 20 filas x 129 columnas, correspondiente a dicho banco.
- *Cálculo de la densidad espectral de potencia y aplicación del banco de filtros a la señal:* éste corresponde al siguiente paso en el cálculo de los MFCC o LFCC y se hace determinando la densidad espectral de potencia como el cuadrado de la magnitud de la matriz obtenida del cálculo de la FFT de *fframe*, y multiplicando el resultado de este cálculo por el banco de filtros *fb*, de lo cual se obtiene la matriz *"z"*. La razón para utilizar el cepstrum real como herramienta para el cálculo de los coeficientes cepstrales, radica en el hecho de que ésta representa una forma más fácil y rápida de parametrizar la señal de voz, ya que, aunque se pierde información sobre la fase de la señal y, a partir de estos coeficientes no se puede recuperar la señal de voz, se obtienen características de la misma que son especialmente útiles para las posteriores tareas de reconocimiento.
- *Cálculo de la transformada discreta de coseno (DCT):* este cálculo corresponde a pasar toda la información obtenida de la etapa previa al dominio del tiempo, lo cual es un paso fundamental para la obtención de los coeficientes cepstrales. En este caso se aplica la DCT, que representa la parte real de la Transformada Inversa de Fourier, ya que la matriz "z" es totalmente real en el dominio de la frecuencia y mediante la DCT se lleva al dominio del tiempo.

✓ Predicción Lineal de los Coeficientes Cepstrales: LPCC

El cálculo de los LPCC se realiza dentro de la función *parametrizar* en un bloque que se basa la ecuación 1, en la cual se puede observar que, partiendo del cálculo de los LPC de una señal, se pueden obtener sus LPCC. Usando la función *lpc* de MATLAB se calcularon los coeficientes de predicción lineal a la matriz *"M2"* y luego, a partir de esta matriz *"a"*, se realizó el cálculo de los LPCC, los cuales se guardaron en la matriz *"v"*.

Reconocimiento del habla

La etapa de reconocimiento del habla se realiza dentro de la función *test*, donde se compara la señal de voz dicha por el usuario y las señales pertenecientes a la base de datos de entrenamiento del sistema. Dicha función basa su funcionamiento en dos grandes bloques:

- *Cuantización vectorial:* este proceso se realiza mediante el uso de la función *VQLBG*. La matriz correspondiente al *codebook* de la palabra parametrizada tiene 20 filas y 16 columnas, éstas últimas correspondientes al número de centroides calculados por el LGB.
- *Distancia euclidiana:* este cálculo se realiza mediante el uso de la función *disteu,* la cual devuelve un valor numérico que representa la menor distancia entre la voz del hablante parametrizada y la señal de voz guardada en la base de datos de entrenamiento del sistema que más se parece a la primera.

Recolección de señales de voz

Para crear las bases de datos para el entrenamiento y pruebas del funcionamiento del sistema de reconocimiento se hizo con un total de 20 personas, de las cuales 11 eran hombres con edades comprendidas entre 20 y 24 años y 9 eran mujeres con edades comprendidas entre 20 y 22 años. La recolección de las voces se realizó en ambientes con diferentes niveles de ruido, por lo que se obtuvo una gran diversidad entre las señales de voz grabadas. Se grabaron las vocales desde la *a* hasta la u un total de tres veces: las primeras dos secuencias para agregarlas a la base de datos utilizada para el entrenamiento del sistema y la tercera secuencia para agregar a la base de datos usada para simular el funcionamiento "en vivo" del sistema de reconocimiento. Posteriormente se le aplicó a cada una de ellas el algoritmo correspondiente al *endpoint detection*, para guardar en las bases de datos sólo la información útil de la señal de voz.

Pruebas comparativas entre los métodos de parametrización estudiados

El criterio establecido para realizar la comparación entre los diversos métodos de parametrización desarrollados fue el porcentaje de error que arroja cada uno en cuanto a la exactitud del reconocimiento de los vocablos. Para realizar esta comparación se utilizaron las señales obtenidas en la fase de recolección de señales de voz para entrenar al sistema y para simular el funcionamiento "en vivo" del mismo.

Hardware encargado del control de los dispositivos eléctricos

Se siguieron varias fases de diseño, las cuales se describen a continuación.

Programación del microcontrolador

El papel del microcontrolador dentro del hardware es el de interpretar las señales provenientes del puerto serial del PC y transformarlas en salidas que el circuito de encendido y apagado pueda entender. El modelo del microcontrolador empleado fue el PIC 16F873, el cual posee un módulo de comunicación serial y varios puertos de entrada y salida, de los cuales sólo 5 de ellos fueron empleados para controlar el encendido y apagado de los dispositivos eléctricos conectados al hardware.

Circuito de encendido y apagado de los dispositivos eléctricos

Un circuito aparte del que se implementó para el microcontrolador fue necesario para generar el suministro de corriente necesario para la polarización de los relés encargados del encendido y apagado de los dispositivos. Dicho circuito está basado en el funcionamiento de un transistor como *suiche*, un diodo y un relé, tal y como se muestra en la figura 6

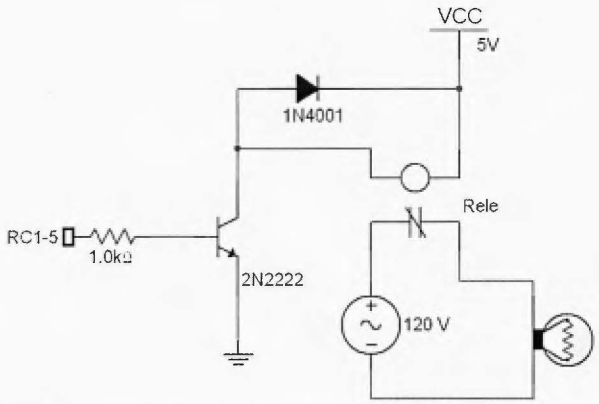


Figura 6. Diagrama circuital del módulo de encendido y apagado de los dispositivos eléctricos

El funcionamiento del circuito mostrado en la figura 6 es el siguiente: Se polariza la base del transistor con 5 voltios y el colector cae a tierra. Por su parte, la diferencia de potencial en el diodo es de 5 voltios (provenientes directamente de la fuente regulada) y la corriente suministrada por éste es suficiente para cambiar el estado del relé y, por ende, encender el dispositivo eléctrico.

Circuito impreso para el hardware de control de los dispositivos eléctricos

El montaje del hardware para el control de los dispositivos eléctricos se realizó en una placa de cobre, sobre la cual se diseñó un circuito impreso mediante el uso del software PCBExpress. Dicho circuito impreso incluye en su diseño el módulo de alimentación, además de todas las conexiones necesarias para la unión del circuito relacionado al microcontrolador y el circuito de encendido y apagado de los dispositivos eléctricos.

Resultados

Los resultados obtenidos del desarrollo del proyecto se describen a continuación.

Diseño en MATLAB de la aplicación encargada del reconocimiento del habla

La estructura de la aplicación encargada del reconocimiento del habla se basa en el esquema mostrado en la figura 2. Para ejemplificar los resultados obtenidos de las etapas que conforman el esquema mencionado, se utilizó la señal de voz que se muestra en la figura 7

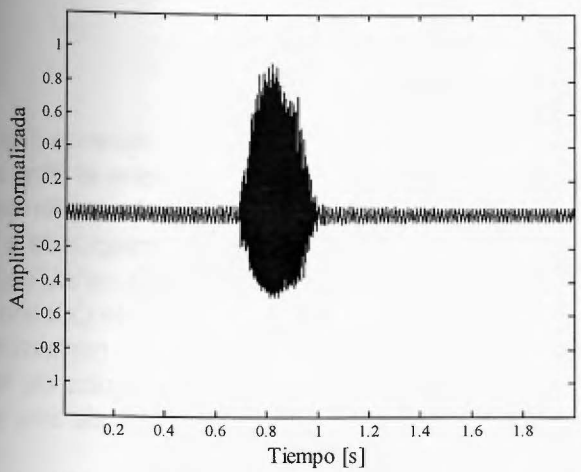


Figura 7. Representación en tiempo de la señal de voz correspondiente a la vocal "a"

Pre-procesamiento

La primera etapa del pre-procesamiento corresponde a la aplicación del filtro de *preénfasis* a la señal de voz. La finalidad de aplicar el filtro de preénfasis dentro del contexto de este proyecto es la de disminuir el nivel de ruido para que la detección del inicio y fin de la información útil *(endpoint detection)* sea más precisa. Una vez obtenidos el comienzo y final, la señal que se truncó fue la original (figura 7).

El paso siguiente a la obtención de la información útil de la señal es el de *segmentación y aplicación de la ventana*, paso que permite aprovechar que la señal es estacionaria en períodos cortos de tiempo y establecer un equilibrio entre la resolución temporal y frecuencial de la señal, tal y como se observa en la figura 8

Evaluación de las técnicas de parametrización estudiadas

A cada una de las tramas de voz obtenidas del pre-procesamiento, se le aplica el método de parametrización correspondiente. Los resultados que se muestran a continuación corresponden a pasar la trama que se muestra en la figura 7 por cada uno de los bloques correspondientes al esquema de parametrización de la señal de voz.

Al pasar dicha señal por el esquema mostrado en la figura 5, se obtuvieron los resultados que se muestran en la figura 9, la cual representa dos de las 20 dimensiones que se generan con el cálculo de los coeficientes cesptrales, tanto de escala Mel

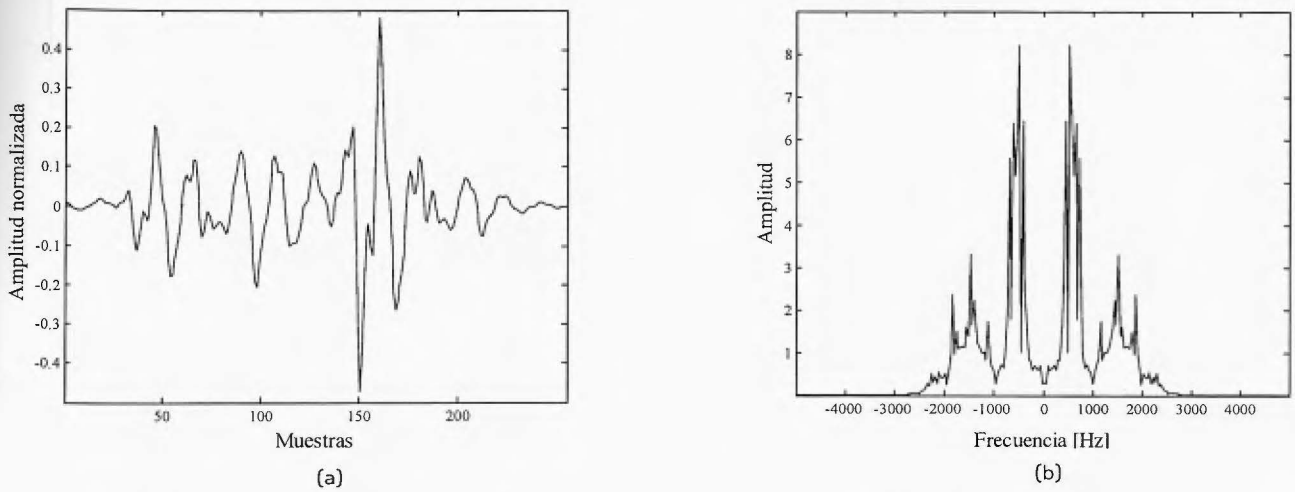


Figura 8. Representación de un segmento de la señal de voz con ventana Hamming: (a) En tiempo, (b) En frecuencia

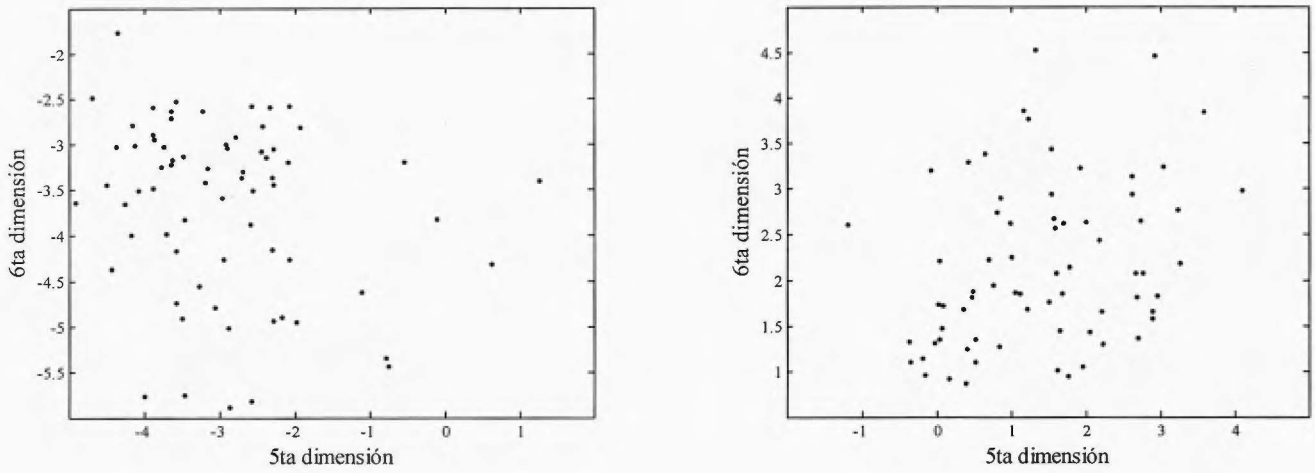


Figura 9. Coeficientes cepstrales: (a) Escala Mel (MFCC), (b) Escala lineal (LFCC)

(figura 9a) como de escala lineal (figura 9b). La ubicación de los MFCC y los LFCC es la principal diferencia entre ellos, y el tamaño de la matriz que se genera del proceso de paramatrización depende de la cantidad *n* de tramas en las que se divide la señal de voz.

Por su parte, para el caso del LPCC, los resultados obtenidos representan características diferentes, ya que se calculan primero los coeficientes del filtro de predicción lineal mediante la técnica del LPC y a partir de estos coeficientes se realiza la predicción del espectro de Fourier, tal y como se muestra en la figura 10. Esto equivale al cálculo de la Transformada de Fourier de la trama con la que se está trabajando.

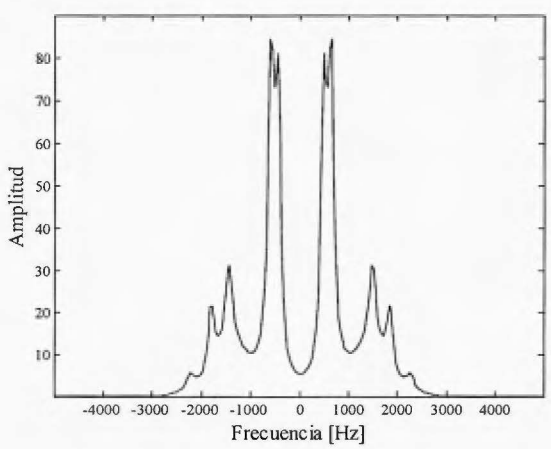


Figura 10. Predicción del espectro de Fourier de una trama de voz mediante la técnica del LPC

En este caso, el número de LPCC depende directamente del número de polos que tiene el filtro de predicción lineal, que para el proyecto implementado es de 20. El paso siguiente corresponde a la obtención del espectro logarítmico de la señal, luego se le aplica la Transformada Discreta de Coseno y se obtiene una matriz que depende del número de tramas. La representación gráfica de dos de los coeficientes calculados para la señal de voz se muestra en la figura 11

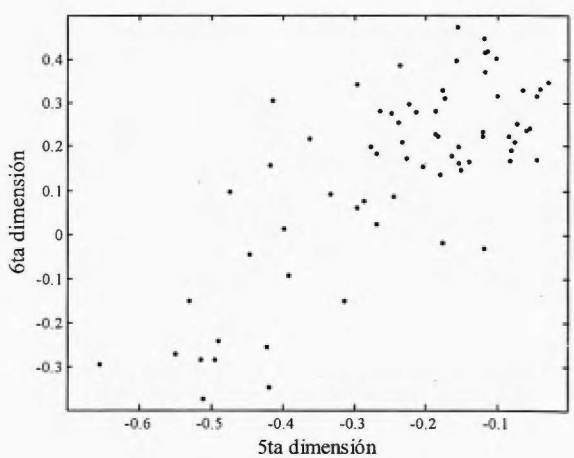


Figura 11. Representación de dos de los LPCC's calculados a la señal de voz

La diferencia entre la ubicación de los MFCC, LFCC y LPCC radica en los pasos realizados para su cálculo y en el hecho de que cada técnica extrae características diferentes para parametrizar la señal de voz.

Reconocimiento del habla

La señal que se obtiene del proceso de parametrización pasa por el proceso de cuantización vectorial. La representación gráfica de los VQ calculados a los MFCC de la figura 9a se muestra en la figura 12.

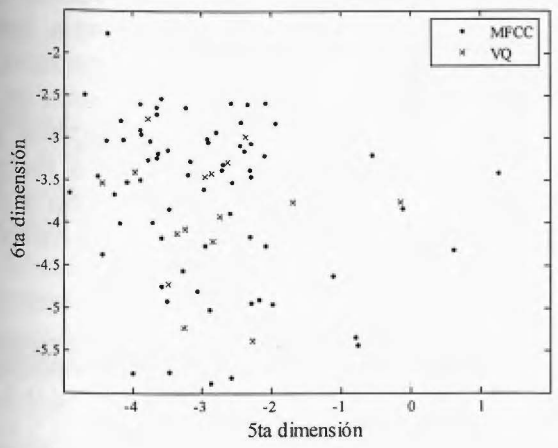


Figura 12. Cuantización vectorial de los MFCC's

Las cruces rojas que aparecen en la figura representan los 16 centroides calculados para la señal de voz de la figura 28, mientras que los puntos azules corresponden a los MFCC calculados anteriormente. Estos centroides corresponden a un vector que contiene los 16 puntos ubicados a la menor distancia de todos los coeficientes cesptrales obtenidos del proceso de parametrización (bien sea mediante MFCC, LFCC o LPCC).

Comparación entre los métodos de parametrización estudiados

Usando la base de datos para simular el funcionamiento en vivo del sistema, se realizaron las pruebas para comparar los métodos de parametrización evaluados. De acuerdo a la metodología seguida para la recolección de señales de voz, se obtuvieron 200 muestras para la base de datos de entrenamiento y 100 muestras para la base de datos de prueba del sistema. Se realizaron pruebas comparativas entre los tres métodos de parametrización estudiados: MFCC, LFCC y LPCC, usando como criterio de evaluación el porcentaje de acierto en el reconocimiento de los vocablos. El resultado obtenido para las muestras de voz correspondientes a las mujeres se muestra en la tabla 1.

La celda sombreada en la tabla 1 representa el número de aciertos que tuvo el algoritmo del LPCC para la vocal "i". Según este resultado, de 9 muestras de voz correspondientes a la vocal "i" para las mujeres, este algoritmo identificó 8 de ellas correctamente, lo que corresponde a un 88,9% de acierto para esta palabra y así se puede observar para el resto de los métodos y las vocales. Por su parte, en la tabla 2 se muestran los resultados obtenidos de las pruebas para los hombres.

	a		e		i		O		u	
	Muestras	%	Muestras	%	Muestras	%	Muestras	%	Muestras	%
MFCC	9	100	9	100	9	100	9	100	8	88,9
LFCC	9	100	9	100	9	100	6	66,7	7	77,8
LPCC	9	100	9	100	8	88,9	8	88,9	6	66,7

Tabla 1. Resultados de las pruebas para las mujeres

	a		E		i		O		u	
	Muestras	%	Muestras	%	Muestras	%	Muestras	%	Muestras	%
MFCC	11	100	10	90,9	10	90,9	8	72,7	10	90,9
LFCC	10	90,9	11	100	8	72,7	6	54,5	8	72,7
LPCC	11	100	10	90,9	11	100	10	90,9	9	81,8

Tabla 2. Resultados de las pruebas para los hombres

	a		e		i		O		u	
	Muestras	%	Muestras	%	Muestras	%	Muestras	%	Muestras	%
MFCC	20	100	19	95	19	95	17	85	18	90
LFCC	19	95	20	100	17	85	12	60	15	75
LPCC	20	100	19	95	19	95	18	90	15	75

Tabla 1. Resultados de las pruebas independiente del sexo del hablante

Según el ejemplo que se resalta en la tabla 2 se puede ver, de manera similar a la descrita para la tabla 1 que, para el caso de la vocal "a" y el método MFCC, de 11 muestras de voz recolectadas, el algoritmo acertó las 11 identificaciones correspondientes. Este mismo análisis se puede hacer para el resto de los métodos y vocales. En la tabla 3 se observa el comportamiento general de cada uno de los métodos, independientemente del sexo del hablante.

De forma similar a la que se menciona para las tablas anteriores, en la celda que se destaca en la tabla 3 se puede observar que de un total de 20 muestras de voz recolectadas para la letra "o" (9 de mujeres y 11 de hombres), el algoritmo LFCC identificó sólo 12, lo que corresponde a un 60% de acierto, bastante bajo en relación con los porcentajes de acierto del resto de los métodos.

Un aspecto importante que vale la pena destacar de los resultados mostrados en las tablas anteriores es que, independientemente del método o sexo del hablante, la mayor parte de los errores en el reconocimiento se dan para las vocales "o" y "u". Para comprender el por qué de esta situación, se muestra en la figura 13 las zonas donde se ubican los MFCC de las vocales a, e, i, o, u

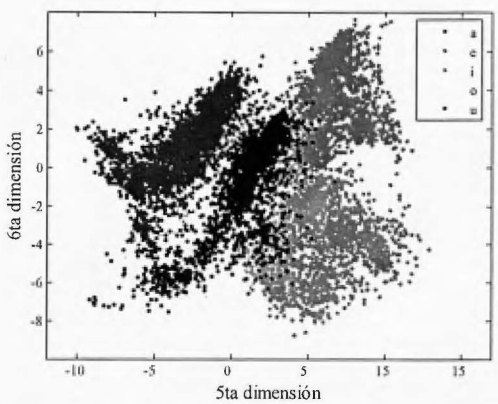


Figura 13. Principales formantes de las vocales a, e, i, o, u

La figura 13 corresponde al cálculo de los MFCC de las 20 personas que participaron en la recolección de señales de voz realizada. Se puede notar que los MFCC de las vocales "o" y "u" están ubicados en la misma zona, lo que hace muy probable que la aplicación encargada del reconocimiento al realizar el cálculo de la distancia euclidiana, tienda a confundir estas dos vocales. Por el contrario, se observa también cómo los MFCC del resto de las vocales se encuentran en diferentes lugares, por lo que se presentan escasos errores en el reconocimiento de éstas. En la tabla 4 se muestra el porcentaje promedio de acierto para los tres métodos de parametrización evaluados

	MFCC	LFCC	LPCC
Mujeres	97,8 %	88,9 %	88,9 %
Hombres	89,1 %	78,2 %	92,7 %
Total	93 %	83 %	91 %

Tabla 4. Porcentaje promedio de acierto de cada uno de los métodos de parametrización

De la tabla 4 se puede observar que el método con el mayor porcentaje de acierto total es el MFCC, por lo que se escogió esta técnica de parametrización para la implementación de la aplicación encargada del reconocimiento del habla. Se observa también que este porcentaje de acierto está seguido de cerca por la técnica del LPCC, pero la información encontrada en la revisión bibliográfica sobre el gran uso de la técnica MFCC para la parametrización de señales de voz en aplicaciones para el reconocimiento del habla, permitió hacer la escogencia de esta técnica como definitiva.

Diseño e implementación del hardware encargado del encendido y apagado de los dispositivos eléctricos

El hardware encargado del control de los dispositivos eléctricos conectados a él consta de varios

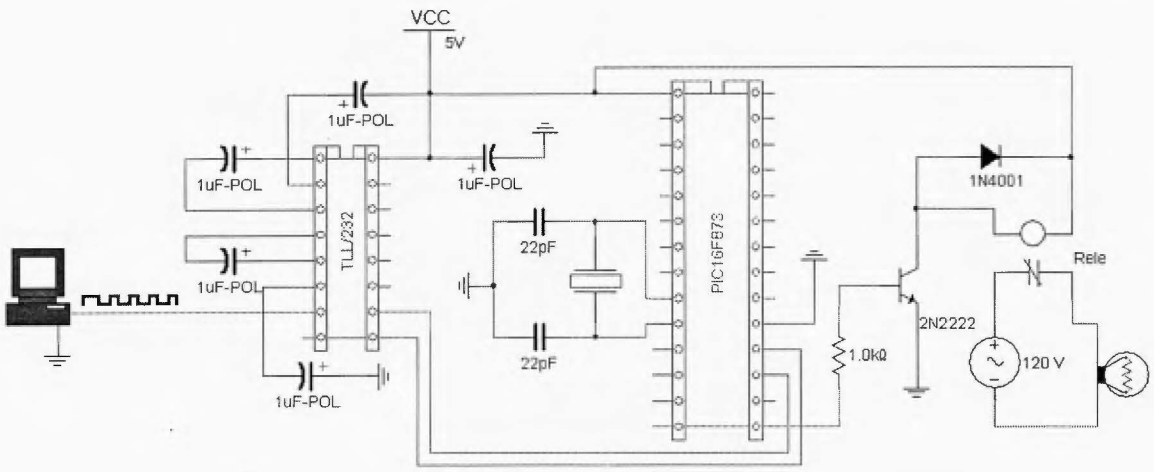


Figura 14. Diagrama circuital del hardware encargado del control de los dispositivos eléctricos

módulos: alimentación, comunicación serial con el PC y el módulo de encendido y apagado como tal, los cuales fueron combinados en el mismo circuito impreso de la forma que se observa en la figura 14

El módulo de comunicación serial con el PC representa la parte fundamental del hardware, ya que éste se basa en el uso del microcontrolador 16F873 como elemento para el procesamiento de la información proveniente del PC como resultado de la interpretación del software encargado del reconocimiento del habla. El PIC recibe por unos de sus pines de comunicación serial una señal en formato TTL, provenientes del conversor HIN232 e interpreta la misma como un comando de encendido o apagado, verifica a cuál dispositivo se hace referencia y su estado actual (encendido o apagado) y luego niega el valor lógico presente a la salida de uno de sus 5 pines (dependiendo del dispositivo, se usa un pin diferente del puerto de salida c). Tal y como se explicó en la sección II.8.2, dependiendo de la salida que produce el microcontrolador (0 ó 1 lógico), el transistor se comporta como un interruptor que cierra el circuito para polarizar la bobina del relé para encender o apagar el dispositivo eléctrico involucrado (dependiendo del estado inicial del dispositivo) o, por el contrario, mantiene el estado de actividad del mismo.

Por su parte, el módulo de alimentación se incluyó como parte integral del hardware para prescindir de alimentación externa para polarizar la bobina del relé. Este módulo consiste en una fuente fija de 5 voltios que proporciona la corriente que el PIC no puede suministrar al circuito de polarización. La unión de los módulos mencionados se realizó en una placa metálica mediante el diseño e implementación de un circuito impreso.

Conclusiones y recomendaciones

Se determinó mediante el estudio teórico-práctico que la etapa de pre-procesamiento es de gran importancia, ya que en ésta se resaltan y mejoran sustancialmente las características de la señal a parametrizar. Dentro de esta fase, la segmentación y aplicación de la ventana son especialmente importantes para asumir que la señal de voz es estacionaria en esa ventana y aplicar técnicas que permitan el Análisis de Fourier en tiempo corto. La ventana utilizada fue la Hamming, ya que ofrece el equilibrio entre la resolución en tiempo y frecuencia requerida.

Del estudio realizado para la etapa de parametrización, se encontró una gran variedad de técnicas. Se consideró desde un principio el uso de técnicas basadas en el cálculo de los coeficientes cepstrales por su gran uso en aplicaciones que involucran el procesamiento de señales de voz y por la gran variedad de combinaciones que entre ellos se pueden obtener. Como resultado del estudio comparativo, se escogió la técnica del MFCC porque su porcentaje de acierto del 93% indica el excelente desempeño que este algoritmo tiene para la aplicación desarrollada en este trabajo.

De las pruebas realizadas con la aplicación para determinar el desempeño de cada una de las técni-

cas de parametrización evaluadas, se observó que para el MFCC, el porcentaje de acierto en el reconocimiento de la vocal "a" fue de 100%, mientras que para las vocales "o" y "u" fue de 85% y 90% respectivamente. Este resultado se debe a la gran cercanía que tienen los MFCC de las dos últimas vocales, lo que crea un margen de ambigüedad en la fase de reconocimiento, mientras que la vocal "a" tiene una zona muy bien definida en la cual se ubican sus MFCC, disminuyendo cualquier posible margen de error en el reconocimiento.

En cuanto al lenguaje de programación utilizado en la aplicación encargada del reconocimiento, la elección fue MATLAB, ya que éste posee una amplia gama de módulos que pueden usarse para el desarrollo de aplicaciones relacionadas con procesamiento de señales de una forma sencilla, en comparación con otros lenguajes de programación tales como C++, Visual Basic, entre otros.

A pesar de que en el sistema no se hace consideración alguna de la influencia de un ambiente ruidoso en el desempeño y porcentaje de acierto en el reconocimiento del habla, se realizó la recolección de las señales de voz para el entrenamiento y pruebas del sistema, en ambientes con distintos niveles de ruido para simular el funcionamiento "en vivo" del sistema.

El entrenamiento del sistema se realiza al inicio de la aplicación, ya que éste consiste en el cálculo de los *codebooks* de cada una de las señales y esto lleva cierto tiempo que es preferible que el usuario espere al comienzo de la aplicación, y no cada vez que quiera realizar el reconocimiento de una palabra. De esta observación se desprende además la conclusión de que mientras más *codebooks* estén guardados en la base de datos de entrenamiento del sistema, más acertado es el reconocimiento, aunque esto implica un incremento en el tiempo de inicialización de la aplicación.

Entre algunas de las recomendaciones que se desprenden del desarrollo de este proyecto se encuentran: estudio de otras técnicas de parametrización de la señal de voz como modelos ocultos de Markov, redes neuronales, algoritmos genéticos, entre otros. Para la etapa de reconocimiento pueden considerarse otros métodos de compresión basados en cuantización vectorial, ya que ayudan a mejorar la comparación; además pueden evaluarse otros métodos de medición diferentes a la distancia euclidiana, como el Hitakura.

En cuanto al hardware empleado para controlar los dispositivos, en lugar de relés, se pueden emplear transistores de potencia y también ampliar el alcance de control de dicho hardware a dispositivos de estado sólido y no solamente electromecánicos.

Referencias bibliográficas

Davis, S. & Mermelstein, P. (Agosto de 1980). "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". *IEEE Transactions on Acoustics, Speech, and Signal Processing*. Volumen 28, páginas: 357 – 366.

Do, M. (n.d). *An Automatic Speaker Recognition System.* Extraído el 25 de noviembre de 2005 desde http://lcavwww.epfl.ch/~minhdo/asr_project/asr_project.pdf

Faúndez, M. y Rodríguez, D. (10 de febrero de 2005). *Estudio comparativo de diferentes distancias en sistemas basados en VQ para identificación automática de locutores.* Extraído el 27 de junio de 2006 desde http://veu.eupmt.es/publicacions/docs/ursi98-1.pdf

Kornhauser, D. (4 de marzo de 1999). *Diseño de un reconocedor de voz de palabras aisladas público*. Extraído el 3 de febrero de 2006 desde http://bq.unam.mx/~daniel/gvoice/tesis/node22.html

Oppenheim, A. y Schafer, R. *Discrete-Time Signal Processing.* Englewood Cliffs, NJ: Prentice Hall, 1989.

Reyes, J. A. y Herrera, J. A. (2005). *Reconocimiento de comandos de voz para Windows y Word.* Extraído el 10 de mayo de 2006 desde http://www.fip.unam.mx/ simposio investigacion2005/ponencia29 ext.html

Slavinsky, J. (19 de diciembre de 1999). *Speaker Verification*. Extraído el 16 de febrero de 2006 desde http://www.owlnet.rice.edu/~elec301/Projects99/ wrcocee/endpt.htm