Diseño de un sistema de comunicación entre personas sordas y personas oyentes

Design of a communication system between deaf people and hearing people

José Miguel Martínez-Aponte
Ingeniero en Telecomunicaciones
Universidad de Pamplona
Pamplona, Colombia
jmiguel_29ma@hotmail.com

Sergio Stivenson-Pinto
Ingeniero en Telecomunicaciones
Universidad de Pamplona
Pamplona, Colombia
ing.sergio.elk@gmail.com

Resumen– El siguiente artículo presenta el diseño de un *software* prototipo capaz de interpretar dos tipos de lenguaje, el lenguaje articulado y el lenguaje de señas. Se pretende mediante técnicas de procesamiento digital de señales y establecimiento de patrones encontrar la manera de decodificar los lenguajes, de tal forma que una persona sorda pueda entender lo que le comunica una oyente y viceversa.

Palabras clave– Reconocimiento de señas, cuantificación, caracterización y modelos ocultos de Markov.

Abstract– This article presents the design of a software prototype, able to allow the communication between two languages, the voice and the signs language. It is pretended to decode both languages by using digital signal processing techniques as well as the establishment of common patterns. The goal is that a deaf person be able to understand what a hearing person says and vice versa.

Keywords– Signs recognizing, quantification, characterization, Hidden Markov Models.

1. INTRODUCCIÓN

El propósito de este proyecto es diseñar un sistema decodificador de lenguaje que logre el reconocimiento de voz y el de señas por medio de una computadora, usando como herramienta de desarrollo el *software* MatLab. Este sistema es un prototipo, el objetivo principal es el de lograr el reconocimiento de un número discreto de palabras y frases, pretendiendo la mayor robustez posible. Se utiliza el análisis cepstral en la escala mel y los modelos ocultos de Márkov para el reconocimiento de voz. En cuanto al reconocimiento de señas, la extracción de parámetros característicos se lleva a cabo gracias a la detección de color y a patrones de movimiento.

2. DESARROLLO DEL ARTÍCULO

2.1 Acerca de los lenguajes

Con el fin de ampliar la perspectiva que se tiene acerca de ambos lenguajes, se indagó sobre las características generales de cada uno de ellos.

2.1.1 Lenguaje articulado (voz)

La voz es el conjunto de sonidos que emplea el ser humano para comunicarse por medio de ondas mecánicas. Es producida por el aparato fonador, compuesto por las cuerdas vocales y el tracto vocal. Las cuerdas vocales son las encargadas de generar la frecuencia fundamental y sus armónicos, mientras el tracto vocal es el conjunto de cavidades resonantes, las cuales se acomodan de tal manera que amplifiquen o atenúen dichos armónicos, y puedan dar forma a los distintos sonidos de la voz [1].

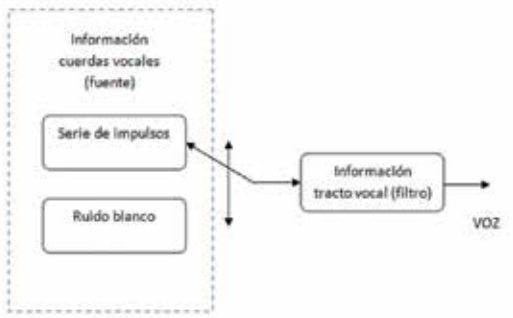
2.1.2 Modelo source-filter de la voz

Se puede considerar la voz como una combinación aleatoria entre una señal periódica generada cuando las cuerdas vocales vibran y un ruido blanco (característico de los sonidos sordos como el de la 'k') cuando hay ausencia de esta vibración. Esta combinación es denominada fuente o source, es llamada así porque esta contiene la energía e información cruda de la voz y es la entrada a un filtro de características variables cuya función es terminar de darle forma a la señal de voz.

Recibido: 28/10/2014 Aceptado: 15/12/2014 ITECKNE Vol. 12 Número 2 • ISSN 1692 - 1798 • ISSN Digital 2339-3483 • Diciembre 2015 • 138 - 145

La Fig. 1 muestra un esquema del modelo source-filter, un modelo representativo del origen de la voz. De esta manera se puede incorporar al modelo la información de las cuerdas vocales más la información del tracto vocal, cuyas piezas móviles están consideradas como las características variables del filtro. La información del tracto vocal está contenida en la envolvente del espectro resultante [3].

Fig. 1. MODELO SOURCE-FILTER



Fuente: J.M. Martínez [2].

2.1.3 *Lenguaje de señas*

Es un lenguaje basado en el movimiento corporal y gestual y se percibe de forma visual. El movimiento de las manos es el componente más influyente en las señas y es la base del establecimiento de patrones por reconocer.

2.1.4 *Modelo de reconocimiento*

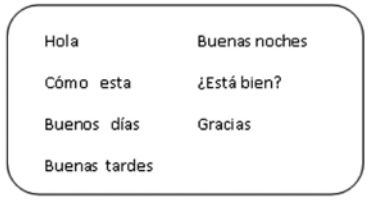
En el reconocimiento automático de ambos lenguajes se establecen tres pasos importantes, la caracterización del lenguaje, el entrenamiento o aprendizaje del sistema y el proceso de reconocimiento.

2.2 *Reconocimiento de voz*

El reconocimiento de voz en este proyecto es aplicado a palabras aisladas, el discurso por parte del usuario debe ser pausado entre palabra y palabra. Se reconocen diez palabras, suficientes para establecer las frases de saludo mostradas en la Fig. 2.

La señal de voz es caracterizada matemáticamente por medio de los coeficientes Mel-cepstrum (MFCC por sus siglas en inglés). Estos coeficientes logran captar la información del tracto vocal, información que corresponde a las bajas frecuencias de la señal de voz en el dominio cepstral [4].

Fig. 2. MUESTRA LAS PALABRAS QUE EL SISTEMA RECONOCE

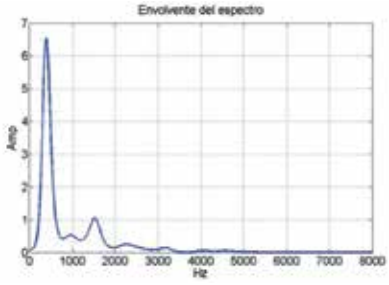


Fuente: [2].

2.2.1 *Análisis cepstral en la escala Mel*

Considerando el modelo source-filter, enunciado anteriormente, la información del tracto vocal se encuentra en la envolvente del espectro en frecuencia de la señal de voz. En la Fig. 3 se puede ver la envolvente de una señal de voz en el dominio de la frecuencia.

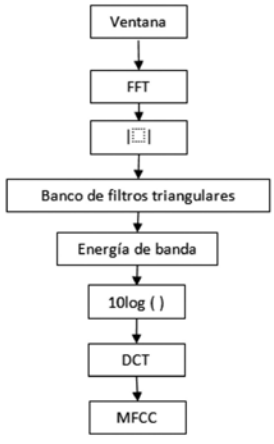
Fig. 3. ENVOLVENTE DEL ESPECTRO EN FRECUENCIA DE UNA SEÑAL DE VOZ



Fuente: [2].

Para poder captar las diferencias cepstrales entre un fonema y otro a medida que evoluciona una palabra en el tiempo, se analiza la señal de voz por ventanas de corta duración, estas ventanas están solapadas entre sí en un 50%. A cada una de estas ventanas se les calcula los MFCC como se muestra en la Fig. 4 [5], [6].

Fig. 4. PROCESO DE OBTENCIÓN DE LOS MFCC



Fuente: [2].

El espectro de amplitud de cada ventana, obtenido con la transformada discreta de Fourier, es multiplicado por un banco de *N* filtros triangulares equi-espaciados en la escala Mel. Luego a la salida de cada filtro se le calcula el logaritmo de la energía y, por último, la transformada discreta del coseno. En el proyecto se usó un banco de 20 filtros y los primeros 17 MFCCs para crear el vector característico. Con el fin de hacer más precisa la caracterización del vector, se le añaden también algunos coeficientes dinámicos, obtenidos de la primera y segunda derivada de los MFCCs (1) y (2) respectivamente [7].

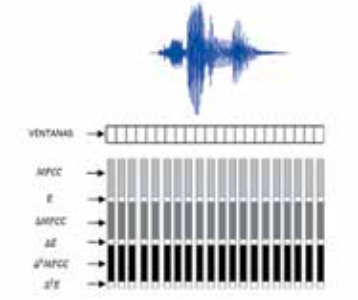
$$\delta_m(t) = \frac{\Sigma_{k=1}^{D} k\left(c_m(t+k) - c_m(t-k)\right)}{2\Sigma_{k=1}^{D} k^2} \quad (1)$$

$$\delta_m^2(t) = \frac{\Sigma_{k=1}^{D} k\left(\delta_m(t+k) - \delta_m(t-k)\right)}{2\Sigma_{k=1}^{D} k^2} \quad (2)$$

Con $1 \leq m \leq 17$ y D=2

Donde c_m es cada uno de los 17 MFCC, δ_m los coeficientes de la derivada y δ^2_m los de la aceleración (o segunda derivada), t es el parametro tiempo y D es la cantidad de vectores adyacentes sobre los cuales se calculan los coeficientes dinámicos. La Fig. 5 muestra la señal de voz dividida en ventanas, cada una con su vector característico, cada vector conformado por los MFCC, la derivada y la aceleración de los MFCC y tres valores de energía, uno por cada grupo de coeficientes.

Fig. 5. COMPOSICIÓN DE LOS VECTORES CARACTERÍSTICOS



Fuente: [2].

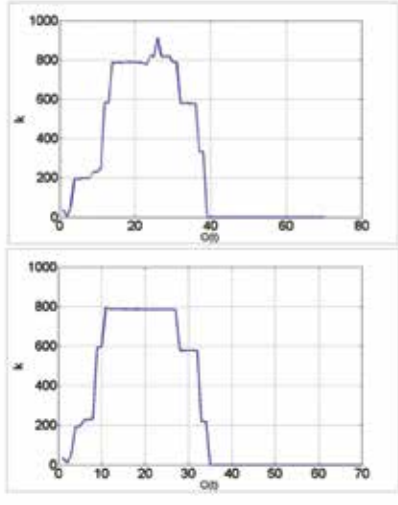
2.2.2 Proceso de cuantificación

Con este proceso se pretendió representar los vectores característicos con un índice, de tal forma que aquellos vectores cercanos entre sí pudieran ser cuantificados con un mismo valor y reducir la cantidad de información para utilizar y por ende el costo computacional. Para este propósito se tomaron suficientes muestras de voz y con los vectores característicos se empleó el algoritmo LGB para obtener un codebook, o libro de regiones representativas, con 1024 índices o codewords [3].

El proceso de cuantificación consiste entonces en asignar por medio de la distancia euclidiana, uno de los 1024 índices a cada uno de los vectores característicos de la señal de voz capturada, como resultado queda un vector de observación $O=\{O_1,O_2,O_3 \ldots O_T\}$. Donde O_n es la observación de la ventana . De esta manera datos por analizar se reducen de una matriz de MxT a un vector de observación de 1xT, siendo M la longitud del vector característico y T la cantidad de ventanas temporales de la señal de voz capturada.

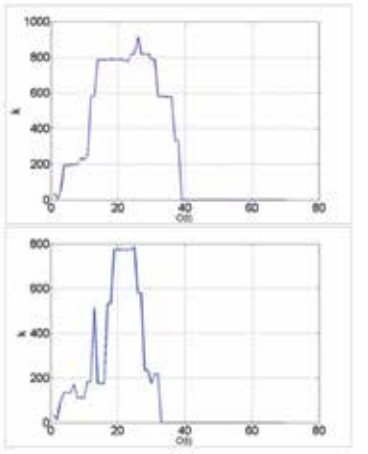
Como puede verse en las figs. 6 y 7, la cuantificación permite notar similitud y diferencias entre las distintas muestras de voz. El eje de abscisas corresponde al instante t, y las ordenadas al índice k o codeword correspondiente al vector en ese instante de tiempo. Una vez cuantificada la señal de voz se usan los modelos ocultos de Markov para determinar a qué palabra corresponde la secuencia de observación. Para esto se hizo necesario entrenar un modelo por cada palabra, de tal forma que se hiciera máxima la probabilidad condicional, $P(O|\lambda)$ que es la probabilidad de que dado el modelo se dé la observación O.

Fig. 6. VECTORES DE OBSERVACIÓN DE UNA MISMA PALABRA (HOLA)



Fuente: [2].

Fig. 7. VECTORES DE OBSERVACIÓN DE LAS PALABRAS *HOLA* Y *DÍAS*



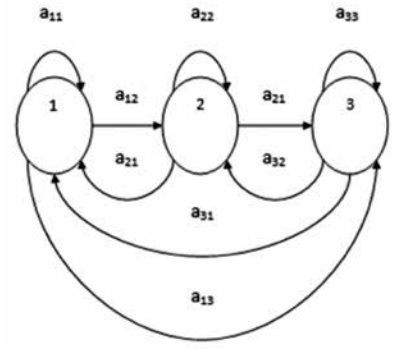
Fuente: [2].

2.2.3 Modelos ocultos de Márkov (HMM) t3

Son modelos probabilísticos que permiten representar un proceso doblemente estocástico, donde la probabilidad de estado asociado no es observable, de ahí la palabra oculto. Los HMM son útiles en el reconocimiento de patrones si estos cumplen con las características de un proceso aleatorio [8], [9].

La Fig. 8 muestra un ejemplo de un modelo de tres estados, representados gráficamente por medio de círculos, las líneas que unen estos círculos corresponden a la probabilidad de transición de un estado a otro. Estas probabilidades de transición son acomodadas en la matriz de transición *A* de NxN posiciones, donde N es la cantidad de estados del modelo y junto con el vector de probabilidad inicial y la matriz de densidad de emisión *B* son los tres componentes de un HMM. La cantidad de estados de un modelo determina la robustez y complejidad al momento del entrenamiento, entre más estados tenga más complejo es [10].

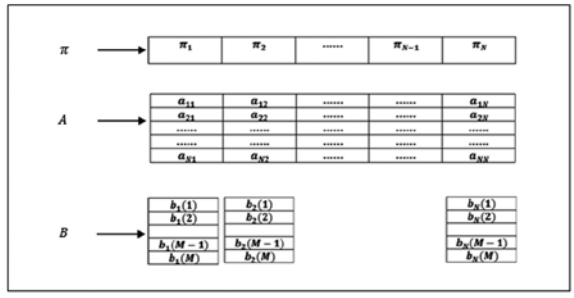
Fig. 8. REPRESENTACIÓN GRÁFICA DE UN PROCESO MÁRKOV



Fuente: [2].

La Fig. 9 muestra los elementos de un modelo oculto de Márkov de N estados y son definidos por (3), (4) y (5). Donde q_t es el estado en el que el modelo se encuentra en el instante t.

Fig. 9. ELEMENTOS DE UN HMM



Fuente: [2].

$$\pi_i = P(q_1 = i) \;\; con \;\; \sum_{i=1}^{N} \pi_i = \quad (3)$$

$$a_{ij} = P(q_{t+1} = j \mid q_t = i) \;\; con \;\; \sum_{j=1}^{N} a_{ij} = 1 \,\forall i \quad (4)$$

$$b_j(k) = P(O_t = k \mid q_t = i) \;\; con \;\; \sum_{k=1}^{M} b_j(k) = 1 \,\forall j \quad (5)$$

Lo que se hizo fue, a partir de cada observación de una palabra determinada, entrenar un HMM inicial de componentes λ=(A,B,π) con valores iniciales aleatorios hasta obtener un modelo más especifico para esa observación. A través de distintos algoritmos, que serán nombrados en la siguiente sección, se optimizó un modelo por cada una de las palabras por reconocer, teniendo así un diccionario de 10 HMMs.

2.2.4 Los tres problemas de un HMM y su solución

Existen tres problemas en un HMM [6]. El primero de ellos es hallar P(O/λ), la solución está en el algoritmo Forward o Backward:

- *Algoritmo Forward*

  Inicio:

$$\alpha_1(i) = \pi_i b_i(O_1) \;\; 1 \le i \le N \quad (6)$$

  Iteracion:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1}) \quad (7)$$

  Fin:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i) \quad (8)$$

- *Algoritmo Backward*

Inicio:

$$\beta_T(i) = 1 \quad 1 \le i \le N \quad (9)$$

Iteración:

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(O_{t+1}) \quad (10)$$

con 1 ≤ i ≤ N y t=T-1, T-2,..1

Fin:

$$P(O|\lambda) = \sum_{i=1}^{N} \beta_1(i) \pi_i b_i(O_1) \quad (11)$$

El segundo problema es encontrar la secuencia de estados óptima q_t, para determinar los estados por los cuales transita un modelo al ser evaluado y en qué momento lo hacen. La secuencia de estados y la probabilidad P(O/λ) son los dos criterios usados para el correcto reconocimiento de palabras y para el entrenamiento. Con el algoritmo Viterbi se obtiene esa secuencia:

- *Algoritmo Viterbi*

Inicio:

$$\delta_1(i) = \pi_i b_1(O_1) \quad 1 \le i \le N \quad (12)$$

Iteración:

$$\delta_{t+1}(j) = \left[\max_{1 \le i \le N} \delta_{t+1}(i) a_{ij} \right] b_j(O_{t+1}) \quad (13)$$

con 1 ≤ j ≤ N y 1 ≤ t ≤ T-1

$$\varphi_{t+1}(j) = \arg \left[\max_{1 \le i \le N} \delta_t(i) a_{ij} \right] \quad (14)$$

con 1 ≤ j ≤ N y 1 ≤ t ≤ T-1

Fin:

$$q_T^* = \arg \left[\max_{1 \le i \le N} \delta_T(i) \right] \quad (15)$$

Secuencia de estados óptima:

$$q_t^* = \varphi_{t+1}(q_{t+1}^*) \quad (16)$$

con t=T-1, T-2,... 1

El tercer problema es el de entrenar un modelo a partir de una secuencia de observación dada. Para este propósito se utiliza el algoritmo Baum-welch y un algoritmo de re-estimación [11], [12].

La tabla I muestra la cantidad de estados de cada uno de los modelos del sistema, entre más estados tenga un modelo mayor será el tiempo y la dificultad de su entrenamiento.

TABLA I

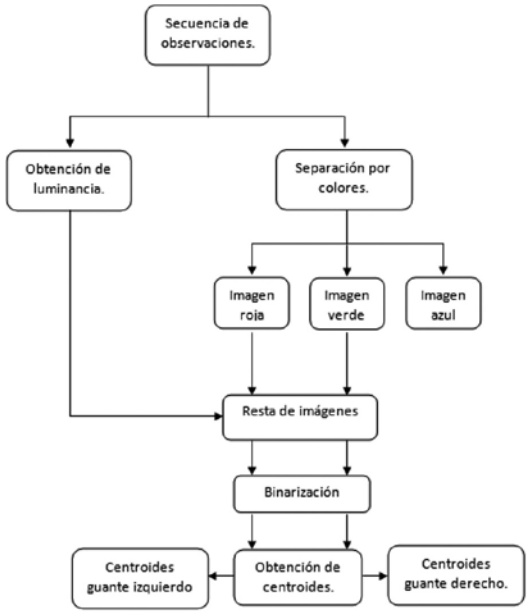
NÚMERO DE ESTADOS DEL HMM DE CADA PALABRA POR RECONOCER

Modelo	No de estados
Hola	4
Cómo	4
Está	4
Buenos	5
Buenas	5
Días	4
Tardes	5
Noches	5
Bien	4
Gracias	5

2.3 Reconocimiento de señas

Se tomó como unidad de caracterización el centroide de las manos. Para ayudar a la ubicación de las manos en la imagen se usaron dos guantes de distinto color, uno para cada mano, y por medio de las técnicas mostradas en la Fig. 10 se llevó a cabo esta tarea [13], [14].

Fig. 10. PROCESO DE EXTRACCIÓN DE LOS CENTROIDES



Fuente: [2].

El sistema reconoce cuatro señas y al igual que con la voz se capturan de forma aislada. Las se-

ñas son los saludos: hola, buenos días, buenas tardes y buenas noches. Algunas señas varían dependiendo del país por lo que se establecen para este artículo las correspondientes al lenguaje de señas de Colombia [15].

2.3.1Extracción de patrones representativos

Para la extracción de patrones, por cada muestra de una seña se toman 40 fotogramas en 3 segundos, velocidad suficiente para captar el movimiento de las manos y establecer el patrón de movimiento de los centroides de cada guante.

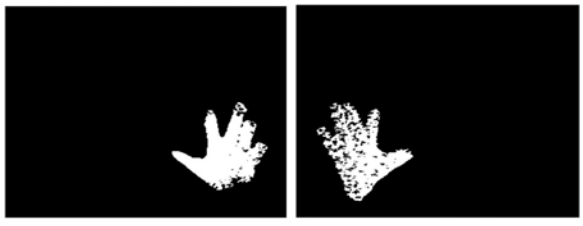
La Fig. 11 es una secuencia de imágenes RGB de resolución 640x480, las cuales contienen los movimientos correspondientes a la seña.

Fig. 11. SECUENCIA DE IMÁGENES DE UNA SEÑA

Fuente: [2].

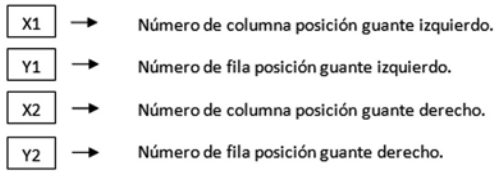
A cada una de estas 40 imágenes se les hace el filtro por color para detectar los guantes por separado y luego son binarizadas, se puede apreciar un ejemplo de esta binarización en la Fig. 12, a cada imagen binaria se le encuentran los centroides de cada guante y se almacenan en el vector de características de la Fig. 13. En la Fig. 14 se observa una matriz que contiene las 40 observaciones representadas por los vectores característicos.

Fig. 12. IMAGEN BINARIA DE LOS GUANTES



Fuente: [2].

Fig. 13. COMPOSICIÓN DEL VECTOR CARACTERÍSTICO



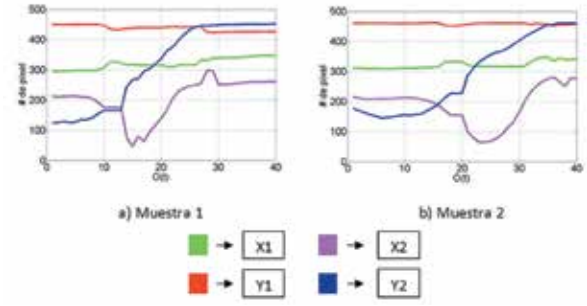
Fuente: [2].

Fig. 14. PATRONES DE MOVIMIENTO

Fuente: [2].

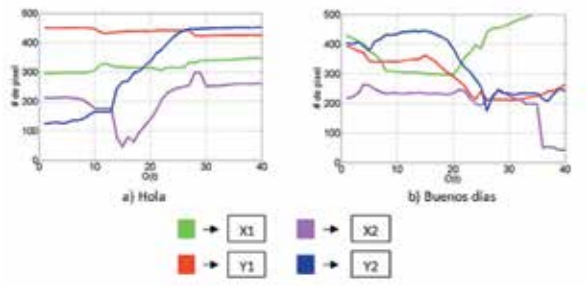
En las figs. 15 y 16 se puede ver el movimiento de los centroides en señas iguales y en señas distintas, respectivamente.

Fig. 15. PATRONES DE DOS MUESTRAS DE LA MISMA SEÑA (HOLA)



Fuente: [2]

Fig. 16. PATRONES DE DOS SEÑAS DISTINTAS



Fuente: [2].

2.3.2 Reconocimiento por medio de la correlación

Se empleó la correlación, ya que es un método muy útil para medir la semejanza entre vectores de una misma longitud. La correlación mide la relación de cambio entre dos variables aleatorias y da como resultado un valor entre -1 y 1 siendo 1 el valor cuando las dos variables aleatorias varían en igual proporción. Se utilizó un umbral de acierto de 0.9 al momento de comparar y reconocer las señas.

2.3.3 Diccionario de señas

Al igual que con la voz es necesario que el sistema aprenda a reconocer señas o tenga una base de datos con la cual comparar y determinar la seña más parecida a la obtenida en una observación. Para esto se creó un diccionario con 10 muestras por cada una de las cuatro señas del sistema, así cuando se capture una nueva seña y se compare, se tendrá mayor probabilidad de acierto si se pregunta cuál fue la seña que más se repitió entre las del diccionario que dieron buena correlación.

3. RESULTADOS

El reconocimiento de voz fue realizado sobre una sola voz, la del autor, ya que es un prototipo el objetivo principal el de indagar y probar las técnicas mencionadas.

En las tablas II y III se ven los resultados luego de realizar las pruebas al sistema en un ambiente silencioso; se llevaron a cabo 100 pruebas por cada palabra y 50 por cada frase.

Para el reconocimiento de señas se hicieron 50 pruebas por cada seña en un ambiente de luminosidad normal, y los resultados en cuanto al acierto se pueden ver en la tabla IV.

Los tiempos de respuesta en el reconocimiento de voz fueron bastante buenos, al haber pocos modelos en el diccionario el reconocimiento fue casi inmediato (tiempo inferior al segundo). En cuanto al reconocimiento de imagen el procesamiento fue un poco más pesado, en consecuencia tuvo tiempos de respuesta más lentos, de 2 a 4 segundos aproximadamente en cada reconocimiento realizado.

TABLA II
RESULTADOS DEL RECONOCIMIENTO DE VOZ EN PALABRAS

Palabra	% de reconocimiento
Hola	72
Como	88
Esta	98
Bien	90
Gracias	99
Buenos	80
Buenas	84
Días	90
Tardes	99
Noches	96

TABLA III
RESULTADOS DEL RECONOCIMIENTO DE VOZ EN FRASES

Frase	% de reconocimiento
“¿hola cómo está?”	80
“buenos días”	84
“buenas tardes”	96
“buenas noches”	86
“¿está bien?”	90
“bien gracias”	92

TABLA IV
RESULTADOS DEL RECONOCIMIENTO DE SEÑAS

Seña	% de reconocimiento
Hola	96
Buenos días	96
Buenas tardes	94
Buenas noches	92

4. CONCLUSIONES

Trabajar con modelos probabilísticos en el área de reconocimiento de voz permite al sistema una mayor flexibilidad, ante las posibles diferencias o errores existentes entre secuencias de observación de una misma palabra.

El uso de centroides como elementos característicos de las señas, y el establecimiento de patrones basados en la secuencia de movimiento dio muy buenos resultados en el reconocimiento de señas.

Las pruebas realizadas mostraron que los métodos utilizados para el reconocimiento de ambos lenguajes dieron un porcentaje de acierto alto y un tiempo de respuesta relativamente bueno.

AGRADECIMIENTOS

Dedicado a mis padres José A. Martinez y María R. Aponte, quienes han sido los pilares de mi formación y mi vida, a mis hermanos, a todos ellos les debo su apoyo, también agradecer a la Universidad de Pamplona por la formación y al director de mi proyecto de grado sobre el cual fue basado este artículo, ingeniero Sergio S. Pinto.

REFERENCIAS

[1] American Academy of Otolaryngology-head and neck surgery, "How the Voice Works", [online]. Disponible en http://www.entnet.org/content/how-voice-works.

[2] J. M. Martinez, "Diseño de un sistema de comunicación entre personas sordas y personas oyentes", tesis pregrado, Depto. de Ingenierías y Arquitectura, Universidad de Pamplona, Villa del Rosario, 2013.

[3] F. Martinez, G. Portale, H. Klein y O. Olmos, "Reconocimiento de voz, apuntes de cátedra para Introducción a la Inteligencia Artificial," [online] tutorial, Universidad Tecnológica Nacional, Argentina. Disponible: http://www.secyt.frba.utn.edu.ar/gia/IA1_IntroReconocimientoVoz.pdf

[4] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," [online] artículo, Cambridge research laboratory and Compaq Computer Corporation, Disponible en: ftp://gatekeeper.dec.com/pub/Compaq/CRL/publications/logan/musicir_paper.pdf.

[5] K. Prahallad, "Speech Technology: A Practical Introduction," [online] laboratorio, Carnegie Mellon University & International Institute of Information Technology Hyderabad, Disponible en: http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

[6] "Mel Frequency Cepstral Coefficient (MFCC)," [online] tutorial en internet, Disponible en: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[7] L. Weisi, D. Tao, J. Kacprzyk, Z. Li, E Izquierdo and H. Wang, "A comprehensive study of sports video analysis," [online] in Multimedia analysis processing and communications, Springer 2011 edition, May 12, 2011. Disponible en https://books.google.es/

[8] I. Villamil, "Aplicaciones en reconocimiento de voz utilizando HTK," [online] tesis de grado, Facultad de Ingeniería, Pontificia Universidad Javeriana, 2005, Disponible en: http://repository.javeriana.edu.co/bitstream/10554/7588/1/tesis95.pdf

[9] P. Blunsom, "Hidden Markov Models," [online] tutorial, Utah State University, august 19, 2004. Disponible en http://digital.cs.usu.edu/~cyan/CS7960/hmm-tutorial.pdf.

[10] F. Salcedo C. "Sistemas de reconocimiento basados en modelos ocultos de Márkov," [online] en Modelos ocultos de Márkov: del reconocimiento de voz a la música, Ed. Lulu Press, (2009). Disponible en https://books.google.es/

[11] L. Bergasa, "Introducción a los modelos ocultos de Márkov", [online] tutorial, Depto. de Electrónica, Universidad de Alcalá, Argentina. Disponible en: http://www.bioingenieria.edu.ar/academica/catedras/metestad/Introduccion%20al%20Modelo%20oculto%20de%20Markov.pdf.

[12] L. Bahi, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Acoustics, Speech, and Signal Processing*, IEEE International Conference on ICASSP 86, vol. 11.

[13] R. Buksh, S. Routh, P. Mitra, S. Banik, A. Mallik and S. Gupta, "MATLAB based image editing and color detection," *International Journal of Scient ific and Research Publications*, vol. 4, Issue 1, January 2014 ISSN 2250.

[14] OTSU, Nobuyuki. "A Threshold selection method from Gray-level histograms," *IEEE Transactions on systems, man, and cybernetics* 9, nº 1 (1979), 62-66.

[15] Colombia Aprende, "Vocabulario por categorías," [online]. Disponible en:http://mail.colombiaaprende.edu.co:8080/recursos/lengua_senas/