IMPLEMENTACIÓN DE UNA METODOLOGÍA PARA LA DETECCIÓN DE COMANDOS DE VOZ UTILIZANDO HMM

IMPLEMENTATION OF A METHODOLOGY FOR DETECTING VOICE COMMANDS USING HMM

William Acosta Bedoya[1], Milton Sarria Paja[1], Leonardo Duque Muñoz[1]

[1] Instituto Tecnológico Metropolitano, Medellin- Colombia. williamacostabedoya@gmail.com, miltonsarria@gmail.com, leonardoduque@itm.edu.co.

Recibido: Marzo 5 de 2012
Aceptado: Mayo 21 de 2012
*Correspondencia del autor . Instituto tecnológico Metropolitano Campus Fraternidad, Barrio Boston, CLL 54 A 30-01, oficina 102 (laboratorio MIRP), Medellín Antioquia. Leonardo Duque Muñoz. Teléfono 3212162924 E- mail: leonardoduque@itm.edu.co.

RESUMEN

En este artículo se plantea una solución que permite la identificación de comandos de voz, con un diccionario reducido. Se construyó una base de datos con los comandos: adelante, atrás, derecha, izquierda y pare. Se realizó la caracterización de los comandos de voz con características frecuenciales obtenidas de la transformada de Fourier en tiempo corto y se realiza el reconocimiento de dichos comandos con clasificadores estocásticos GMM y HMM. Los resultados obtenidos con este método presentan una alta tasa de acierto en el orden de 98-100% para el reconocimiento de estos comandos. Una posible aplicación de esta herramienta será asumir el mando de los motores acondicionados a una silla de ruedas. Lo que daría una solución al problema de movilidad en personas con discapacidad motora, con lo cual se logrará dar una mayor autonomía en su movilidad; mejorando así su calidad de vida.

Palabras clave: Coeficiente cepstral de Mel, GMM, HMM, reconocimiento de voz.

ABSTRACT

This article presents a solution that allows the identification of voice commands. We built a database with the commands: atrás, adelante, derecha, izquierda, pare and silla. The characterization of the voice commands was conducted with frequency characteristics obtained from the Short Time Fourier Transform, and the recognition of these commands with stochastic GMM and HMM classifiers. The results obtained with this method have a high performance in the order of 98-100% for the recognition of these commands. One possible application of this tool will be the coupling of the recognition system with the engine control conditioned to a wheelchair. This would represent a solution to the problem of mobility in people with motor disabilities, thereby giving greater autonomy achieved in their mobility, improving their quality of life.

Key words: Cepstral Coeficient, GMM, HMM, voice command recognition.

INTRODUCCIÓN

El campo del procesamiento de la señal de voz ha sido sujeto de estudio intenso en las últimas tres décadas, debido principalmente a los avances en las técnicas de procesamiento digital de señales y reconocimiento de patrones, además de la capacidad de proceso de los sistemas de cómputo. Su objetivo final es desarrollar interfaces hombre-máquina, que permitan al ser humano comunicarse de manera natural con los diferentes dispositivos electrónicos de uso diario (1-5). La comunicación entre personas tiene en cuenta gran diversidad de conocimiento, lo que permite sortear dificultades como el ruido ambiental, el acento y la concatenación de palabras, además de asuntos gramaticales (6). El reconocimiento automático del habla o ASR (Automatic Speech Recognition), es un campo difícil de tratar, debido principalmente a las variaciones de fonación (los locutores no hablan igual), las ambigüedades en la señal acústica (no toda la información presente está relacionada con el habla), la falta de cuidado del hablante, la variación en la frecuencia y duración de los fonemas, y la presencia de ruido o interferencias (7- 10). Los sistemas actuales están restringidos a ambientes controlados, a ser utilizados con un grupo de hablantes reducido o requieren de posicionamiento especial del micrófono resultando en interfaces poco naturales.

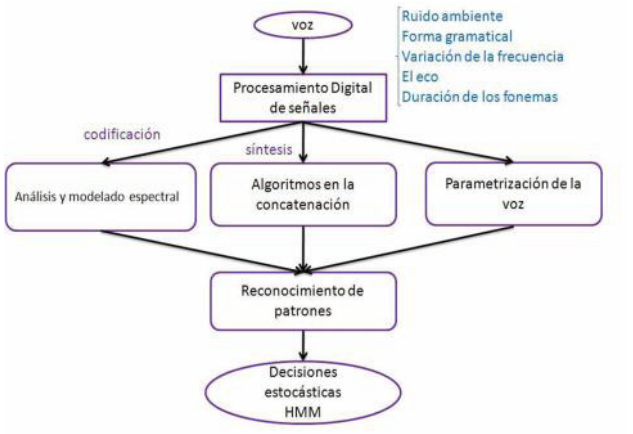

Figura 1. La representación del procesamiento de la señal de voz.

El procesamiento de voz se divide en tres temas de interés: codificación, síntesis y reconocimiento del habla, los cuales aun son problemas abiertos de investigación. La codificación de voz fue ampliamente estudiada en la década de los ochenta y principios de los noventa, los esfuerzos se concentraron en el desarrollo de algoritmos de parametrización de la señal (11, 12), la extracción de la frecuencia fundamental (13), y el análisis y modelado de la envolvente espectral (14). Mientras que en la síntesis, el gran paso se produjo a mediados de los noventa, cuando se introdujeron los algoritmos basados en la concatenación de unidades pregrabadas (12- 17). Con respecto a los sistemas de reconocimiento del habla, los dos bloques fundamentales consisten en un sistema de parametrización de la señal de voz y un sistema de reconocimiento de patrones (1, 7, 18). Aunque el grado de desarrollo alcanzado en los sistemas de codificación es satisfactorio, no lo es para la síntesis ni para el reconocimiento del habla. De cara al problema de reconocimiento automático del habla se han propuesto estrategias basadas en varias aproximaciones, pero la técnica de reconocimiento de patrones que mejores resultados ha ofrecido para descifrar la señal de voz hasta el momento, es aquella basada en la teoría de decisión estadística. Esta técnica permite encontrar la secuencia de patrones que tiene la mayor probabilidad de estar asociada a la secuencia de observaciones acústicas de entrada. Los modelos ocultos de Markov o Hidden Markov Models (HMMs por sus siglas en inglés) son modelos estadísticos cuya salida es una secuencia de símbolos o cantidades y poseen mejores tasas de identificación de habla distorsionada y normal que aquellas basadas en plantillas o en otras aproximaciones (19). Entre las razones de su popularidad sobresale el hecho de que la señal de voz puede ser vista como una señal estacionaria a trozos, es decir, se puede asumir que en un corto tiempo la señal puede ser modelada como un proceso estacionario (18, 20). Además, los HMMs pueden ser entrenados automáticamente, y es factible su sistematización (18).

En este trabajo, se utiliza la transformada de Fourier en tiempo corto (STFT), la cual permite entregar un valor en un espacio de representación que puede integrarse con herramientas como: SVM, GMM y HMM. Los dos primeros presentan inconvenientes en el manejo de las señales dinámicas, lo cual se convierte en fortaleza para los HMM; a su vez esta técnica presenta la dificultad en hallar un punto óptimo de inicialización. Se presentan posibles soluciones con la utilización de métodos como el de K-media, J-medias, para obtener una señal de control que puede ser aplicada a una máquina de estados para el reconocimiento de los comandos de voz.

MATERIALES Y MÉTODOS

A. Definición de modelos ocultos de Markov

Un Modelo Oculto de Markov para una clase en particular está definido por el conjunto de parámetros $\lambda m = \{ A^{(m)}, B^{(m)}, \pi^{(m)}\}$, donde $A^{(m)}$ es la matriz de transición de estados, y está compuesta por las probabilidades discretas $a_{ij}^{(m)}$ que representa la probabilidad de pasar del estado si al estado s_j , $B^{(m)}$ corresponde a la función densidad de probabilidad de observación, $\pi^{(m)}$ corresponde al vector de probabilidad de estado inicial (18).

Existen dos formas de distribuciones de salida que pueden ser consideradas:

La primera es una suposición de observación discreta donde se asume que una observación es una de nv posibles símbolos de observación $v = \{v_k : k = 1, ..., n_v\}$

La segunda forma de distribución de probabilidad, es considerar una mezcla de M funciones de distribución para cada estado. Convencionalmente las funciones utilizadas son Gaussianas multivariadas, debido a sus propiedades y a la facilidad que representan y a que toda la descripción matemática está en función de éstas. En este caso,

$$b_j(\varphi_n) = \sum_{l=1}^{M} c_{jl} N\left[\varphi_n, \mu_{jl}, \Sigma_{jl}\right],$$

Donde μ_{jl} y Σ_{jl} son el vector de medias y la matriz de covarianzas respectivamente de la componente normal *ln*, el estado *j*, y C_{jl} es el peso que pondera la componente Gaussiana l del estado *j* y φ_n la observación en el tiempo t_n .
Los modelos que asumen la primera forma de distribución, son llamados modelos ocultos de *Markov* discretos, mientras que los que asumen la segunda forma de distribución de salida son llamados modelos ocultos de *Markov continuos*.

El desarrollo de los modelos ocultos de Markov está relacionado con tres tareas estadísticas:

1. Evaluación: Dada una secuencia de observación $\varphi = \{\varphi_1,...\varphi_{n\varphi}\}$ con longitud n_φ y el modelo λ, cómo calcular de manera eficiente la probabilidad $P(\varphi \mid \lambda)$ de la secuencia de observación.

2. Decodificación: Dada una secuencia de observación $\varphi = \{\varphi_1,...\varphi_{n\varphi}\}$ con longitud n_φ y el modelo λ, cómo escoger de forma óptima la correspondiente secuencia de estados $\Theta = \{\Theta_0, \Theta_1,..., \Theta_{n\varphi}\}$ para un criterio de medida fijado a priori.

3. Entrenamiento: El ajuste de los parámetros del modelo λ que brinden el máximo valor $P(\varphi \mid \lambda)$.

El entrenamiento de los HMM implica el ajuste de los parámetros de un modelo, tal que se extraiga la máxima información de las secuencias de observación. Entre los métodos conocidos están el criterio basado en la estimación de máxima verosimilitud (Maximum Likelihood Estimation - MLE), donde se optimiza la descripción del respectivo modelo para un conjunto dado de observaciones (función de verosimilitud), sin tener relación explícita con el rendimiento del clasificador, por lo cual, este es un criterio de entrenamiento generativo. Por otro lado están los métodos de entrenamiento discriminativo, por ejemplo, la técnica de Máxima Información Mutua (Maximum Mutual Information – MMI), donde se busca optimizar la probabilidad a posteriori de los datos de entrenamiento y, por lo tanto, la separabilidad entre clases o el criterio de Mínimo Error de Clasificación (Minimum Classification Error - MCE) donde se minimiza el error de clasificación mediante la formulación de una función de error empírica.

B. Arquitectura de HMM

Un HMM puede ser representado como un grafo dirigido de transiciones/emisiones. La arquitectura específica que permita modelar de la mejor forma posible las propiedades observadas depende en gran medida de las características del problema. Las arquitecturas más usadas son:

1) Ergódicas o completamente conectadas en las cuales cada estado del modelo puede ser alcanzado desde cualquier otro estado en un número finito de pasos.

2) Izquierda-derecha, hacia adelante o Bakis las cuales tienen la propiedad de que en la medida que el tiempo crece se avanza en la secuencia de observación asociada φ, y en esa misma medida el índice que señala el estado del modelo permanece o crece, es decir, los estados del sistema van de izquierda a derecha. En secuencias biológicas y en reconocimiento de la voz estas arquitecturas modelan bien los aspectos lineales de las secuencias.

3) Izquierda-derecha paralelas, son dos arquitecturas izquierda-derecha conectadas entre sí.

C. Base de Datos

Una primera etapa, es la construcción de una base de datos propia; que tiene las siguientes características: velocidad de muestreo de 44100Hz, a 16 bit, el tiempo de grabación para cada palabra está fijado de forma constante en 1.2 S, en un formato tipo WAV; entregando una señal normalizada, para las palabras: **silla, adelante, atrás, derecha, izquierda y pare**. Cada clase (palabra) se ha almacenado en cien ocasiones, para obtener una base de datos de 600 palabras. Para realizar esta tarea; se dividió en subgrupos de 10 elementos de cada clase. En donde cada procedimiento de captura la señal esta bajo condiciones controladas de ruido, con distancias entre el locutor y el elemento de captura variando en el rango de 0,2 a 0,5 m, además se hace un desplazamiento angular de ± 30°.

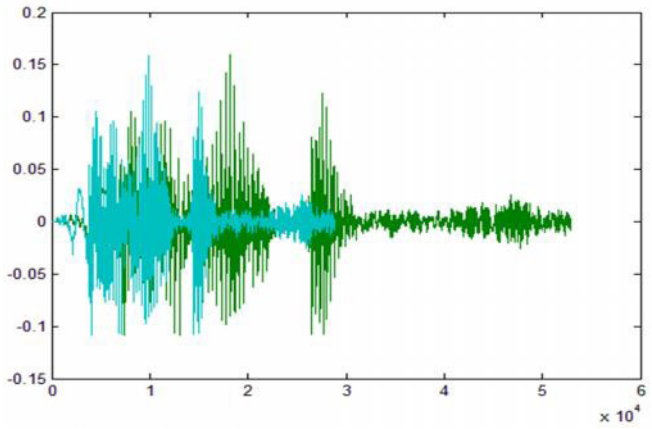

Figura 2. Representación en el tiempo de la señal para la palabra **adelante**.

RESULTADOS Y DISCUSIÓN

Todas las capturas fueron realizadas a un hombre mayor de 50 años. El recinto donde se realiza la grabación está ubicado en un sector residencial; con bajo tráfico vehicular. El elemento de captura; es el micrófono (estéreo) de la tarjeta de sonido incorporada en un PC portátil y el software utilizado, son los comandos del Matlab propios de esta aplicación.

A. Metodología Propuesta.

Se realizan tres etapas:

1. Preprocesamiento.
2. Procesamiento.
3. Reconocimiento de patrones de la voz.
4. Máquina de estados y potencia.

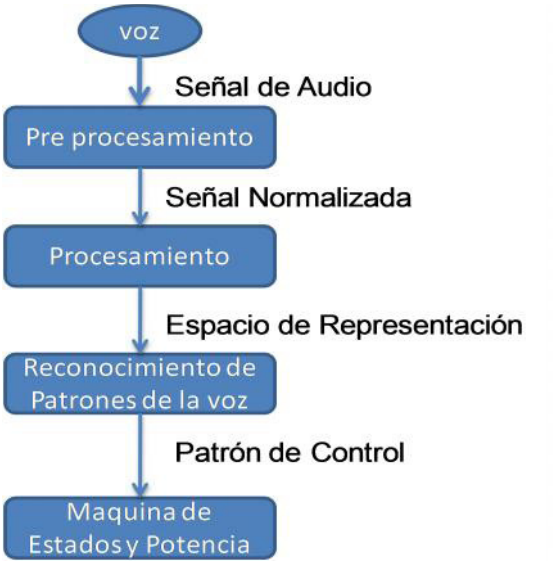

Figura 3. Diagrama a bloques del procesamiento de la señal de voz para su identificación.

Preprocesamiento. Se convierte el archivo de audio en una matriz (***Y***) de 2 columnas; por ser una señal "estéreo" y 52920 filas, de muestras que forman la señal. El siguiente paso consiste en realizar una ponderación de los datos, que permite generar un vector columna de 52920 filas. Luego se hace un filtrado mediante un filtro digital de primer orden con una frecuencia de corte en 6kHz.

Procesamiento: Los coeficientes de Mel son calculados realizando un ventaneo Hamming de 50ms con un traslapamiento de 25 ms entre ventanas, seguidamente la señal en el tiempo es representada mediante la STFT. A la representación obtenida, se le aplica un banco de filtros distribuidos en la escala de frecuencias Mel, que en el caso de estudio son 20, luego se calcula la energía de cada uno de los filtros y finalmente se aplica el logaritmo y la Transformada Coseno (DCT), obteniendo así los coeficientes MFCC y la derivada del mismo que generan la matriz de características de cada señal, que para este caso son 20 coeficientes MFCC y 20 que corresponden a su derivada.

Reconocimiento. Este programa tiene como núcleo el Toolbox de Kevin Murphy, para Matlab. Se inicia leyendo la celda ***X*** que contiene todas las matrices de características, se utiliza una validación cruzada; tomando el 70% para realizar el entrenamiento y el 30% para realizar la validación. A continuación se realiza la normalización, con el proceso de z-score. La siguiente etapa permite realizar las configuraciones iniciales, para

generar HMM con mezclas de Gaussianas, éstas son: número de estados que de forma arbitraria se le asigna un valor de 2, número de Gaussianas; que variará desde 2 hasta 20. La densidad inicial, la definición de la matriz de transición, definir los pesos de las Gaussianas, capturar la cantidad de características por cada palabra, se inicializa de forma aleatoria, se establece el máximo valor de iteraciones para el algoritmo EM y finalmente se elige el valor para el criterio de convergencia del EM.

El siguiente proceso es generar el modelo para la clase I, con datos que se hallan en (X{I}), para inicializar las gaussianas se deben poner los datos de forma matricial, mediante el Toolbox mixgauss_init se inicializa gaussianas con Kmeans, se realiza un acondicionamiento de los datos para trabajar con HMMall. Con los datos de la clase I, se entrena el modelo de la clase I con el algoritmo EM. Se generan seis(X) modelos que corresponden a los parámetros del modelo de la clase I.

La siguiente etapa es evaluar el logaritmo de la probabilidad de una palabra (cualquiera) con cada uno de los modelos, y a través de un sistema para toma de decisión se define la pertenencia o no a una clase determinada.

Máquina de estados y potencia. Este ítem hace referencia al sistema que va a ser controlado por medio de los comandos de voz reconocidos mediante la metodología propuesta.

Inicialmente se realizó un procedimiento con los algoritmos de GMM, la tabla I muestra los resultados de promedio y dispersión de los aciertos para las diferentes pruebas con esta metodología.
Para la verificación fueron realizadas 600 pruebas de cada clase de palabra, ejecutando variaciones de las mezclas gaussianas desde 10 hasta 200. De la figura 4 se observa que se presenta una buena repuesta alrededor de las 150 mezclas gaussianas; pero al aumentar el número de Gaussinas el sistema entrega de nuevo resultados con pocos aciertos. Bajo el proceso de GMM, se observa que la palabra con mayor cantidad de aciertos es ***adelante*** con un porcentaje de98%, y una dispersión del 2,4%. Por otra parte la palabra que mayor dificultad presenta es ***pare*** con un porcentaje de acierto de 87% con una dispersión de 9,3%.

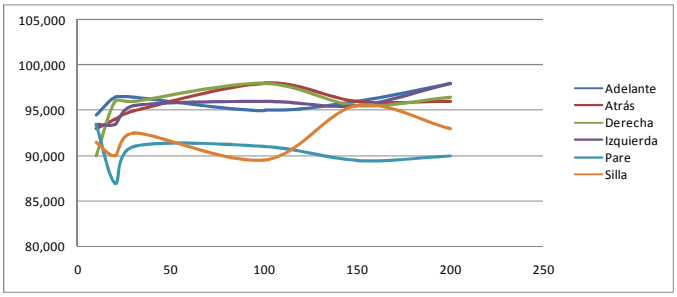

Figura 4. Representación de la relación entre porcentaje de aciertos Vs el número de gaussianas para GMM

La segunda metodología se realiza con algoritmos de HMM y utilizando de nuevo los coeficientes cepstrales en la escala de Mel, la tabla 2 recoge los resultados de este procedimiento.

Para la verificación fueron realizadas 1200 pruebas de cada clase de palabra, ejecutando variaciones de las mezclas gaussianas desde 2 hasta 20. Al analizar la figura 5, se observa una buena repuesta alrededor de 12 mezclas gaussianas y a partir de este valor se obtienen respuestas del 100%, es de anotar que existe un punto de inflexión cuando se tienen 8 mezclas gaussianas, y además se observa que la palabra con mayor cantidad de aciertos es derecha con un porcentaje de 100%, dispersión de 0,0, en 2 mezcla gaussianas. La palabra que presenta mayor dificultad es adelante, pues logra obte-

Tabla 1. Porcentaje de aciertos, variando el número de gaussianas utilizando GMM

Gaussinas	Adelante		Atrás		Derecha		Izquierda		Pare		Silla	
	Pro.	Dis.	Pro.	Dis.	Pro.	Dis.	Pro.	Dis.	Pro.	Dis.	Pro.	Dis.
10	94,5	4,7	93,0	9,0	90,0	8,7	93,5	5,9	93,5	3,9	91,5	7,1
20	96,5	3,2	94,0	6,6	96,0	3,7	93,5	7,1	87,0	9,3	90,0	5,0
30	96,5	3,2	95,0	7,1	96,0	4,9	95,5	3,5	91,0	5,8	92,5	7,8
100	95,0	3,2	98,0	2,4	98,0	2,4	96,0	4,9	91,0	4,9	89,5	7,9
150	96,0	5,4	96,0	3,7	95,5	3,5	95,5	4,7	89,5	4,2	95,5	4,2
200	98,0	2,4	96,0	4,9	96,5	6,3	98,0	3,3	90,0	5,0	93,0	7,8

Tabla 2. Porcentaje de acierto, variando el número de gaussianas utilizando HMM

Gaussinas	delante		Atrás		Derecha		Izquierda		Pare		Silla	
	Pro.	Dis.	Pro.	Dis.	Pro.	Dis.	Pro.	Dis.	Pro.	Dis.	Pro.	Dis.
2	98,8	0,5	99,3	0,5	100,0	0,0	99,1	1,0	99,8	0,3	99,8	0,3
5	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0	99,9	0,2	99,9	0,2
7	97,2	0,4	100,0	0,0	100,0	0,0	100,0	0,0	99,9	0,2	99,9	0,2
8	97,4	0,4	99,6	0,3	100,0	0,0	99,8	0,2	100,0	0,0	100,0	0,0
9	96,8	4,4	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0
10	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0
12	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0
15	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0
20	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0	100,0	0,0

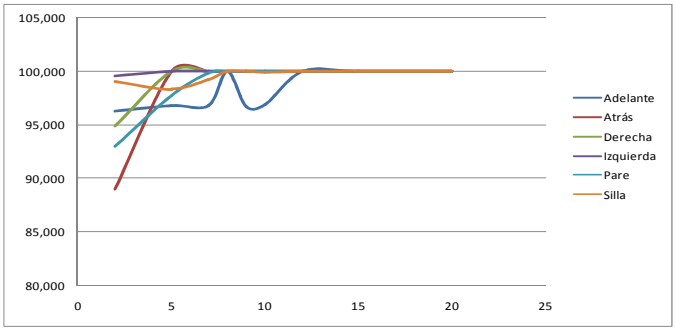

Figura 5. Representación de la relación entre porcentaje de aciertos Vs el número de gaussianas para HMM

ner el 100% de aciertos y una dispersión de 0,0; con 10 mezclas gaussianas.

CONCLUSIONES

En el presente proyecto de investigación, se presenta una metodología para el reconocimiento de comandos de voz para el control de una silla de ruedas, basado en los HMM, que ha sido caracterizado mediante los coeficientes cepstrales en la escala de Mel; que permiten la caracterización de las señales dinámicas.
Basados en esto, se hallaron los siguientes resultados: la palabra con mayor cantidad de aciertos es derecha con un porcentaje de 100%, y dispersión de 0,0, con 2 mezclas gaussianas. La palabra que presenta mayor dificultad es adelante, pues logra obtener el 100% de aciertos y una dispersión de 0,0; empleando un número mayor de mezclas gaussianas (alrededor de 10).

Es posible emplear un algoritmo de extracción de características, basado en los coeficientes cepstrales de Mel con un clasificador HMM, que permite el reconocimiento de comandos de voz para el control de dispositivos de ayuda a personas en estado de discapacidad. Esta metodología permitió establecer que los HMM presentan una muy buena respuesta comparada con la respuesta de las GMM, aunque persiste el problema de un mayor costo computacional.

AGRADECIMIENTOS

Este trabajo se enmarca en el proyecto de investigación: "Metodología para el reconocimiento de voz basado en dinámicas estocásticas orientado al control de una silla de ruedas" código PM11129, del Instituto Tecnológico Metropolitano y el programa de Maestría en Control y automatización industrial de la misma institución.

BIBLIOGRAFÍA

1. Grimm M, Kroschel K, editors. Robust Speech Recognition and Understanding. Vienna: I -Tech Education and Publishing 2007; 3, 8.
2. Homayounfar K. Rate adaptive speech coding for universal multimedia access. Sig Proc Mag, 2003; 3, p. 30–39.
3. Shaughnessy DO. Interacting with computers by voice: Automatic speech recognition and synthesis. LNCS 4233, 2006; 3, pp. 489–498.
4. Mellorf B, Baber C, Tunley C. Evaluating automatic speech recognition as a component of a multi-input device human-computer interface. Proceedings of the International Conference on Spoken Language Processing, 1996; 3, pp. 1668–1671.
5. Kubik T, Sugisaka M. Use of a cellular phone in mobile robot voice control. Proceedings of the 40th SICE Annual Conference International Session Papers SICE 2001, 2001; 3, pp. 106–111.
6. Shamma S. Relevance of auditory cortical representations to speech processing and recognition. Automatic Speech Recognition and Understanding p. 5, 2005.
7. Juang B, Chen T. The past, present, and future of speech processing. Sig Proc Mag, 1998; 15 (3), pp. 24–48.
8. Goecke R. Current trends in joint audio-video signal processing: A review. IEEE Eight International Symposium on Signal Processing and Its Applications ISSPA2005, Sydney, Australia, 2005; p. 70–73.
9. Campbell R. Audio-visual speech processing. Elsevier, 2006; pp. 562–569.
10. Campbell R. The processing of audio-visual speech: empirical and neural bases," Philosophical Transactions of The Royal Society B, 2008; 363, p. 1001–1010.
11. Hurtado J, Castellanos G, Suarez J. Effective extraction of acoustic features after noise reduction for speech classification. In International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science, 2002; 3, pp. 245–248.
12. Schroeter J, Larar J, Sondhi M. Speech parameter estimation using a vocal tract/cord model. In IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, 1987; 12 (3), pp. 308–311.
13. Cheng N, Mabiner L, Rosenberg A, Mooonegal C. Some comparisons among several pitch detection algorithms. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '76, IEEE, 1976; 3(1), pp. 332– 335.
14. Hoory S, Sagi A, Shechtman A, Sorin A, Shuang A, Bakis R. High quality sinusoidal modeling of wideband speech for the purposes of speech synthesis and modification. In IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP'06), IEEE, 2006; 1(3), pp. I877–I880.
15. Termens RG, Lafont JO, Portabella FG, Roca JM. Síntesis de voz utilizando difonemas: uniones entre vocales. Procesamiento del lenguaje natural, 1997; 21, pp. 69–74.
16. Rothweiler J. Noise-robust 1200-bps voice coding. Proceedings of the Tactical Communications Conference: Technology in Transition, 1992; 1(7), pp. 65–69.
17. Macres J. Real-time implementations and applications of the US federal standard CELP voice coding algorithm. Proceedings of the Tactical Communications Conference: Technology in Transition, 1992; 1, pp. 41–45.
18. Rabiner LR. A tutorial on hidden markov models and selected application in speech recognition. Proceedings of the IEEE, 1989; 77(2) , pp. 257–286.
19. Anderson S, Kewley-Port D, "Evaluation of speech recognizers for speech training applications. IEEE Transactions on Speech and Audio Processing, 1995; 3(4), pp. 229–241.
20. Rabiner L, Juang B, editors. Fundamentals of Speech Recognition. Prentice-Hall, 1993.