Reconocimiento de palabras aisladas para control de navegación de robot móvil

Recognising isolated words for mobile robot navigation control

Reconhecimento de palavras isoladas para controle de navegação de robô móvel

Zuleika Alezones - Campos[1], Yeison Baquero-Romero[2], HenrryBorrero-Guerrero[3], Marcelo Becker[4]

[1,2*] Ingeniero de sistemas, Grupo de investigación en robótica – GIRO, Universidad de los Llanos
[3] Ingeniero electrónico, PhD(c), Grupo de investigación en robótica-GIRO,
[4] Ingeniero mecánico, MSc, PhD, Grupo de Mecatrónica, Laboratorio de Robótica Móvil, Universidad de Sao Paulo, Brasil
Email: zuleika@ieee.org

Recibido: septiembre 9 de 2010 **Aceptado**: marzo 12 de 2012

Resumen

El uso de herramientas computacionales que permiten la implementación de modelos bio-inspirados generó el desarrollo de un sistema dedicado a la caracterización de comandos de voz para controlar la navegación de un robot móvil hexápodo prototipo. A nivel general se implementó un subsistema en software, así como otro subsistema a nivel mecánico y electrónico. El subsistema relacionado con el software consta de un programa prototipo desarrollado en lenguaje java, que comprende cinco módulos: (Obtención de la señal hablada; Extracción de características; Comparación de características; Definición de los comandos para transmitir a los actuadores del mini-robot usando un Gene Digital y el módulo de comunicación de las acciones de control) de modo que la aplicación fue capaz de reconocer palabras aisladas y predefinidas en castellano emitidas por un hablante, de manera que se generen los comandos correspondientes a la navegación del robot móvil. Los comandos relacionados con la navegación del mini-robot deben ser trasmitidos a los respectivos actuadores, para este propósito se utilizó una interface hecha con un circuito electrónico simple.

Palabras clave: Codificación lineal predictiva (Linear Predictive Coding - LPC), alineación temporal dinámica (Dinamic Time Warping - DTW), robot móvil, descriptores.

Abstract

Using computational tools facilitating bio-inspired models has led to developing a system for characterising voice commands for controlling a prototype hexapod mobile robot's navigation. A software subsystem was set up, as well as another mechanical and electronic subsystem. The software subsystem consisted of a prototype programme written in Java consisting of five modules (obtaining the spoken signal, extracting characteristics, comparing characteristics, defining commands for transmitting the mini-robot actuators using a digital gene and the control action communication module) so that the application was able to recognise isolated, predefined words in Spanish emitted by a native speaker. Commands corresponding to the mobile robot's navigation were thus generated; such commands had to be transmitted to the respective actuators and thus an interface made with a simple electronic circuit was thus used for this purpose.

Key words: linear predictive coding (LPC), dynamic time warping (DTW), mobile robot, descriptor.

Resumo

O uso de ferramentas computacionais que permitem a aplicação de modelos bio-inspirados levou ao desenvolvimento de um sistema dedicado à caracterização de comandos de voz para controlar a navegação de um robô móvel hexápode de protótipo. Em geral foi implementado um subsistema software e outro subsistema no nível mecânico e eletrônico. O subsistema relacionado com o software consiste de um programa protótipo desenvolvido em linguagem Java, que inclui cinco módulos: (Obtenção do sinal falado, extração de características, comparação de características, definição dos comandos para transmitir os atuadores do robô usando um gene digital e um módulo de comunicação das ações de controle), de modo que a aplicação foi capaz de reconhecer palavras isoladas e predefinidas em idioma castelhano eemitidas por um locutor, para assim então gerar os comandos correspondentes à navegação do robô móvel. Os comandos relacionados com a navegação do mini robô tem que ser transmitidos aos atuadores respectivos, para este fim foi utilizada uma interface feita de um circuito eletrônico simples.

Palavras chave: Codificação preditiva linear (Linear PredictiveCoding - LPC), alinhamento de temporal dinâmica (Dynamic Time Warping - DTW), robô móvel, descritores.

Introducción

El problema del reconocimiento automático del habla implica el desarrollo de técnicas y sistemas capaces de lidiar con múltiples aspectos, tales como, la frecuencia de muestro de la señal de voz; el método de extracción de características que describen la señal; la comparación de patrones, la cual a su vez puede ser dependiente de la extracción de características(Bernal *et al.*, 2000) y en general un sin número de procesos los cuales también dependen del objetivo por el cual se esté desarrollado una aplicación de reconocimiento de voz.

Actualmente es común reconocer los últimos avances en el procesamiento del habla, ya que esta técnica se ha vuelto factible en la sociedad desde aplicaciones de identificación automática del locutor, servicios autónomos en llamadas, escuchar en remplazo de leer mientras se navega en un entorno, dictar en lugar de escribir o software diseñado a la medida de múltiples requerimientos. Todo ello generado a partir de la evolución de las técnicas de reconocimiento del habla.

Grandes investigaciones se han realizado por Google Research, el cual es un centro de investigación científico-electrónico; el laboratorio de sistemas de Língua Falada de Portugal; universidades como la de Edimburgo, con The Center for Speech Technology, la Universidad de Zaragoza o la Universidad Autónoma de Madrid. Adicional a ello se han desarrollado aplicaciones desde los sistemas Android o Google Voice for Mobile, para reconocimiento de voz y conversión de texto a voz en dispositivos móviles, las cuales son muy destacadas en este campo, ya que interactúan en condiciones acústicas no controladas o mientras un interlocutor realice alguna actividad, precisando alternativas y aportando avances al sistema tradicional de comunicación, mediante la puesta en marcha de la implementación de sistemas de reconocimiento de voz en múltiples campos y la vida cotidiana.

En Colombia el desarrollo de técnicas de procesamiento de voz se genera desde los centros de investigación en universidades, aportando desde allí a empresas, y estas a su vez se centran en el comercio de portales de voz, ejemplo de ello Multienlace o Tuxstone entre otras.

El estudio en detalle de las técnicas y pasos del procesamiento de la voz para su posterior duplicación, es un factor crucial para desarrollos en el tema. En este artículo se expone la implementación de un programa prototipo desarrollado en lenguaje java que permite reconocer palabras aisladas emitidas por un locutor de lengua castellana, que está constituido por cinco módulos. El módulo para la obtención de la señal hablada, Módulo de extracción de características utilizando el método de codificación por predicción lineal LPC (Linear Predictive Coding), Módulo de comparación de características usando el método de alineación temporal dinámica DTW (Dinamyc Time Warping), Módulo de definición de los comandos para transmitir a los actuadores del mini-robot usando un Gene Digital y el módulo de comunicación de las acciones de control. El módulo de comunicación de las acciones de control genera los comandos requeridos para accionar los actuadores de un mini-robot móvil hexápodo para así gobernar su respectiva navegación.

Mostrando en el presente trabajo un análisis referente a cada una de las etapas básicas y necesarias al mo-

mento de desarrollar un sistema de reconocimiento computacional de comandos de voz, adicionando a ello el gene digital. Exponiendo las particularidades sobre la implementación y puesta en marcha, tanto en software como en hardware, útiles para aplicar desde diferentes técnicas a las tratadas, abordando por ultimo una discusión sobre la aplicación desarrollada, en la que se centraliza el conocimiento trabajado, enfocándolo finalmente a la exposición de unas conclusiones obtenidas del sistema.

Materiales

En el desarrollo del sistema de reconocimiento de voz, se tuvo en cuenta el lenguaje de programación en el cual se deseó desarrollar según la necesidades, para este caso fue JAVA; de aquí se observaron los formatos de audio soportados y el que más se ajustó en el sistema, en este caso WAV ya que es sin compresión; posterior a ello se analizaron las frecuencias a trabajar, en el caso de la voz, su frecuencia máxima está por debajo de los 8000 Hz, y mediante teorema de Nyquist que indica que al trabajar con señales es adecuado hacerlo al doble de su frecuencia es claro que el valor que se trabajo es de 16 KHz, sin embargo los valores utilizados fueron de 11 KHz, 22 KHz, 44 KHz, etc.; así que el valor implementado en la frecuencia de muestro fue de 22 KHz. Se seleccionaron aquellos materiales que dieran valor agregado al sistema, por lo que se tuvo especial cuidado en la elección del micrófono y la tarjeta de sonido; el micrófono dependiendo sus características puede ayudar o entorpecer la labor, ya que se necesita uno que capture especialmente la voz y no información del ambiente, en cuanto a la tarjeta de sonido, su importancia radica en que algunas ofrecen filtros que ayudan en el proceso de reconocimiento.

El sistema trabaja directamente con un computador de características básicas, en el cual se realizan todos los procesamientos, enviando finalmente una señal de control para el mini robot móvil por puerto paralelo en prototipos iníciales, y actualmente mediante módulos XBee.

Finalmente, la locomoción del mini-robot se baso en un modelo trípode de movimiento, igual como sucede con los seres vivos, el mini-robot debe ser capaz de soportar su propio peso y superar la fuerza de gravedad., con este modelo trípode se pretendió mantener tres extremidades en el suelo y darle libertad de movimiento a las demás; una ventaja de este modelo caminador consiste en la estabilidad que se genera para el mini-robot y que permite aislar el cuerpo del terreno empleando puntos discretos de soporte. Así mismo, mediante patas, es posible conseguir cierta omnidireccionalidad y el deslizamiento ocasionado por la locomoción es mucho menor.

Como se muestra en la figura 1 (A), en la posición inicial el prototipo debe mantener todas sus patas en el suelo, seguido de esto (figura 1 (B)) se reafirma la posición fija para tres de las patas las demás avanzan;

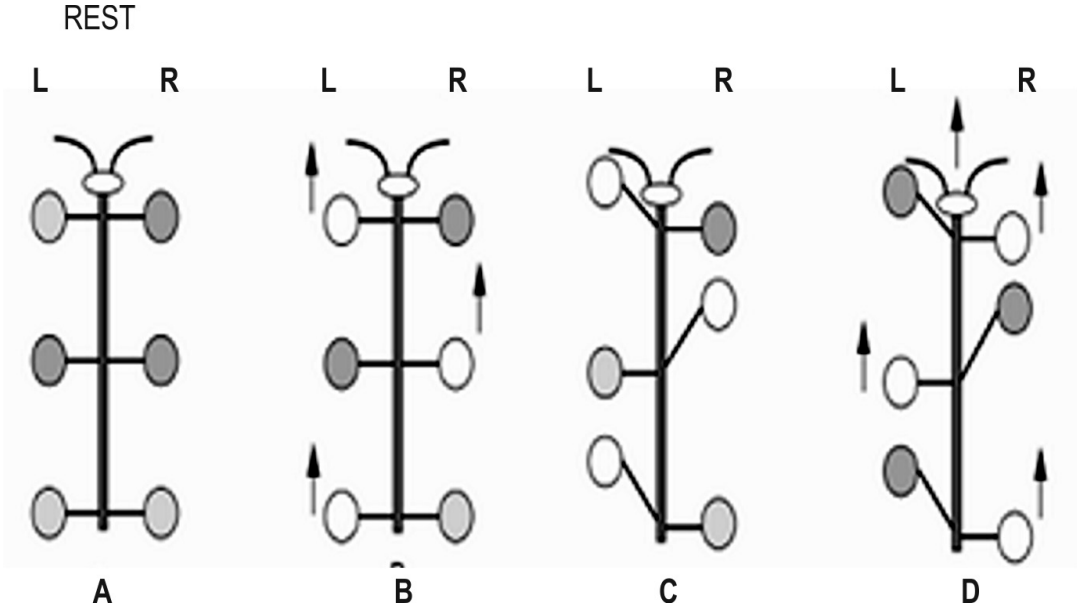

Figura 1. Representación sobre el modelo trípode para la locomoción del hexápodo.

el siguiente paso es fijar las patas que avanzaron lo que permitirá el avance de las demás. Esta secuencia de movimientos fue reiterativa hasta el momento en el que se reinicie el algoritmo que controla la cinemática que realiza el robot. En la figura 10 se observa una imagen de la estructura mecánica implementada para el funcionamiento del robot móvil

Metodos

El objetivo principal de un sistema electrónico reconocimiento automático del habla haciendo uso de sistemas computacionales es convertir con precisión y eficiencia una señal hablada en una representación binaria que resulte fácil de procesar para diversas aplicaciones, independientemente del ambiente en el cual se encuentre localizado el locutor Rabiner, (2007).

Un modelo simple sobre un sistema de reconocimiento del habla se puede apreciar en la figura 2, donde se asume que un locutor emite sentencias lingüísticas propias del habla que son capturadas utilizando para ello un transductor que permita convertir la señal de voz en una señal eléctrica, la cual es digitalizada. En el caso de la figura 1 las sentencias corresponden a la señal de entrada al bloque de digitalización que en la salida proporciona la señal s[n]. La señal de voz digitalizada [n] cuenta con las condiciones adecuadas para ser procesada por un sistema computacional.

La señal s[n] requiere de tratamiento adecuado comprendido en general por las etapas de pre-procesamiento, extracción de características y reconocimiento todo ello representado en la figura 3.

Pre-procesamiento

La unidad básica del habla según sea la aplicación puede ser conformada por fonemas, vocales, silabas, palabras, frases, etc y que estas unidades son las que se pretende reconocer (Bernal *et al.*, 2000). Se cuenta con un conjunto de grabaciones realizadas en un ambiente adecuado las cuales son denominadas como *muestras,* sobre las *muestras* se realiza el preprocesamiento.

El análisis de una señal de voz requiere tener en cuenta que dicha señal presenta por naturaleza una atenuación en frecuencias altas, es por esto que se debe realizar un filtraje que permita obtener información suficiente sobre dichas frecuencias para no concentrarse únicamente en la información contenida en las frecuencias bajas, resaltando así que el oído humano es más sensible a frecuencias en la zona de los 3000Hz.

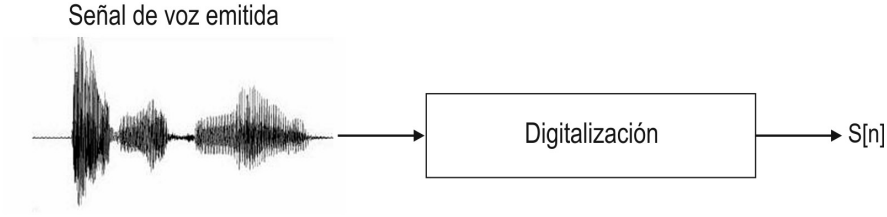

Figura 2. Esquema general de un sistema de reconocimiento del habla.

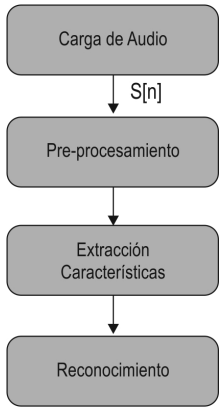

Figura 3. Etapas en el procesamiento de señal de voz.

$$y[n] = x[n] - \alpha[n-1] \quad (1)$$

Un filtro preénfasis es representado matemáticamente por la ecuación 1, donde x[n] es el vector de amplitud de la señal de voz, y[n] es la señal de salida del filtro preénfasis. Si $\alpha < 0$, se tiene un filtro de paso bajo y si $\alpha > 0$ se tiene un filtro de paso alto, para los fines requeridos se utiliza $\alpha = 0.97$.

Debido a que las señales de voz contienen numerosas variaciones debido a la suma de distintas frecuencias, el proceso de extracción de características se realiza en intervalos cortos en el tiempo. A este proceso se le conoce como ventaneo Rabiner, (2007).

Durante la etapa de pre-procesamiento se debe realizar segmentación y ventaneo de la señal ahora obtenida a partir del filtro pre-énfasis. La aplicación de segmentación y ventaneo se observa en la figura 4, en la parte inferior muestra una señal de voz, haciendo notorio sus múltiples variaciones. Para poder analizar esta información es indispensable tomar segmentos (dividir la señal), que son los tramos que se observan en la parte superior de la imagen (fragmentos de 30ms), sin no es suficiente con esto es necesario desarrollar un solapamiento del 50% y no tomar la información una detrás de la otra, ya que existirían tramos fundamentales sin analizar y por último a esta información se le aplica una función ventana. Se segmenta debido a que se requiere tener en cuenta las características fundamentales de las palabras emitidas por el locutor que están siendo analizadas. Se aplicó una función ventana para resaltar la información central de cada segmento y se solapo para hacer un corrimiento y no dejar fragmentos de la señal sin analizar. La elección sobre la grandeza del incremento se ve observa en la figura 4, la cual se realiza teniendo en cuenta la dimensión de la ventana.

Para evitar los efectos negativos provocados al tomar un segmento que contenga una transición de una zona de la señal cuasi-estacionaria a la siguiente, se trabajó la técnica del solapamiento de segmentos. Este método consiste en tomar una separación entre los comienzos de cada segmento menor que la longitud de los segmentos, con lo que se produce un solapamiento entre los mismos.

Una vez elegido el tamaño de ventana, a cada una se le asignó una función, con el fin de disminuir la importancia de los valores que se encuentran a los extremos de la misma, para evitar que características de estos valores varíen la interpretación de la parte central del bloque que es la más significativa. Los tipos de funciones ventana más utilizados son el tipo Hann representado en la ecuación 2 y el tipo Hamming en la ecuación 3 (Bernal *et al.*, 2000) en donde en V(n) corresponde a los valores y rangos característicos para cada tipo de ventana para todos los n valores de la señal, ajustados a N tiempos de duración en las muestras de la ventana aplicada.

$$V(n) = \begin{cases} 0.5 - 0.5 \times \mathrm{Cos}\left(\frac{2\pi n}{N}\right) \text{ si } 0 \le n \le N \\ 0 \text{ en otro caso} \end{cases} \quad (2)$$

$$V(n) = \begin{cases} 0.54 - 0.46 \times \mathrm{Cos}\left(\frac{2\pi n}{N}\right) \text{ si } 0 \le n \le N \\ 0 \text{ en otro caso} \end{cases} \quad (3)$$

Extracción de características (descriptores)

Una vez realizado el ventaneo, los segmentos de voz ya son aptos para la aplicación de técnicas de extracción de patrones como es el caso de LPC (codificación por predicción lineal), el cual está basado en la producción del habla. Se utiliza este modelo debido a que proporciona un patrón adecuado de la señal de voz; sus parámetros se ajustan a las características del tracto vocal; representa la envolvente espectral de la

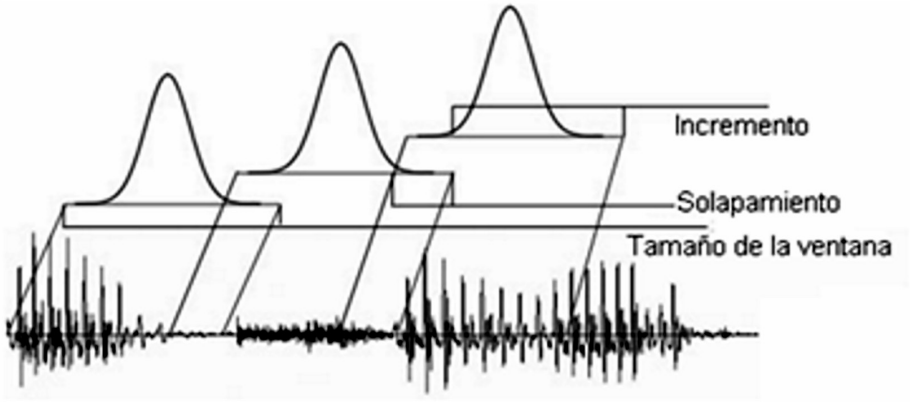

Figura4. Ventaneo[1] de una señal.

[1] Ventaneo: En inglés "*Windowin*", (significa fragmentar una señal en ventanas).

señal de forma comprimida; los parámetros obtenidos mediante predicción lineal muestran un espectro suavizado que proporciona la información más representativa de la voz y es un método preciso, adecuado para computación, tanto por su sencillez como su rapidez de ejecución.

El concepto básico de predicción lineal (LPC) se centra en que una muestra de una señal de voz *x*(n) puede ser predicha por las *k* muestras anteriores de la misma señal, generando una señal aproximada $\tilde{x}(n)$, representada por medio de la ecuación 4.

$$\tilde{x}(n) = \sum_{i=1}^{k} \alpha_i * x(n-1) \quad (4)$$

Presentando en esta fase un error de predicción, el cual se establece en la ecuación 5.

$$e(n) = x(n) - \tilde{x}(n) \quad (5)$$

Para hallar los coeficientes α_i de la ecuación 4 minimizando el error, se aplican mínimos cuadrados al intervalo de N muestras que se desee considerar, como se aprecian en las ecuaciones 6,7 y 8.

$$L = \sum_n e^2(n) = \quad (6)$$

$$\sum_n [x(n) - \tilde{x}(n)]^2 \quad (7)$$

$$\sum_n \left[x(n) - \sum_{i=1}^{k} \alpha_i * x(n-i) \right]^2 \quad (8)$$

Para obtener el valor mínimo de *L,* se deriva parcialmente la expresión 8 respecto a cada una de las variables α_j $1<j<k$ y se iguala a cero.

$$\frac{dL}{da_j} = \frac{d\sum_n \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right]^2}{da_j} = 0 \quad (9)$$

$$\frac{dL}{da_j} = 2\sum_n \left(x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right) * (0 - x(n-j) = 0 \quad (10)$$

$$\frac{dL}{da_j} = \sum_n \left(x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right) * (x(n-j) = 0 \quad (11)$$

para $1 \le j \le k$

Desarrolando la ecuación 11 se tiene lo siguiente:

$$\sum_n x(n-j) * x(n) - \sum_{i=1}^{k} a_i * \sum_n x(n-j) * x(n-i) \quad (12)$$

Si se define

$$C_{ij} = \sum_n x(n-j) * x(n-i) \quad (13)$$

Reemplazando en la ecuación 12 se obtiene:

$$C_{j0} - \sum_{i=1}^{k} a_i * C_{ji} \quad (14)$$

Método de auto correlación: Se puede observar en la ecuación 14 que la correlación existente en i-j con lo que $C_{ij} = C_{ji} = r_{|i-j|}$, donde los $r_{|i-j|}$ son los coeficientes de correlación, de esta manera la expresión (13) se puede expresar:

$$\sum_n x(n-j) * x(n-i) = \sum_n x(n) * x(n + |j-j|) = r_{|j-j|} \quad (15)$$

Reemplazando $r_{|i-j|}$ en la ecuación (14) se obtiene:

$$\sum_{i=1}^{k} r(|j-i|) * a_i = r(j), \quad 1 \le j \le k \quad (16)$$

La ecuación (16) puede ser expresada en forma matricial:

$$\begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \dots & r_n(k-1) \\ r_n(1) & r_n(0) & r_n(1) & \dots & r_n(k-2) \\ r_n(2) & r_n(1) & r_n(0) & \dots & r_n(k-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_n(k-1) & r_n(k-2) & r_n(k-3) & \dots & r_n(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_4 \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(k) \end{bmatrix} \quad (17)$$

La matriz de la ecuación 17 de autocorrelación puede ser resuelta aplicando el algoritmo de Levinson-Durbin estudiado en (Alvarado, 2008; Keogh, 2001) u otro método de algebra lineal. Con los coeficientes obtenidos del desarrollo de esta matriz, se generan los descriptores LPC de cada ventana aplicada, que en conjunto darán el grupo de vectores descriptores de la señal de voz.

Reconocimiento

Para la ejecución de la etapa de reconocimiento se tiene pre-establecido el vocabulario del sistema que en caso presente corresponde con las palabras *adelante, atrás, izquierda, derecha,* así como con los parámetros LPC. La palabra pronunciada por el locutor es parametrizada del mismo modo que las palabras definidas como vocabulario. La fase de reconocimiento se inicia con la palabra pronunciada por el locutor y el correspondiente patrón LPC que es comparado con los patrones de referencia previamente almacenados en memoria usando una medida de similitud (Hamming, euclidiana, distancia máxima); La

medida de distancia entre los parámetros usados es la euclidiana tal como se aprecia en la ecuación 18 (Bregón, 2005):

Parámetros $P=(p_1,p_2,\cdots,p_n)$ y $Q=(q_1,q_2,\cdots,q_n)$

$$d=\sqrt{(p_1-q_1)^2+\cdots+(p_n-q_n)^2} \qquad (18)$$

$$d=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

Debido a la variabilidad intralocutor de la señal, hay diferencias no lineales en la duración de los sonidos y la velocidad de pronunciación de los mismos, incluso tratándose de una palabra igual. Por tanto, se realiza un alineamiento temporal de los patrones (los recién calculados y los almacenados en memoria propios a cada palabra del vocabulario), el cual consiste en minimizar la distancia total entre dichos patrones.

Para realizar el alineamiento se utiliza el método de DTW (Dynamic Time Warping) (Alvarado,2008; Keogh, 2001) que se basa en determinar el patrón más similar a la palabra pronunciada, es decir, el que demuestra una menor distancia euclidiana en la etapa de comparación. De manera más explícita, la asimilación se realiza a cada palabra del vocabulario generando un plano de dos ejes el cual, uno lo conforman los parámetros calculados y otro el de los almacenados, en donde, cada punto o intersección en el plano es la distancia euclidiana calculada, teniendo como finalidad encontrar la ruta mínima D desde el origen hasta la última intersección de ambos ejes, mediante la suma de las distancias de la diagonal en el plano.

$$D=\sum d \qquad (19)$$

La ecuación 19 indica la sumatoria de las distancias locales, en donde cada valor d es la distancia euclidiana calculada entre los parámetros obtenidos de la palabra recién pronunciada y los parámetros que relacionados al vocabulario almacenado en memoria.

Como se observa en la figura 4 se implementa un plano en donde un eje corresponde a la información LPC de la palabra a reconocer y otro a la información LPC de la palabra almacenada en memoria, allí cada punto de intersección es el cálculo de la distancia euclidiana entre los valores de los ejes; partiendo desde el origen, se busca entre los vecinos cercanos al que le corresponde menor distancia, se selecciona y se salta a este punto, tomándolo como referencia, luego nuevamente se observa entre los vecinos cercanos y se pasa al que menor distancia ofrece, repitiendo esta misma operación de manera sucesiva; de esta manera se obtiene el trayecto de las menores distancias euclidianas ofrecidas, las cuales se analizan al llegar hasta el último punto de intersección; Por último se suman dichas distancias obtenidas, generando un valor de reconocimiento (D).

Se realiza un límite diagonal o radio para lograr mejores resultados y evitar que la ruta siga en forma vertical o diagonal, manera que los valores seleccionados en el cálculo de las distancias e implementación del algoritmo estén más cerca de la diagonal, como se muestra en la figura 5.

La etapa de reconocimiento proporciona una salida que corresponde con una palabra binaria que repre-

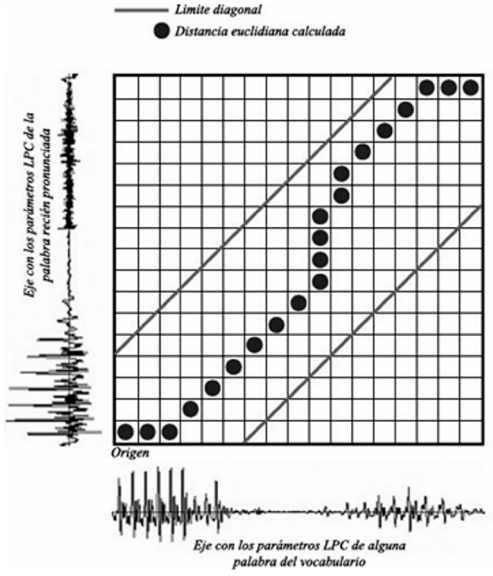

Figura 5. Representación espacial en el plano generado por las características de la palabra adelante.

senta la palabra pronunciada por el locutor. Dicha palabra es ingresada a un gene digital para generar acciones de control sobre los actuadores del robot móvil.

ADN y Gene digital

El ADN está constituido por moléculas denominadas nucleótidos, cada nucleótido está compuesto de un fosfato, un azúcar de cinco carbonos (desoxirribosa) y una base nitrogenada (Campbell,2003; Alberts,1998). Los nucleótidos derivan su nombre de la base que poseen las cuales se clasifican en dos pirimidinas {Citosina (C), Timina (T)} y dos purinas {Adenina (A), Guanina (G)}. En la figura 5(a) se representan los cuatro nucleótidos constituyentes del ADN cada fosfato (cruz) se enlaza con la desoxirribosa (óvalo) y el grupo fosfato – azúcar se enlaza a una base nitrogenada.

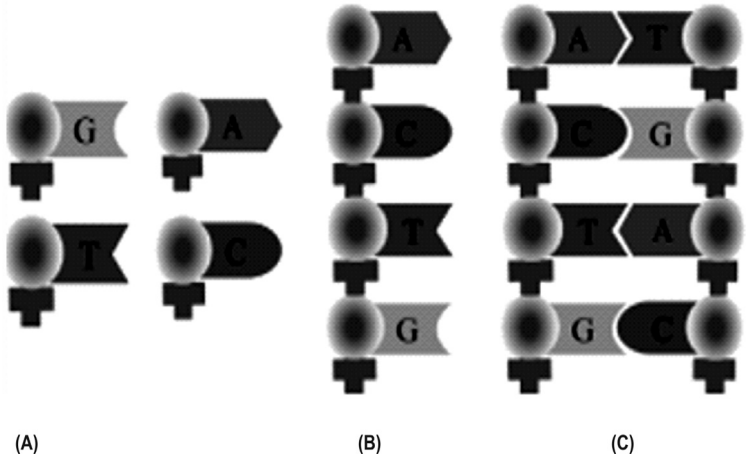

Figura 6. Representación básica del ADN. (a) los nucleótidos, (b) cadena sencilla de ADN, (c) doble hélice de ADN.

Una cadena de ADN está conformada por una secuencia de nucleótidos como se representa en la figura 6 (b). Existe un principio natural denominado complemento Watson – Crick en el cual dos nucleótidos son complementarios si sus bases son complementarias, la Adenina complementa la Timina y la Citosina complementa la Guanina, el complemento corresponde a la unión por enlaces de hidrógeno. Es importante tener en cuenta que para el caso de las cadenas sencillas de ADN (ver figura 5(b)) se cuenta con 4 elementos de base de modo que si se podría en principio que si dispone de *n* bases disponibles para formar una sola cadena, es posible que se forme alguna de 4^n posibles combinaciones.

Los genes son segmentos de ADN que codifican una o más proteínas, en la figura 6 se muestra un diagrama de bloques de las regiones constituyentes de un gene.

La región reguladora (RR) es donde las proteínas y otras moléculas se fijan para iniciar o parar la expresión de los genes, l segunda región corresponde a la región codificadora (CR) que es el segmento de DNA que se transcribe en RNA (Campbell, 2003; Alberts, 1998). Desde el punto de vista de la emulación electrónica, se encuentran varias posibilidades de programación e implementación del gene biológico, conocido como gene digital (Prieto, 2007).

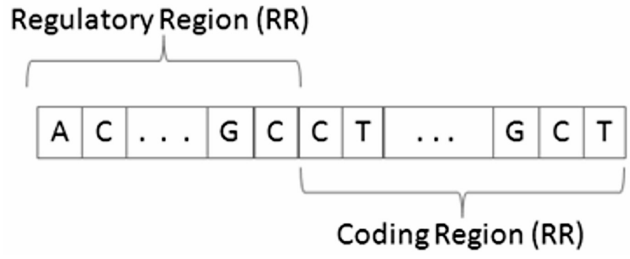

Figura 7. Representación general de un gen

Basados en los aspectos mencionados, se plantea el gene digital como una configuración de registros de forma tal que se logre asociar una acción a un comportamiento determinado en la entrada del sistema.

Con el apoyo de la figura 8 se puede describir la conformación del gene digital que está compuesto por cuatro secciones: un registro de entrada (u); una serie de registros de prueba (p_i); registros asociados a los anteriores (y_i); un registro de salida (y). De acuerdo con la figura 7 el sistema ahora presentado incorpora en un solo registro que se le denomina *Registro de entrada u,* la palabra generada desde la etapa de *reconocimiento* del programa de reconocimiento del habla.

La palabra en el registro de entrada proviene de la salida del sistema de reconocimiento del habla y representa un problema que hay que resolver, en este caso se pretende que el robot realice desplazamientos básicos hacia adelante, atrás, a la derecha y a la izquierda, así, los registros de prueba almacenan palabras que puedan complementar o incluso hacer hibridación exitosa con la palabra de entrada. La operación consiste en realizar comparaciones de forma paralela entre el contenido de registro de entrada y cada uno de los registros de prueba. A partir de esta comparación y según alguna restricción, el registro de prueba asociado correspondiente, es o no concatenado al registro de salida.

En el registro de salida residen los comandos de control correspondientes a los actuadores del mini-robot. La comparación entre el registro de entrada (U) y los registros de prueba se realiza detectando la distancia de Hamming, que básicamente mide el número de bits diferentes entre los registros. Para determinar la condición de concatenación, se define un parámetro llamado umbral de Hamming, si la distancia es menor al umbral, el contenido del registro de prueba asociado correspondiente se concatenará al registro de salida.

La figura 9 permite explicar el concepto con más detalle: La palabra de entrada (*U*) que corresponde a una cadena de bits asociada a un comando de voz caracterizado a vector binario desde el sistema de reconocimiento automático del habla. El vector binario del registro *U* es comparada paralelamente con un cierto número de palabras cargadas en los registros P_{r1} a P_{rN}, Si alguna de estas palabras genera sobrepasamiento de un umbral θ_N (umbral de Hamming), entonces se habilita el correspondiente bloque "*Enable*" y por lo tanto la palabra contenida en el respectivo registro e_N pasa al registro de salida (*exón concatenation*). El *Exon* contiene el contenido habilitado hacia

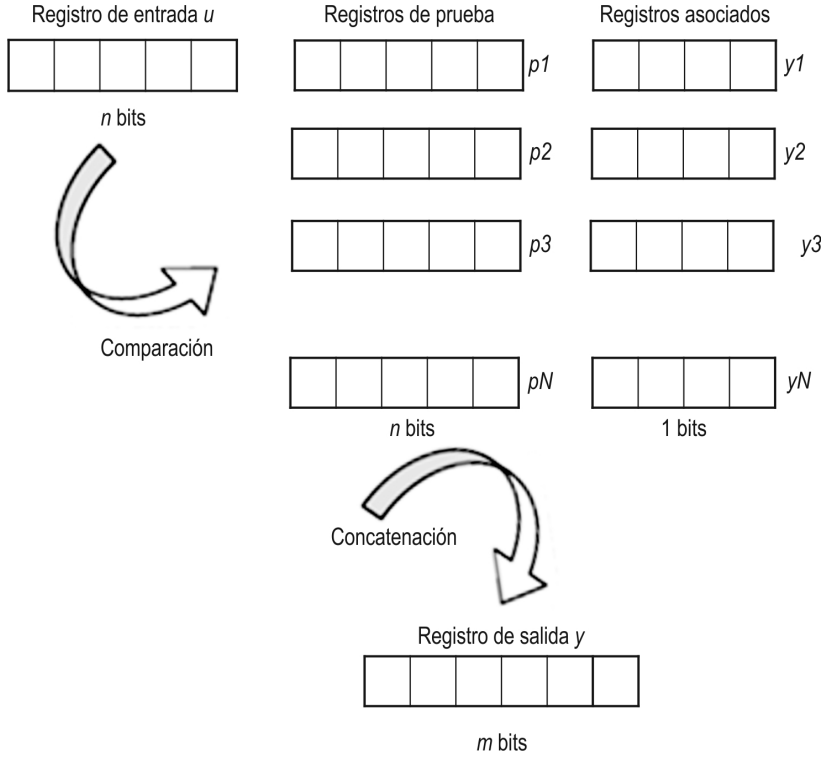

Figura 8. Diagrama de registros en el gene digital.

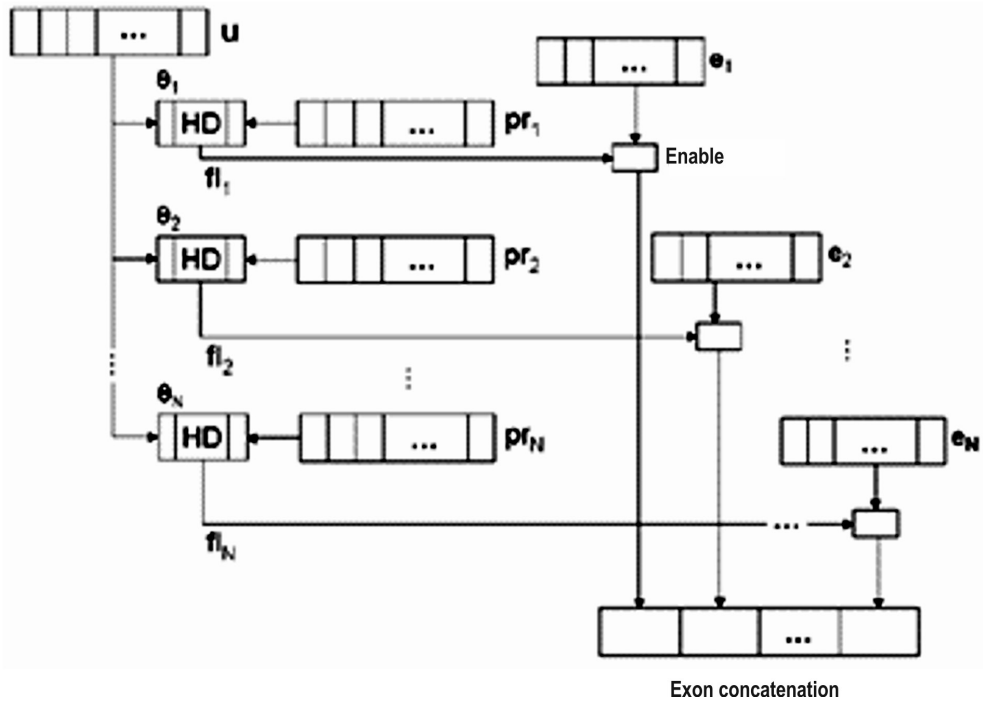

Figura 9. Diagrama gene digital

la salida desde los registros e_N, dicho contenido está asociado a las acciones de control del mini-robot.

Fundamentalmente, lo que se busca es que las palabras que hagan parte de cada uno de los registros de prueba (P_{r1} a P_{rN}) representen una solución a cada problema que se manifiesta en el registro de entrada *U*, con el objetivo de conseguir que se realicen acciones de control exitosas gracias a la concatenación adecuada de acciones de control (e_1 a e_N) en el registro de salida (*Exon con catenation* en la figura 8). No deja de ser importante anotar que si una palabra en algún registro de prueba no genera la superación del umbral de Hamming, no habrá concatenación de la acción de control (*e*).

Asumiendo que la palabra reconocida fue "*derecha*", ya que esta palabra se encuentra contenida en el registro en entrada *U* del gene digital de la figura 8, así "*derecha*" es comparada paralelamente con los contenidos de cada uno de los registros (P_{r1} a P_{rN}). Si la acción de control que hace que el robot móvil se desplace a la derecha está localizada en el registro e_2, entonces la palabra que genere superación del umbral de Hamming, cuando la entrada manifieste "*derecha*" debe ser ubicada en el registro P_{r2}; Si la palabra que hace superar el umbral para el caso expuesto se encuentra en otro registros de prueba, entonces no se realizará acción de control exitosa.

De acuerdo con lo anterior, resulta evidente que el contenido de los registros de prueba puede ser definido por auto-aprendizaje, realizando un entrenamiento para que el sistema se adapte aprendiendo de los errores cometidos para así colocar un contenido adecuado en los registros de prueba usando un algoritmo genético, de forma similar al entrenamiento del chip ADN emulado electrónicamente (Borrero, 2008). En el presente estudio, el contenido de los registros de prueba fue definido previamente e instalado por el diseñador.

Implementacion

El sistema de reconocimiento del habla fue entrenado para reconocer las palabras *"adelante", "atrás", "derecha", "izquierda"* generadas por un mismo locutor, cada palabra reconocida se caracterizó en forma de un vector binario.

El vector binario caracterizado en el subsistema de reconocimiento del habla se transmitió al registro de entrada del gene-digital, donde en cuatro registros de prueba se almacenaron las expresiones binarias que generan un umbral adecuado para activar las acciones de control residentes en los registros (figura 8).Las palabras de control que residen en el exón de salida del gene digital se trasmiten a las respectivas entradas de los actuadores del mini-robot (motores).

Figura 10. Prototipo inicial mini-robot.

En el diagrama de bloque de la figura 11 se representan los desplazamientos que puede realizar el mini-robot. Dichos desplazamientos obedecen a la programación de secuencias de ejecución para cada actuador, estas secuencias se encuentran previamente programadas y se puede decir que cada movimiento es como tal un programa.

En la figura 11 se aprecia la entrada llamada exón, que corresponde al registro de salida del gene digital, si el exón contiene la palabra para desplazarse hacia *adelante*, entonces el programa correspondiente ha de ser habilitado activando el bloque *E*, mientras que los programas concernientes a los otros desplazamientos deben ser inhabilitados. En la figura 11 también se aprecia que existe una conexión entre los bloques de desplazamientos y lo tres motores (actuadores) del robot, aquí es bueno resaltar que los motores solo recibirán comandos de actuación desde el bloque habitado gracias a la respectiva activación de bloque *E*.

En cuanto a la aplicación software, la interfaz gráfica consta de tres componentes, el primero destinado a la captura o carga de audio, el segundo para extracción de características y reconocimiento de voz y el último de análisis. Adicionalmente, cuenta con un panel inferior en el cual se visualizan los resultados de aquellos procesos que lo requieran (figura 12).

La figura 13 muestra las etapas ejecutadas en el reconocimiento automático del habla, con los tipos de datos que se manejan en forma específica para la aplicación: En la primera parte se obtiene el vector de voz en el tiempo, ya sea desde el modulo del micrófono o desde archivo, esta información es entregada en forma binaria, con lo cual hay que realizar el respectivo cast para poder ser trabajado directamente en las posteriores etapas. En la siguiente etapa se realiza el filtro preénfasis sobre los datos de voz.

Posteriormente se obtienen los descriptores por cada ventana de voz que en conjunto darán una matriz

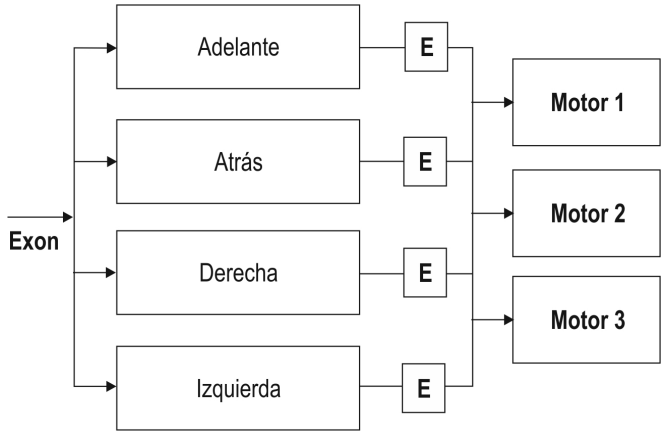

Figura 11. Esquema sobre la secuencias de movimiento del mini-robot

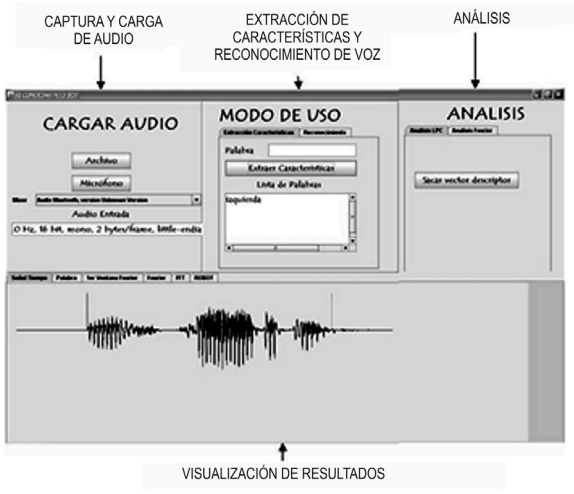

Figura 12. Interface gráfica, aplicación de Reconocimiento de voz.

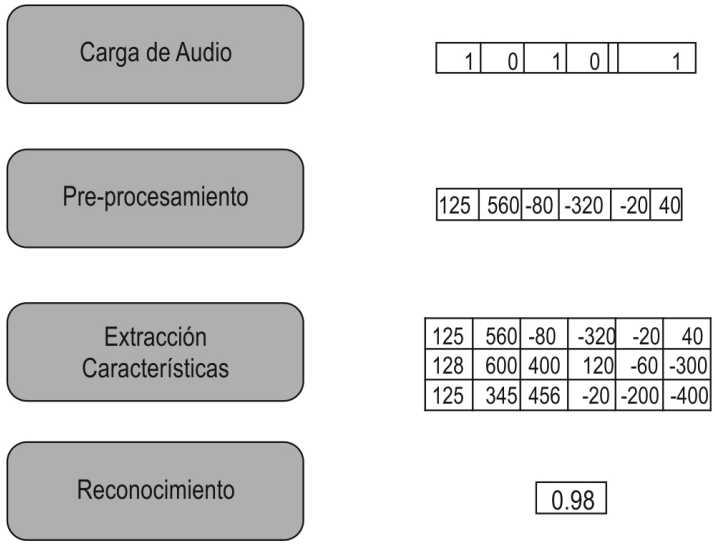

Figura 13. Salida de la información en cada fase del sistema.

de elementos propios de la palabra, dependiendo si el siguiente paso que es el reconocimiento, las características son guardadas o no, para ser comparadas.

Por último se igualan las características de las palabras que se tienen almacenadas y las que se desean comparar, lo cual mostrara cuales de estas arrojan la menor distancia y por consiguiente será la palabra reconocida.

Resultados

La aplicación fue probada en una población de 25 hombres y 25 mujeres (6 menores de 12 años, 11 de 12 a 19 años, 12 de 20 a 35 años y 21 mayores a 35 años); de los cuales se tomaron 5 muestras de cada palabra (*"adelante", "atrás", "derecha"* y *"izquierda", adicionando además el reconocimiento de la palabra "alto"* para un desarrollo posterior de una asociación para respuesta). Teniendo en total 25 muestras por persona (adelante 1-2-3-4-5, atrás 1-2-3-4-5, izquierda 1-2-3-4-5, derecha 1-2-3-4-5 y alto 1-2-3-4-5). El sistema fue probado mediante el entrenamiento de las muestras n 1 inicialmente y recopilando los resultados mediante el reconocimiento de las 4 restantes, posteriormente se realizó la misma acción de entrenamiento para las muestras 2, 3, 4, 5 y sus respectivas pruebas con las muestras sobrantes.

El reconocimiento para los hombres fue 86,8% (figura 14) y para las mujeres 82% (figura 15) y un reconocimiento para la población general de 84,45% (figura 16).

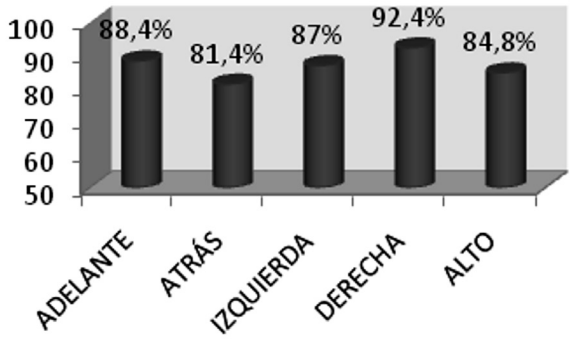

Figura 14. Porcentaje de reconocimiento para hombres.

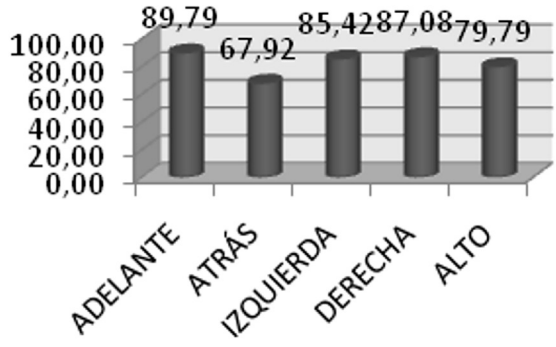

Figura 15. Porcentaje de reconocimiento para mujeres.

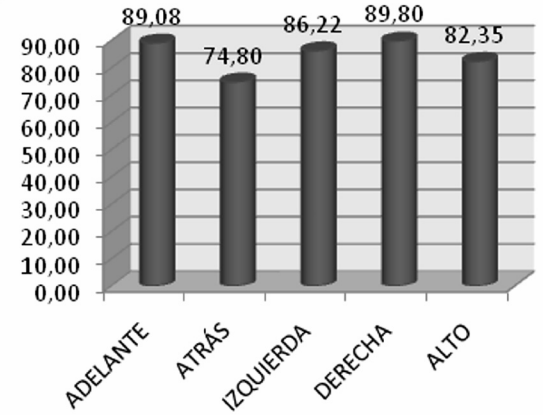

Figura 16. Porcentaje de reconocimiento total de la aplicación.

Las muestras fueron tomadas en un ambiente bastante natural, por tanto la efectividad de la aplicación que reside en un 84,45% puede aumentar considerablemente si se tiene en cuenta condiciones lo más ideales posibles.

La toma de las señales de voz, se realizó de manera continua, es decir, grabaciones en las cuales el locutor repetía una palabra varias veces y posteriormente recortadas las palabras para formar un conjunto de muestras; lo que género como resultado posibles errores humanos al momento de manipularlas.

A pesar de las condiciones adversas en que fueron tomados y manipulados los datos, el porcentaje de reconocimiento es bastante aceptable para el sistema.

Al momento de desarrollar un reconocedor de voz sin filtros de ruido, hay que tener en cuenta el ambiente en el cual se va a aplicar; ya que en este caso, el ruido

sería necesario, porque al momento de reconocer las muestras presentaran un mayor grado de semejanza.

Discusión

La aplicación desarrollada es el principio para un abordaje más profundo en el ámbito del procesamiento digital de señales usando circuitos electrónicos que permitan embeber software, pues en la robótica se desea liberar peso, de modo que la utilización de software usando un computador de escritorio o incluso de características portables a pesar de generar resultados aceptables, afectan la movilidad de un robot, lo cual se soluciona con el uso y aplicación de sistemas embebidos.

La arquitectura del gene digital está orientada a su implementación en hardware para así explotar sus capacidades de paralelismo y memoria asociativa. En este documento se expone su aplicación a nivel de software, de manera que se considera que vale la pena utilizar en el futuro, sistemas reconfigurables como en el cao e los FPGA (Field Programmable Gates Array). Así mimo, el gen digital puede ser entrenado usando algoritmo genéticos para la solución de tareas específicas, de modo que se propone que para el futuro desarrollar sistemas de reconocimiento del habla en FPGA entrenando el gen digital.

Se concluye finalmente, que el procesamiento digital de las señales correspondientes al habla exige para ser realizado de manera segmentada comprender la evolución de dichas señales en el tiempo, ya que si se usan tamaños de ventana demasiado grandes se omiten cambios locales, contrario a si se toman tamaños de ventanas demasiado pequeños pues se reflejan demasiado los cambios puntuales.

El tamaño del incremento entre ventanas influye directamente en los tiempos de repuesta de los algoritmos implementados y a su vez en la calidad de los resultados, con un incremento demasiado pequeño, el tiempo de respuesta es mayor y los resultados poco favorables. LPC es un método adecuado al tratamiento del habla ya que presenta una aproximación a la producción de la misma.

El algoritmo de alineación temporal dinámica (DTW) es ideal para el reconocimiento de señales de voz, porque trata de reducir las diferencias temporales naturales del habla y genera un modelo adecuado de la producción de la misma en el tracto vocal.

El algoritmo DTW ofrece buenos resultados para un conjunto pequeño de palabras a reconocer, si se requiere realizar el reconocimiento para un vocabulario extenso, esta solución no es la más óptima computacionalmente.

El entrenamiento con características de uno o pocos hablantes, hace que el reconocimiento de voz sea dependiente del hablante, para un reconocedor de voz general, se deben realizar estudios con una muestra considerable de distintas voces, tratando de analizar características generales.

Agradecimientos

Los autores agradecen a la Dirección General de Investigaciones de la Universidad de los Llanos por el apoyo financiero y al grupo de investigaciones en robótica de la misma Universidad. Al Laboratorio de Mecatrónica de la Universidad de Sao Paulo (EESC), así como a los Ingenieros Juan Fajardo y Rubén Darío Ángel por sus valiosas orientaciones y al grupo CIS (Control Inteligente de Sistemas) de la Universidad Nacional, por ser pionero en el desarrollo de aplicaciones del chip ADN emulado y el gene digital.

Referencias

Alberts B, Bray D, Johnson A, Lewis J, Raff M, Roberts K, Walter, P .“Essential cell biology.” Garland Publishing, 1998.

Alvarado J. “Reconocimiento De Palabras Aisladas Utilizando MFCC y Dinamic Time Warping”. Universidad Nacional De Trujillo. 2008.

Bernal J, Bobadilla S, Gómez P. “Reconocimiento de voz y fonética acústica”. Ed RA-MA. 2000.

Borrero H, Delgado A. “Evolución de chip ADN emulado con algoritmo genético en FPGA para control de navegación de un robot móvil” Orinoquia. 2008;12(1):117-129.

Bregón A, Alfonso A. “Un Sistema De Razonamiento Basado En Casos Para La Clasificación De Fallos En Sistemas Dinámicos”. Universidad de Valladolid. 2005.

Campbell AM, Heyer LJ.: “Discovering genomics, proteomics, and bioinformatics, Pearson Education”, 2003.

Farfan, A., Herreño J. y Delgado, A.: “Gene digital y chip ADN electrónico: aplicaciones en robótica móvil,” 3rd Colombian Works hopon Robotics and Automation (CWRA), Universidad Tecnológica de Bolivar, Cartagena, Agosto 21- 22, 2007.

Keogh, J. “Derivative Dynamic Time Warping”. 2001.

Prieto, J., Ramos, O. y Delgado, A.: “Diseño de un gene digital en FPGA y MATLAB con aplicaciones en robótica móvil,” XIII Taller Iberchip IWS-2007, Lima – Perú, Marzo 14-16, 2007.

Rabiner, L., Schafer, R.: “introduction to Digital Speech Procesig”, Now publishers Inc., 2007.