RECONOCIMIENTO DE COMANDOS POR VOZ CON MÁQUINAS DE SOPORTE VECTORIAL A TRAVÉS DE BANDAS ESPECTRALES

Voice commands recognizing using support vector machines and spectral bands

RESUMEN

En este artículo se propone una metodología para reconocimiento de comandos hablados, utilizando Máquinas de Soporte Vectorial (SVM). Esta tarea es importante en sistemas autónomos y semi-autónomos por ser un medio natural y práctico de interacción en situaciones en las cuales el individuo tiene ocupadas las manos, en ambientes con poca visibilidad o cuando el contacto táctil es poco práctico o imposible.
Como ejemplo de aplicación, las señales de voz se caracterizan empleando bandas espectrales y luego se clasifican usando SVM. La metodología se prueba con la identificación de vocales de un locutor con resultados promedio del 98% de acierto.

PALABRAS CLAVES: bandas de frecuencia, máquinas de soporte vectorial, reconocimiento de comandos.

ABSTRACT

In this paper a voice command recognizing methodology using Support Vector Machines (SVM) is proposed. This is an important task in autonomous and semi-autonomous systems, because it is a natural and useful interaction way, especially in situations where there are special limitations as low visibility, low or any possibility of physical contact, among others.
As application example, voice registered signals are characterized by using spectral bands and next these are classified by using SVMs. The proposed methodology is tested in vowels identification, having obtained a 98% of average successful results

KEYWORDS: *Frequency bands, support vector machines, voice commands recognizing*

GERMÁN ANDRÉS MORALES ESPAÑA *
Ingeniero Electricista, M.Sc (c)
Universidad Industrial de Santander
german.morales.e@gmail.com

RENÉ ALEXANDER BARRERA CÁRDENAS*
Ingeniero Electrónico, M.Sc (c)
Universidad Industrial de Santander
abarrera@uis.edu.co

JUAN JOSÉ MORA FLÓREZ **
Ingeniero Electricista, Ph.D.
Profesor asistente
Universidad Tecnológica de Pereira
jjmora@utp.edu.co

*** GRUPO DE INVESTIGACIÓN EN SISTEMAS DE ENERGÍA ELÉCTRICA - GISEL**

**** GRUPO DE INVESTIGACIÓN EN CALIDAD DE ENERGÍA ELÉCTRICA Y ESTABILIDAD - ICE[3]**

1. INTRODUCCIÓN

La interacción hombre–máquina por medio de la voz cubre muchas áreas de investigación, en [1] se resalta entre otras, el reconocimiento del habla, la síntesis e identificación de discurso, la verificación e identificación del hablante y la activación por voz (comandos) de sistemas robóticos. Todas estas áreas concurren en la utilización de métodos estadísticos y técnicas de inteligencia artificial basadas en el reconocimiento de patrones, las cuales han tenido un gran auge en los últimos tiempos [1], [2]. Los beneficios de este tipo de interacción frente a las demás, cobran especial importancia en situaciones en las cuales el individuo tiene ocupadas las manos, en ambientes donde se tiene poca visibilidad o cuando el contacto táctil es poco práctico o imposible [3].

El uso de comandos verbales para la activación de sistemas robóticos en particular, ofrece un amplio campo de aplicación [1], [3]-[7]. Más recientemente en [8], se plantea un prototipo del VOIC (Voice Operated Intelligent Wheelchair), una silla de ruedas inteligente operada por comandos verbales.

Reconocer el discurso es natural y simple para las personas, pero es un trabajo complejo para las computadoras e, históricamente, no se ha encontrado solución definitiva al problema [8]. En esta área cabe destacar las técnicas basadas en los modelos ocultos de Markov (Hidden Markov Model, HMM), los modelos de mezclas Gaussianas (Gaussian Mixture Model, GMM) y las redes neuronales artificiales (ANN), esta última ha mostrado un camino alentador [2].

En este artículo se presenta una metodología para el reconocimiento de comandos hablados empleando máquinas de soporte vectorial (SVM) como técnica de clasificación. El éxito de las SVM en muchas aplicaciones frente a las redes neuronales (Neural Networks, NN), revela la posibilidad de encontrar una alternativa ante los métodos empleados hoy [9]. Se identifican las cinco vocales habladas por un único locutor, un problema típico en el reconocimiento de comandos hablados [7]. La caracterización de las palabras se hace por medio de bandas de energía en el

Fecha de Recepción: 06 Septiembre de 2007
Fecha de Aceptación: 26 Noviembre de 2007

dominio de la frecuencia. El análisis de Fourier es usado como base en el proceso de caracterización de la voz, el cual presenta resultados satisfactorios.

Como contenido de este artículo, en la sección dos se presentan los fundamentos básicos de las máquinas de soporte vectorial. En la sección tres se presenta la metodología propuesta para el reconocimiento de comandos. En la sección cuatro se muestran resultados de la metodología propuesta. Finalmente, en la última parte se presentan las conclusiones derivadas de esta investigación.

2. MÁQUINAS DE SOPORTE VECTORIAL

Para problemas de clasificación simples, la teoría estadística de aprendizaje puede identificar con mucha precisión los factores a tener en cuenta para un aprendizaje exitoso, pero las aplicaciones reales demandan el uso de modelos y algoritmos mas complejos (ej. redes neuronales, técnicas Bayesianas, etc), que son difíciles de analizar.

Las SVM, a diferencia del método Bayesiano, presentan la ventaja de no requerir ningún tipo de hipótesis sobre la densidad de probabilidad de los rasgos, mientras que sobre las redes neuronales ofrecen la ventaja de ser convenientes en términos de la dimensionalidad del problema. Como se verá más adelante, la arquitectura de las SVM sólo depende de la constante de penalización *C*, y el parámetro σ de la función kernel para el caso de la función de base radial (RBF), tal como se muestra en la sección 2.3 [10]. Esta facilidad de configuración evita la selección de requerimientos sobre parámetros exclusivos de arquitectura, tales como número de nodos y capas, tipo de conexión entre capas, entre otros.

2.1. Clasificación lineal

Las SVM están basadas en hiperplanos que separan los datos de entrenamiento en dos subgrupos que poseen una etiqueta propia. En medio de todos los posibles planos de separación entre las dos clases etiquetadas $y \in \{-1,+1\}$, existe un único *hiperplano de separación óptimo* (OSH), de forma que la distancia entre el hiperplano óptimo y el patrón de entrenamiento más cercano sea máxima, con la intención de forzar la generalización de la máquina de aprendizaje [10], [11]. El OSH se expresa tal como se presenta en (1).

$$P_0 : \langle w, x \rangle + b = 0 \tag{1}$$

En este problema, lo que se desea es maximizar el margen, tal como se presenta en la figura 1.

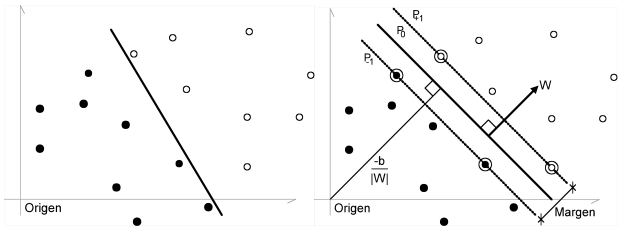

Figura 1. Hiperplanos que separan correctamente los datos. El OSH de la derecha tiene un mayor margen de separación entre clases, por lo tanto se espera una mejor generalización.

La función decisión $f_{w,b}(x_i) = y_i$, se puede definir como el signo que resulta de evaluar un dato en la ecuación del OSH (2), tal como se presenta en la ecuación (2).

$$f_{w,b}(x_i) = \operatorname{sign}\left(\langle w, x_i \rangle + b\right) \tag{2}$$

Si existe un hiperplano como se muestra en la figura 1, se dice que los datos son *linealmente separables*

.

2.2. Clasificación con margen débil

En casos donde existen datos de entrada erróneos, ruido o alto solapamiento de clases en los datos de entrenamiento, se puede afectar el hiperplano clasificador óptimo. Por esta razón se cambia un poco la perspectiva y se busca el mejor hiperplano clasificador que pueda tolerar ruido en los datos de entrenamiento (ver figura 2), introduciendo la variable de relajación que se presenta en la ecuación (3).

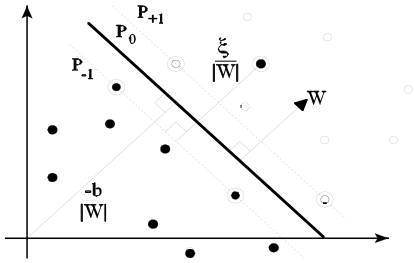

Figura 2. Hiperplano de separación permitiendo ruido

$$\xi_i \geq 0, \forall i \tag{3}$$

Con el objeto de definir de forma única el hiperplano óptimo (forma canónica), se deben añadir las restricciones, tal como se presenta en la ecuación (4).

$$y_i\left(\langle w, x_i \rangle + b\right) \geq 1 - \xi_i, \forall i \tag{4}$$

2.3. Caso no lineal

Las SVM no lineales, tienen la posibilidad de mapear el espacio de entrada en otro de representación de dimensión alta. En este nuevo espacio, los datos son linealmente separables y luego construye un OSH sobre este último, cuya representación en el espacio de entrada es una función de separación no lineal, tal como se representa gráficamente en la figura 3.

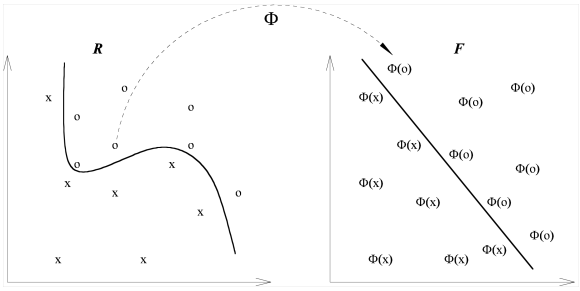

Figura 3. Transformación del espacio de entrada en otro de dimensión más alta donde las clases tienen una separación lineal.

La generalización de la Máquina de Soporte Vectorial a funciones de decisión no lineales consiste en mapear el espacio de entrada sobre un espacio de representación de dimensión alta usando una función no lineal elegida a priori. Esta función es el "kernel" (Φ), relaciona los datos de entrada $x_i \in R^N$ con un espacio de mayor dimensión y en el cual esté definido el producto punto, conocido como espacio característico (*F*), tal como se define en la ecuación (5) [10].

$$\Phi : R^N \rightarrow F \tag{5}$$

Así, de la función en la ecuación (1), la cual depende del producto punto de los vectores en el espacio de entrada, aplicando la transformación de la ecuación (5) se obtiene una función que depende del producto punto de los vectores en el espacio característico, tal como se presenta en la ecuación (6).

$$g(x) = \Phi(w)\Phi(x) + b \tag{6}$$

Se debe definir una función que sea el producto punto de los vectores en el espacio característico, tal como se presenta en (7).

$$k(u,v) = \Phi(u)\Phi(v) \tag{7}$$

Considerando que (F) es de alta dimensión, el lado derecho de la ecuación (7) es costosa en términos computacionales, sin embargo existe una función "kernel" (k), que se puede evaluar eficazmente y demostrar que corresponde a un trazado de (Φ) en un espacio que abarca todos los productos punto [11]. Los kernels más utilizados son el polinomial presentado en la ecuación (8), el de función de base radial (RBF) en (9) y el sigmoide en (10).

$$k(\vec{u},\vec{v}) = \left(\langle \vec{u},\vec{v} \rangle + a\right)^d \tag{8}$$

$$k(u,v) = e^{\left(-\frac{\|u-v\|^2}{\sigma^2}\right)} \tag{9}$$

$$k(u,v) = \tanh\left(\kappa\, u,v + \Theta\right) \tag{10}$$

donde *a*, *d*, σ, κ y Θ son los parámetros de cada función kernel [10].

En resumen, el hiperplano óptimo en forma canónica de margen débil se halla solucionando el problema de optimización restringida dado por la ecuación (11), sujeto a (3) y (4).

$$\min_{w,b} \frac{1}{2}\, w,w + C\sum_{i=1}^{N} \xi_i \tag{11}$$

Utilizando los multiplicadores de Lagrange, el teorema dual de Wolfe, se obtiene el problema final de optimización dado por (13).

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(\vec{x}_i, \vec{x}_j) \tag{12}$$

Sujeto a: $0 \le \alpha_i \le C, \forall i$ y $\sum_{i=1}^{N} \alpha_i y_i = 0$

donde α_i son los multiplicadores de Lagrange.

La ecuación del OSH y la función decisión se puede expresar como (13) y (14)

$$g(x) = \sum_{i,j=1}^{N} (\alpha_i y_i k(x_i, x)) + b \tag{13}$$

$$f(x) = \text{sign}\left(\sum_{i,j=1}^{N} (\alpha_i y_i k(x_i, x)) + b\right) \tag{14}$$

Para resolver el problema de multiclasificación (más de dos clases), se construye una función clasificadora global a partir de un conjunto de funciones biclasificadoras. Existen técnicas de descomposición y reconstrucción que permiten a las SVM manejar problemas de multiclasificación, con mayor simplicidad y/o menor tiempo de respuesta que una SVM generalizada para multiclasificación [10].

Para obtener resultados de clasificación satisfactorios, es necesario escoger adecuadadamente el "parámetro de penalización al ruido" (*C*) y el "parámetro del nivel de no linealidad" (σ para el caso del kernel RBF) de las SVM. La validación cruzada y la búsqueda en malla realizan esta tarea de manera exitosa, abordando a la vez el problema de sobreentrenamiento.

3. METODOLOGÍA PARA EL RECONOCIMIENTO DE COMANDOS

3.1. Adquisición de las señales

Un comando verbal corresponde a una palabra que es pronunciada por el emisor y capturada por algún aparato que digitaliza la señal audible. Para capturar la señal de audio se usa una frecuencia de muestreo de 8000Hz, pero solo se utilizan componentes de frecuencia menores a 3000Hz, por considerarse que en este rango se presentan las características más relevantes de la voz humana.

En el proceso de obtención de la señal de audio se involucran múltiples factores que agregan ruido a la señal, tales como ruido propio del aparato receptor y comúnmente sonidos ambientales que distorsionan la señal haciendo más difícil su identificación. Considerando lo anterior, se debe grabar las voces en distintos momentos del día o diferentes días, para así tener distintos escenarios de ruido y también considerar los distintos registros de la voz, ya que no siempre se habla con el mismo espectro.

3.2. Pre-procesamiento de la voz

En esta etapa, a cada señal de voz se le saca la transformada rápida de Fourier (FFT), y se trabaja únicamente las magnitudes de cada componente de frecuencia. La señal en el dominio de la frecuencia se normaliza dividiendo por la componente de mayor amplitud. De esta manera se independiza la señal del nivel de energía con el que fue grabada, es decir, si varias señales fueron grabadas a distinto nivel de energía (unas con mayor volumen o simplemente unas palabras se han dicho más fuerte que otras), quedan en proporciones a la componente de frecuencia de mayor energía (amplitud), haciéndolas comparables.

De igual manera, se escalan las señales por banda, es decir, para la banda *i*, todas las señales se dividen por la componente de mayor amplitud de esta banda, con la intención de no perder las relaciones de proporcionalidad entre señales. El mismo proceso debe realizarse en cada una de las *N* bandas.

3.3. Caracterización de los comandos verbales

Para los problemas de clasificación, es necesario adquirir rasgos característicos de los distintos comandos de voz, para así diferenciarlos. Se propone una caracterización por medio de bandas de frecuencia en todo el ancho de banda de análisis (3000 Hz). En primer lugar se divide el espectro de frecuencias en *N* bandas y luego se obtiene por cada banda un patrón estadístico que corresponde a un descriptor, por lo tanto por cada comando se tendrán *N* descriptores. El patrón característico de cada banda se define como el promedio de todas las componentes de la banda cuya magnitud sea superior a la media más una desviación estándar (En pruebas preliminares se establece que este estadístico presenta mejores resultados que otras medidas de tendencia como la media o el máximo, entre otros), y se calcula como se presenta en (15).

$$\overline{Xs}_i = \frac{\sum_{k=1}^{Nk} X_{ki}}{Nk} \tag{15}$$

donde $\overline{Xs}_i$ es el descriptor (característica) de la banda *i*, y X_{ki} son las magnitudes de las componentes espectrales que pertenecen a la banda *i* que cumplen con (16).

$$X_{ki} \geq \overline{X}_i + S_i \tag{16}$$

donde $\overline{X}_i$ y S_i son la media aritmética y la desviación estándar de las magnitudes espectrales que pertenecen a la banda *i*.

En caso que la distribución de los datos sea sesgada hacia la derecha, pueden no existir datos que cumplan (16), entonces se calcula un nuevo X_{ki} que cumpla con (17).

$$X_{ki} \geq \overline{X}_i \tag{17}$$

Con esta caracterización se desea discriminar un poco el ruido blanco, debido a que se discriminan componentes de baja energía.

La selección del número de bandas N, es propia de cada conjunto de comandos y se determina por validación cruzada, que básicamente consiste en la búsqueda exhaustiva de la mejor división, a partir del menor error de validación cruzada obtenido en distintas cantidades de bandas [10]. En las figuras 4 y 5 se muestra el comportamiento de los descriptores para el caso de reconocimiento de vocales con 10 y 30 bandas respectivamente.

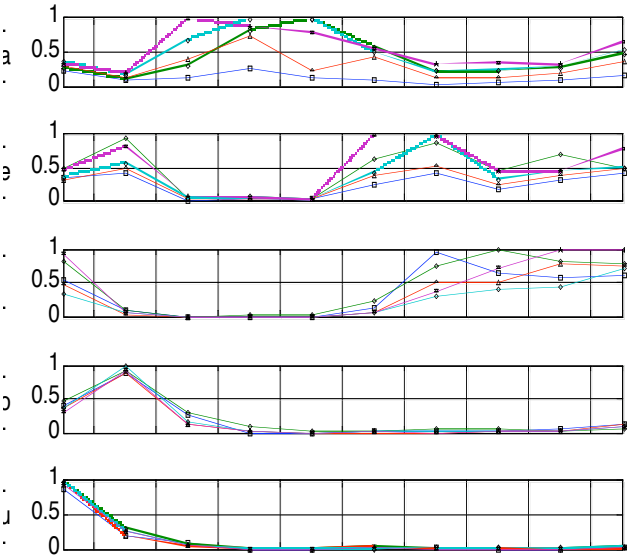

Figura 4. Comportamiento de los descriptores propuestos ante las vocales con 10 bandas.

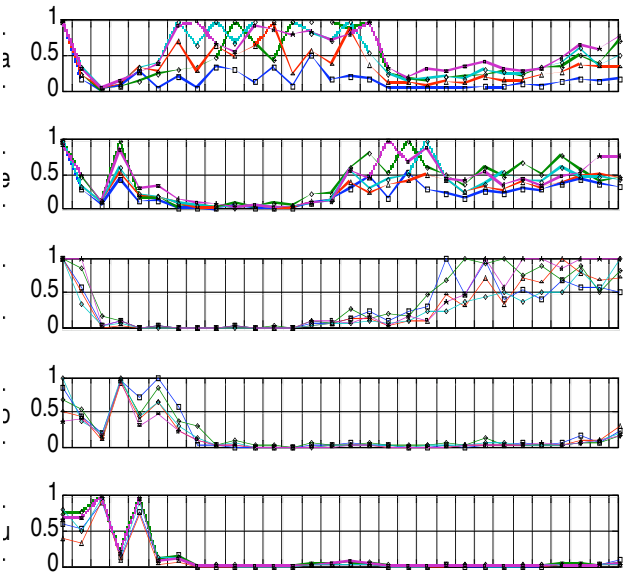

Figura 5. Comportamiento de los descriptores propuestos ante las vocales con 30 bandas.

En la figura 4 se observan los descriptores de 10 bandas adquiridas del espectro, donde cada subfigura contiene los descriptores adquiridos de cada vocal, cada curva es una vocal pronunciada por la misma persona en instantes de tiempo distintos, lo cual significa que el ruido es distinto. En una subfigura se observa la gran variación de los descriptores por cada vocal, lo cual significa que no se pronuncia siempre de la misma manera, pero a pesar de la variedad se nota una clara diferencia entre las cinco vocales. Lo anterior significa que con una herramienta de clasificación se puede resolver este problema, con resultados satisfactorios. En la figura 5 se observa los descriptores de 30 bandas, estos tienen un comportamiento análogo al de 10 bandas, presentado en la figura 4.

El conjunto de descriptores propuestos se puede aplicar a cualquier conjunto de palabras. En caso de que dos palabras presenten comportamientos similares en los descriptores (lo cual puede provocar confusión en la herramienta clasificadora), se debe descomponer la señal con la transformada de Fourier, para obtener una parte real y una imaginaria.

A manera de ejemplo, en caso de 10 bandas, se obtienen los descriptores de la forma propuesta anteriormente en las ecuaciones (15) a (17), pero no se trabaja con el número complejo, sino con la parte real (10 descriptores). El mismo procedimiento se repite para la parte imaginaria, obteniendo de esta manera 20 descriptores de las 10 bandas

3.4. Entrenamiento de las SVM

Las máquinas de soporte vectorial necesitan la definición a priori tanto del parámetro de penalización *C*, como de la función kernel y sus respectivos parámetros. La función Kernel establecida es la RBF, siendo la que mejores características demuestra según se puede observar en [11]. El parámetro de la función kernel y el parámetro de penalización se determinan por el método de búsqueda en malla y validación cruzada por sus ventajas de: fácil implementación, evita el sobre-entrenamiento y alta competitividad con respecto a otros métodos [10].

3.5. Prueba

Con el fin de obtener la precisión final del modelo, se realiza una prueba final con datos desconocidos (Estos datos no deben ser utilizados en la etapa de entrenamiento de la SVM). La cantidad de datos de prueba debe ser por lo menos el 20% del total de los datos y se extraen justo antes de la etapa de entrenamiento de la SVM.

Para medir el desempeño de la metodología, se considera un porcentaje de acierto a partir de los datos de prueba, el porcentaje de exactitud se define como (18).

$$\%\,\text{exact} = \frac{\#\,\text{datos identificados correctamente}}{\#\,\text{total de datos}} \qquad (18)$$

4. PRUEBAS Y RESULTADOS

La metodología propuesta para el reconocimiento de comandos verbales es puesta a prueba en el reconocimiento de las vocales. Los datos de entrenamiento y prueba provienen del mismo locutor (mujer), donde se trabaja con un total de 100 datos (20 por vocal). Se utilizan 75 datos para el entrenamiento (15 por vocal) y 25 para prueba (5 por vocal).

Se hacen pruebas con 5, 10, 15, 20,… y 50 bandas. Los errores de validación cruzada (VC) y de prueba, se muestran en la tabla 1.

No. de bandas	Exact. VC (%)	Exact. Prueba (%)
5	95,00	92,00
10	98,75	100,00
15	93,75	96,00
20	98,75	100,00
25	95,83	100,00
30	96,25	100,00
35	97,08	96,00
40	94,58	100,00
45	96,25	100,00
50	96,25	100,00
Promedio	96,25	98,40

Tabla 1. Porcentajes de acierto de validación cruzada (VC) y prueba ante distintos números de bandas.

En la tabla 1, se observa que en los datos de prueba únicamente con 5, 15 y 35 bandas no se obtiene una clasificación perfecta a diferencia de las demás bandas. El buen desempeño del método es debido a la clara

diferencia de las palabras con la caracterización propuesta, como se observa en las figuras 4 y 5. Según los errores de validación cruzada obtenidos por cada banda, con 10 y 20 bandas se obtiene el mejor desempeño.

5. CONCLUSIONES

En este artículo se presentó un método para identificar palabras habladas, basado en la técnica de Máquinas de Soporte Vectorial para clasificación. Las señales de voz fueron caracterizadas empleando bandas en el dominio de la frecuencia.

Las pruebas realizadas para el reconocimiento de vocales de un único locutor, permiten obtener resultados del 92% en el peor de los casos, 100% en el mejor y 98% como promedio.

Finalmente, con la metodología propuesta se presenta una alternativa sencilla para resolver el problema de reconocimiento de comandos en entornos con limitaciones especiales.

6. BIBLIOGRAFÍA

[1] T. Tran, Q.P. Ha, G. Dissanayake. "New Wavelet-Based Pitch Detection Method for Human-Robot Voice Interface*". Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings IEEE/RSJ Internacional Conference. 2004.*

[2] A. Cichocki, R. Unbehauen. "*Neural Networks for Optimization and Signal Processing*" Ed. John Willey and Sons, 1993.

[3] N. Botros, M.Z. Deiri, P. Hsu. "Automatic Voice Recognition Using Artificial Neural Network Approach". Circuits and Systems - *IEEE. Proceedings of the 32nd Midwest Symposium. 1990*

[4] R. Seireg, A. Barbour, "A New Algorithm for Pattern Recognition of Voices". *Proceedings of the 35th Midwest Symposius on Circuits and Systems, IEEE 1992*

[5] M Bodruzzaman, K. Kuah, T. Jamil, C. Wang, X. Li. "Parametric Feature-Based Voice Recognition System Using Artificial Neural Network". *Southeastcon '93. Proceedings IEEE. Tennessee State University. 1993*

[6] J. Plaza, D. Báez, L. Guerrero, J. Rodríguez. "A Voice Recognition System for Speech Impaired People". *Proceedings of CONIELECOMP'04 - 14th International conference on Electronics, Communications and Computers, IEEE 2004*

[7] S. Kummar, D. Kant, M. Alemu, M. Burry. "EMG Based Voice Recognition". *Intelligent Sensors, Sensor Networks and Information Processing Conference, IEEE 2004*

[8] J. Mora. "*Localización de Faltas en Sistemas de Distribución de Energía Eléctrica usando Métodos Basados en el Modelo y Métodos Basados en el Conocimiento*". Tesis doctoral. Universidad de Girona. 2006.

[9] G. Pacnik, K. Benkič and B. Brečko. "*Voice Operated Intelligent Wheelchair - VOIC*". Faculty of Electrical Engineering and Computer Science, Institute of Robotics, Maribor, Slovenia. IEEE ISIE. Croatia 2005

[10] G. Morales, A. Gómez, *"Estudio e implementación de una herramienta basada en Máquinas de Soporte Vectorial aplicada a la localización de fallas en sistemas de distribución*". Tesis de grado, Universidad Industrial de Santander, Colombia 2005.

[11] V. Vapnik "*The nature of Statistical Learning Theory*" Second Edition, Springer Verlag, 2000.