
XIII Congreso Argentino de Acústica
VII Jornadas de Acústica, Electroacústica y Áreas Vinculadas

Buenos Aires, 29 y 30 de octubre de 2015

AdAA2015-A008

**Sistema automático de reconocimiento de voz
para control de acceso**

Maximiliano Manuel Yommi^(a),

(a) Ingeniería de Sonido, Universidad Nacional de Tres de Febrero. Valentin Gomez 4828, Caseros, Buenos Aires, Argentina. E-mail: maximilianoyommi@gmail.com

Abstract

In this research, the study of the techniques of signal processing of voice were discussed, particularly feature extraction of speaker identification to discriminate between other speakers. The main objective is the control access, through of one human machine interface. The method proposed in this research has a physical layer (hardware) in Arduino and an application layer (software) developed in Matlab 7.10. The Mel Frequency Cepstrum Coefficients, are the best available approximation of speakers features. Backpropagation neural network, with a multilayer perceptron topology, are used to design the classification system. The training and testing of the system was carried out with a total base of 80 different voice samples. Finally, it was concluded that the system has an average success rate of 64.58%.

Resumen

En esta investigación, se aborda el estudio de las técnicas de procesamiento de señales de voz que permiten extraer las características intrínsecas de la identificación del locutor. El objetivo es delimitar el acceso a solo aquellos locutores identificados y habilitados. Dicho procedimiento se realiza por medio de una interfaz máquina-usuario. El método propuesto en esta investigación posee una capa física (hardware) implementada en torno a Arduino Uno y una capa de aplicación (software) desarrollada sobre la plataforma de MatLab 7.10. Para evaluar las características de los locutores se utilizaron los coeficientes cepstrales de frecuencia Mel. Para la clasificación, una red neuronal artificial con topología perceptrón multicapa y un aprendizaje supervisado del tipo Backpropagation. El entrenamiento y prueba del sistema se llevó a cabo con una base total de 80 muestras diferentes de voz. Finalmente el rendimiento arrojado refleja que el sistema posee un porcentaje de acierto promedio del 64,58 %.

1 Introducción

Hoy en día es imposible concebir la idea de proteger objetos, información clasificada o simplemente controlar el registro de acceso de personas, sin un sistema digital de soporte. Los sistemas de seguridad basados en características biométricas son, en la actualidad, lo más confiables, dúctiles y aceptados por la mayoría de sus usuarios. Estos sistemas se fundamentan en características fisiológicas o de comportamiento de personas. Las primeras investigaciones sobre esta disciplina se realizaron en el ámbito forense, en el año 1941, un excelente resumen cronológico realizado por Campell sobre el desarrollo de sistemas SR (Speaker Recognition), se observa en (Campbell Jr, J., P., 1997). Esta investigación tiene como línea de trabajo principal el reconocimiento de locutores o hablantes por voz, para implementarlo como sistema de control de acceso. El campo de aplicación de la investigación se concentra en el procesamiento digital de señales (DSP, Digital Signal Processing) en principio a nivel de software, dado que el hardware que se utiliza solo cumple la función de interfaz de comunicación con el algoritmo, no participa en el procesamiento de la señal de interés. La hipótesis principal de esta investigación es determinar si es posible desarrollar un sistema automático de reconocimiento de voz para control de acceso, basado en redes neuronales. El objetivo principal es diseñar un sistema automático de reconocimiento de voz para control de acceso, compuesto de software y hardware propio, que permita identificar al menos 4 locutores de ambos sexos.

2 Conceptos básicos

2.1 Aspectos fisiológicos

Definimos a los fonemas como la menor unidad del habla, la ciencia que los estudia es la Fonología (Miyara, F., 1999). El idioma español posee alrededor de 24 fonemas diferentes que se definen a partir de los sonidos que es capaz de producir el aparato fonador (Basso, G., 2006). Otro aspecto importante para destacar de los fonemas, es que su duración temporal no es uniforme, por ejemplo en el caso de las vocales su duración promedio ronda los 0,100 s, mientras que las consonantes son mucho más rápidas en el orden de los 0,020 s (Rufiner, H. L., 2009), dato sumamente importante a la hora de considerar el tamaño de la ventana temporal de análisis de la señal de voz, con el fin de no perder ningún fonema (Basso, G., 2006; Rufiner, H. L., 2009; Miyara, F., 1999).

Por otra parte otro aspecto a tener en cuenta es la variabilidad de la voz: una persona presenta muchos cambios en la voz no solo el proceso natural de envejecer, sino además, producto de estados anímicos extremos, enfermedades, entre otros. La capacidad de afrontar este hecho, así como otros factores, ruido de fondo y/o ruido eléctrico, se le denomina robustez (Mammone, R. J., Zhang, X., & Ramachandran, R. P., 1996) (Reynolds, D., A., 1996) (Moreno, P. J.; Stern, R. M., 1994).

2.2 Aspectos técnicos de un sistema de SR

En vista de la gran aleatoriedad que presentan las señales de voz, se usan a lo largo de toda esta investigación modelos estocásticos que permiten estimar propiedades estadísticas de la señal bajo estudio (Marquina, A. Á., 2001). En la figura 1 se identifican las principales faces que intervienen en un sistema SR.

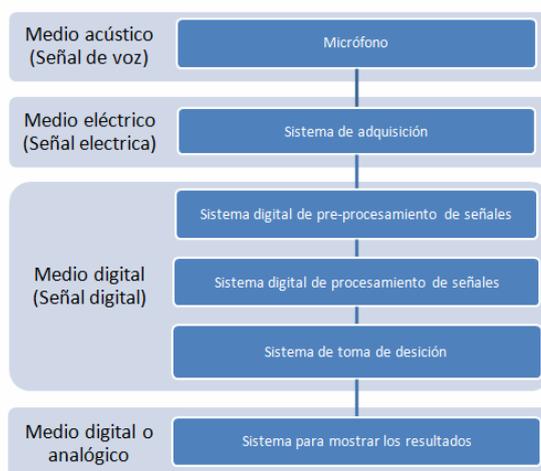


Figura 1. Ramificación del procesamiento del habla.

2.2.1 Sistema de adquisición

2.2.1.1 Muestreo

La señal que recibe el sistema de adquisición por parte de un transductor del tipo Electret (en la mayoría de los casos), es continua en tiempo y amplitud, por este motivo para poder manipular dicha señal es necesario tomar muestras a intervalos fijos y exactos de tiempo. El rango mínimo del habla humana es generalmente inferior a 5000 Hz, por lo que se requiere una tasa de muestreo superior a 10000 Hz, para respetar el teorema de muestreo de Nyquist-Shannon (Xiaoguo, X., 2013).

2.2.2 Sistema digital de pre-procesamiento de señal

Este sistema se compone de un pre-procesamiento que acondiciona la señal para la extracción de sus características intrínsecas donde reside la autenticación del locutor.

La figura 2 describe los pasos que intervienen en esta tarea.



Figura 2. Pre-procesamiento de la señal de voz digitalizada.

2.2.3 Filtrado de ruido

El primer proceso en esta cadena es el filtrado de la señal, con el objetivo de eliminar la mayor cantidad de ruido posible, producto de la adquisición de la señal en el medio analógico. Son varias las técnicas que pueden ser aplicadas para cumplir este objetivo, filtros FIR, resta espectral (Borrás, O, 2006), filtrado Wavelet (Alcon Choque, E. F., 2010; Sanchez Marin, G., D., 2004; Kouro, S., & Musalem, R., 2002), entre otros. En particular en esta investigación se trabaja con filtrado Wavelet, esta técnica se la define como “Denoising” (Kouro, S., & Musalem, R., 2002).

2.2.4 Normalización

La finalidad tras esta etapa es lograr que la media de los valores sea cero y normalizar la amplitud a un rango entre $[-1, 1]$; de esta manera se logra que en extracción de características y la clasificación, se pierda la sensibilidad a los valores máximos de la amplitud (Varela Rincón, J., & Loaiza Pulgarín, J. E., 2008), además la normalización permite ajustar todos los parámetros a una sola escala al momento de ser utilizados por la Red Neuronal Artificial (ANN), evitando de esta manera problema de estabilidad (Cruz Beltrán, L.; Acevedo Mosqueda, M., 2008). Para esto se utiliza la ecuación

$$x_{\text{norm}}[i] = \frac{x[i]}{\max(\text{abs}[x[i]])} \quad (1)$$

donde $x[i]$ hace referencia a las muestras obtenidas del proceso de adquisición, en particular en el entorno de MatLab este cociente se realiza punto a punto. En teoría debido a que se aplica la misma cantidad de ganancia a todo el rango, la relación señal-ruido generalmente no cambia, pero se observa en el análisis tiempo-frecuencia, como este proceso genera aditamentos espectrales, que no existía antes.

2.2.5 Acotamiento

Luego de la normalización se aplica la siguiente ecuación:

$$E_{\text{Prom}} = \frac{1}{N} \sum_{k=1}^N (x[k])^2 \quad (2)$$

La ecuación (2) hace referencia al valor de energía promedio. El proceso consta de dividir la señal en ventanas de un número determinado de muestras, calcular la energía de ese trozo de señal y por último, determinar por medio de un umbral de decisión, si la ventana es eliminada o no, dependiendo si se trata de una ventana que contiene silencio (bajo valor de energía promedio). El umbral se establece empíricamente.

2.2.6 Pre-énfasis

El filtro de pre-énfasis tiene el fin de aumentar la energía de alta frecuencia, para compensar la pérdida durante el mecanismo de producción del habla de los seres humanos, el cual posee un decaimiento de 6 dB por octava (Xiaoguo, X., 2013). Se usa un filtro digital de primer orden cuya función de transferencia es la siguiente:

$$H(z) = 1 - a z^{-1} \quad (3)$$

Por lo general el valor del coeficiente a se encuentra entre 0,9 y 1,0 (Xiaoguo, X., 2013).

2.2.7 Sistema digital de procesamiento de señal

2.2.7.1 Parametrización

La parametrización del sonido es la elección de varios valores, ya sea en plano temporal o frecuencial, para representar aspectos importantes de la señal de voz. Con el objetivo de disminuir considerablemente la tasa de información y ahorrando así espacio de memoria y

tiempo de procesamiento del sistema. Las técnicas de parametrización más utilizadas, en el reconocimiento de locutores son:

- LPC (Linear Predictive Coefficients) (Markel, J. D., & Gray, A. J., 2013) (Varela Rincón, J., & Loaiza Pulgarín, J. E., 2008).
- LPCC (Linear Predictive Cepstrum Coefficients).
- MFCC (Mel-Frequency Cepstrum Coefficients).

El cepstrum de una señal es el resultado de calcular la transformada de Fourier (Fourier Transform, FT) del espectro de la señal estudiada en escala logarítmica (dB). El nombre cepstrum deriva de invertir las cuatro primeras letras de spectrum. La razón principal para utilizar los coeficientes cepstrales es que tienen la ventaja adicional que uno puede derivar de ellos una serie de parámetros que son invariantes sin importar las distorsiones que puedan ser introducidas por el micrófono o por cualquier sistema de transmisión.

2.2.7.2 Coeficientes cepstrales en las frecuencias MEL

Los MFCC (Mel-Frequency Cepstrum Coefficients) son coeficientes para la representación del habla basados en la percepción auditiva humana, a diferencia de los LPC que se basan en la producción del habla, su principal característica es que las bandas de frecuencia están situadas logarítmicamente (según la escala Mel), que modela la respuesta auditiva humana más apropiadamente. En la Figura 3, se muestra el esquema para la obtención de los MFCC.



Figura 3. Esquema para la obtención de los MFCC.

El banco de filtros linealmente espaciado en la escala de Mel tiene en general forma triangular.

2.2.7.3 Ventaneo

Para realizar cualquier análisis derivado de la Transformada de Fourier de Corta Duración (Short-time Fourier transform, STFT), es imprescindible que la señal bajo estudio sea estacionaria en el dominio del tiempo, en el caso particular de las señales de voz humana es posible esta consideración si se toman tramos de 0,010 a 0,030 s (Rufiner, H. L., 2009) (Schaefer, R., W., Rabiner, L., R., 1975). Además el ventaneo resuelve el efecto de borde o fenómeno de Gibbs. Por otra parte, para evitar perder información, producto de la atenuación de los lóbulos secundarios es necesario realizar un solapamiento de las ventanas subsiguientes.

2.2.8 Vectores de observación o característicos

La red neuronal artificial (Artificial Neural Networks, ANN) utilizada para clasificar y determinar los locutores, debe ser alimentada con un vector que contiene la información vocal que representa de alguna manera una colección de características que describen de la mejor manera posible la voz humana. Estos vectores son conocidos en la literatura del SR como vectores de observación. En esta investigación se utiliza los coeficientes Mel, la primera y segunda derivada de los mismos, todos agrupados en un solo vector columna.

2.2.9 Red neuronal artificial

En esta investigación se utiliza una red neuronal artificial (Artificial Neural Networks, ANN) multicapas que dispone de un conjunto de neuronas agrupadas en varios niveles, con todas sus conexiones hacia adelante o “feedforward”, dando a una interconexión de todas las capas (Campbell Jr, J., P., 1997).

Un caso particular de este tipo de ANN son las redes Backpropagation, que es un tipo de red con aprendizaje supervisado, basado en minimizar la función de error mediante un método de descenso de gradiente. Una vez aplicado un patrón de entrenamiento a la entrada de la red, este se propaga desde la primera capa a través de las capas subsecuentes de la red (feedforward), hasta generar una salida, la cual es comparada con la salida deseada y se calcula una señal de error para cada una de las salidas, a su vez esta es propagada hacia atrás, empezando de la capa de salida, hacia todas las capas de la red hasta llegar a la capa de entrada, con la finalidad de actualizar los pesos de conexión de cada neurona, para hacer que la red converja a un estado que le permita clasificar correctamente todos los patrones de entrenamiento (Cruz Beltrán, L.; Acevedo Mosqueda, M., 2008). La principal desventaja de este tipo de redes, es que no se ha podido desarrollar un algoritmo confiable y lo suficientemente rápido por lo que se vuelve extremadamente lento para entrenar. Por ende, este tipo de redes debe de ser escogido cuando se necesiten altas velocidades de ejecución, pero los altos tiempos de entrenamiento no sean problema (Varela Rincón, J., & Loaiza Pulgarín, J. E., 2008), como es el caso de esta investigación.

3 Sistema propuesto

El enfoque del sistema de reconocimiento de locutores, planeado en esta investigación es del tipo identificación dependiente del texto, para un conjunto cerrado. Se presentan en la figura 4 los 3 módulos principales que componen al sistema SR.

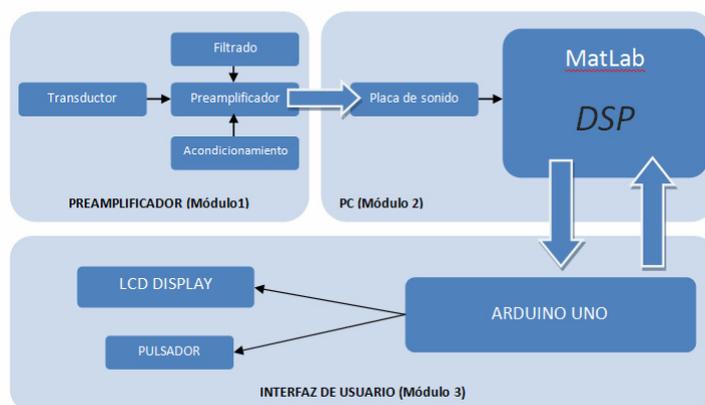


Figura 4. Módulos principales que componen el sistema de SR.

El sistema se compone de un módulo con un transductor y un preamplificador, un segundo módulo lo compone la placa de sonido propia de la computadora y por último, el tercer módulo es el encargado de realizar la comunicación máquina-usuario. En este punto es importante definir precisamente la diferencia entre usuario y administrador. Usuario es el sujeto a identificar por el sistema, que no necesariamente debe tener acceso a todo el sistema de SR. Por otro lado el administrador es el encargado de gestionar, habilitar y modificar el sistema SR, desde una interfaz gráfica.

Retomando al tercer módulo, el controlador principal de éste es un Arduino UNO, que permite la comunicación entre los distintos periféricos que componen al sistema en general. Para describir detalladamente el sistema SR que se propone en esta investigación, se divide al procedimiento en dos capas:

- Capa física: compuesto por el hardware del módulo 1 y 3.
- Capa de aplicación: compuesto por el software del módulo 2.

3.1 Capa física (Hardware)

3.1.1 Módulo 1

El primer módulo es un preamplificador para el micrófono tipo electret utilizado sistema, en vista que dicho transductor se encuentra a una distancia de 5 m de la placa de sonido de la computadora, es necesario amplificar y acondicionar la señal para mantener la relación señal ruido (SNR) por encima de 40 dB, como mínimo. El enlace del preamplificador y la PC se realiza por medio de un cable mallado para evitar la inducción del ruido de fuentes externas. El transductor seleccionado posee una sensibilidad ~ -45 dBV/Pa. Se dividió en dos las etapas de amplificación con el objetivo de no cargar en demasía solo uno de los operacionales. Cada una de estas etapas se encuentra limitada en frecuencia sobre un ancho de banda definido para la palabra (frecuencia de corte inferior ~ 100 Hz y frecuencia de corte superior ~ 9000 Hz). La ganancia de la primera etapa de amplificación es ~ 22 dB, la segunda etapa gana ~ 20 dB, dando una ganancia total ~ 42 dB.

3.1.2 Módulo 3

El tercer módulo de la capa física lo compone el microcontrolador Atmega328, embebido en la plataforma de hardware libre Arduino Uno. La conexión del LCD con el microcontrolador se realizó con el protocolo I²C.

3.2 Capa de aplicación (Software)

El menú principal se divide en varios elementos, como se presenta en la figura 5.

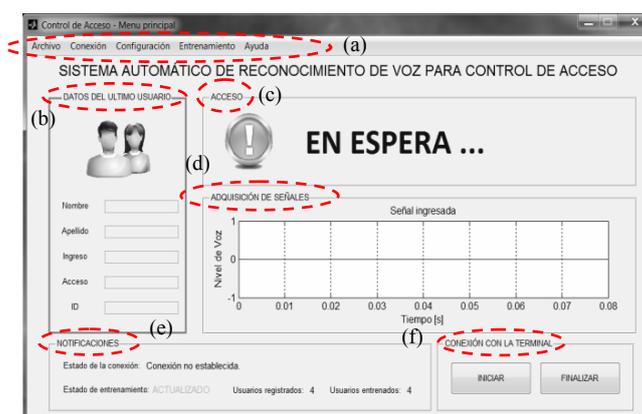


Figura 5. Pantalla principal (Menú principal) del software desarrollado. (a) Barra de submenús. (b) Datos del último usuario. (c) Acceso. (d) Adquisición de señales. (e) Notificaciones. (f) Conexión con la terminal.

- Barra de submenús: acceso a las herramientas de configuración y entrenamiento del sistema.
- Datos del último usuario: muestra los datos del último usuario que solicitó ingreso.
- Acceso: Indica el resultado del último acceso.
- Adquisición de señales: grafico en tiempo real de la señal adquirida.
- Notificaciones: informa el estado de la conexión y el estado del entrenamiento.

Una descripción más detallada de todos los elementos que componen la capa de aplicación se observa en la figura 6.

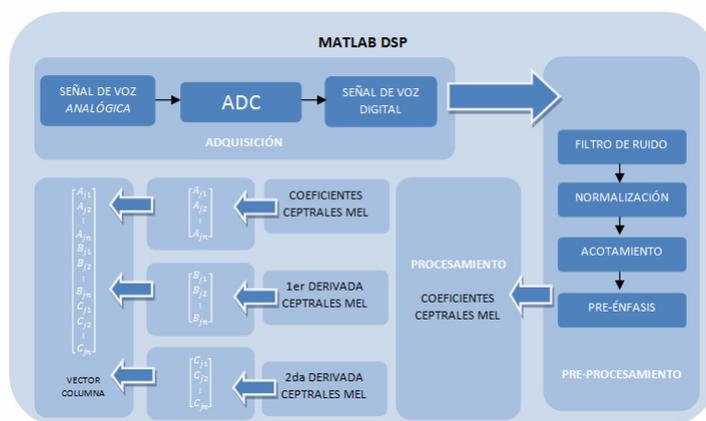


Figura 6. Descripción del módulo 2.

3.2.1 Adquisición

La adquisición en todo el sistema es de 16 bits y una cuantificación PCM uniforme. En la figura 7 se observa la señal adquirida por el sistema a una $F_m = 11025$ Hz y una duración 2 s. Además se observa que el nivel de ruido es bajo, obteniendo 40 dB de SNR.

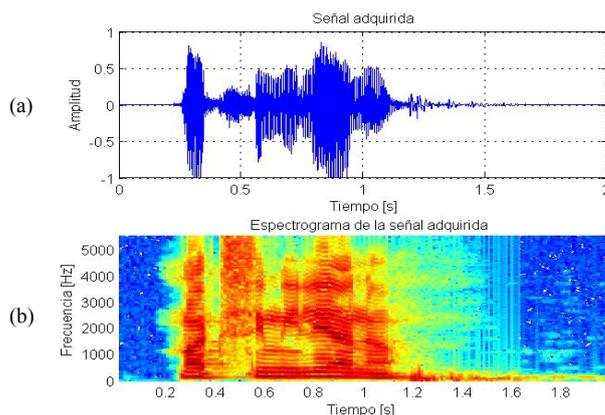


Figura 7. Señal adquirida por el sistema. (a) Ploteo en el dominio temporal. (b) Ploteo en el dominio temporal y frecuencial¹.

¹ El espectrograma se realizó con la función “spectrogram” de Matlab. Con una ventana temporal hamming de 0,020 s y un solapamiento del 50%.

3.2.2 Pre-procesamiento

3.2.2.1 Filtro de ruido

El filtrado de ruido se realiza a través de una técnica denominada *denoising*, la implementación se ejecuta a través del toolbox de Wavelet que dispone Matlab (Misiti, M.; Misiti, Y.; Oppenheim, G.; Poggi, J. M., 1996). En la figura 8, se presentan los resultados de la SNR para las wavelets madres Daubechies (dB) y Coiflets (coif) todas las pruebas se realizaron con umbral Hard y para los 14 niveles de descomposición (Alcon Choque, E. F., 2010).

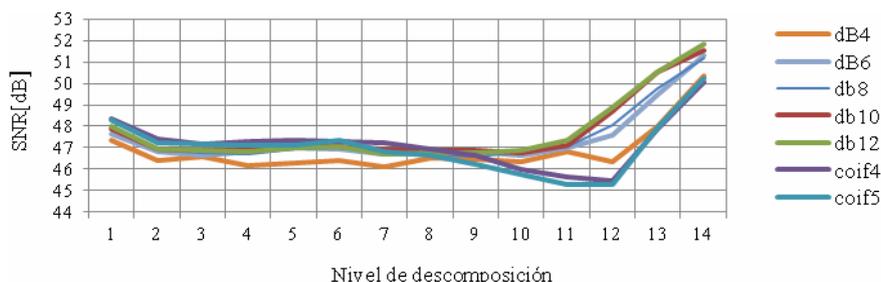


Figura 8. Resultados de filtrado con varias wavelet madres.

Las wavelets madres que mejor responden son: dB10 y dB12. En las pruebas de tiempos de ejecución en función del nivel de descomposición, los resultados fueron parejos en todos los casos. En conclusión en esta investigación se optó por realizar un filtrado con una wavelet madre dB12 con nivel de descomposición 6 y umbral del tipo Hard. A continuación se observan los resultados obtenidos (figura 9).

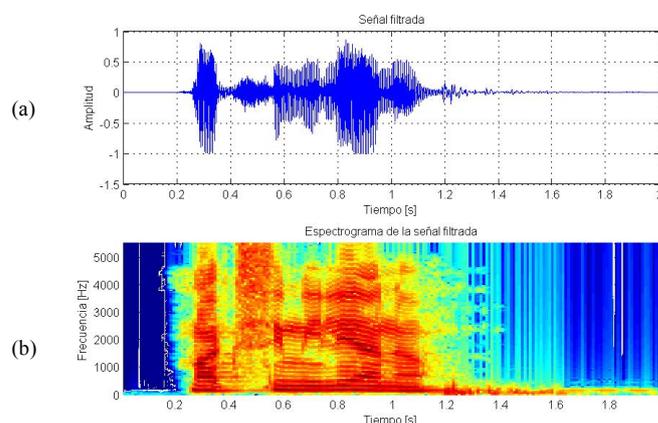


Figura 9. Aplicación del proceso de filtrado. (a) Señal en el dominio del tiempo. (b) Señal en el dominio del tiempo-frecuencia.

3.2.2.2 Acotamiento

La cantidad de muestras para analizar en el proceso de acotamiento lo define la variable ventana:

$$\text{ventana} = \text{round}(0,02321 * F_m) \quad (4)$$

donde *round* es la función de Matlab para obtener el entero más próximo del resultado del argumento.

El coeficiente 0,02321 establece el tamaño de la ventana en 0,020 s, por lo que no existe la posibilidad de que en el intento de eliminar silencio se elimine algún fonema (cuya duración ronda entre los 0,010 y 0,030 s). Si el valor de energía de la ventana analizada es mayor al *umbral* dado por la energía promedio, se conservan las muestras, en caso contrario se las reemplaza por ceros. Luego esos ceros son eliminados. En esta investigación se optó por un *umbral* de 0,09 (9 % de la energía global). El resultado de este proceso se observa en la figura 11.

3.2.2.3 Pre-énfasis

En esta investigación se optó por un coeficiente $a=0,9375$ para la ecuación (3) como figura en (Xiaoguo, X., 2013). El resultado de esto es una amplificación a razón de 5 dB por octava a partir de aproximadamente 2000 Hz.

3.2.3 Procesamiento

3.2.3.1 Ventaneo

Existen varios tipos de ventanas y cada una presenta una característica particular en tiempo y frecuencia, para este caso y en función de las referencias citadas la ventana Hamming es la que mejor resultados produjo. Por defecto el valor correspondiente al solapamiento es de $n/2$ (50 %) (McLoughlin, I., 2009) (con n = cantidad de muestras de la ventana), pero se pueden recurrir a otros valores. El ancho de la ventana, puede ser determinado por la cantidad de muestras o tiempo (en función de la frecuencia de muestreo F_m) por ejemplo a $F_m= 11025$ Hz una ventana de 256 muestras es equivalente a decir una ventana de aproximadamente 0,020 s. En conclusión una ventana más estrecha proporciona una buena resolución en tiempo pero pobre resolución en frecuencia, a medida que aumentamos el ancho de la ventana mejoramos la resolución en frecuencia. Ahora bien, la condición de estacionalidad de la señal de voz limita nuestro tamaño temporal de la ventana a 0,020 s, lo cual también condiciona a una resolución tiempo-frecuencia constante en todo su análisis, definido por el principio de incertidumbre de Heisenberg (Basso, G., 2006) (Rufiner, H. L., 2009). El ventaneo como la mayoría de las funciones de procesamiento se realizó con el toolbox VoiceBox². En esta investigación se optó por una ventana Hamming para el ploteo del espectrograma y en el caso de los coeficientes Mel, se utilizó una ventana triangular. El tamaño de dichas ventanas siempre es proporcional a la frecuencia, para que la cantidad de muestras siempre representan aproximadamente 0,020 s de la señal, esto se logra con la ecuación (4). Con respecto al solapamiento en todos los casos es del 50%.

3.2.4 MFCC

La implementación de este proceso se realizó con el toolbox VoiceBox, con los siguientes parámetros de configuración:

- $F_m = 11025$ Hz.
- Coef = 12.
- Overlap = Size/2 = %50.
- Windows = Triangular.
- Size = round(0,02321* F_m) ~ 256.

² VoiceBox es un toolbox de procesamiento de voz basado en rutinas de MatLab, creadas por Mike Brookes, del Departamento de Ingeniería Eléctrica y Electrónica, del Imperial College en Londres. Todas las funciones son públicas y libres bajo la licencia GNU Public License.

De esta manera se obtiene una matriz de tamaño 69x12 (Filas x Columnas), 69 por el tamaño temporal de la seal y 12 por la cantidad de coeficientes, en efecto esta matriz termina siendo un vector columna de 69x12=828 filas. En la figura 10 se representa graficamente esta matriz y el valor de los coeficientes.

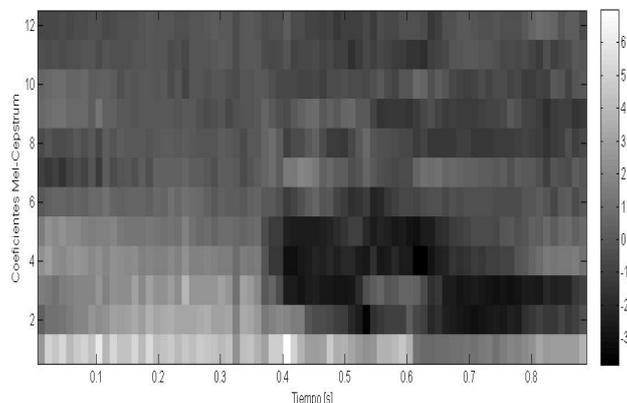


Figura 10. Representacion grafica de una matriz de coeficientes Mel.

Observado la figura 10, que es el caso particular de un solo usuario para una sola palabra aislada, es posible estimar una relativa continuidad en el valor de los coeficientes superiores a 6, se procede a analizar si esta distribucion es normal, por medio del test de Kolmogorov-Smirnov y de esta manera usar el valor medio como valor representativo de estos coeficientes en toda la seal. La tabla 1 muestra los resultados obtenidos de la prueba de normalidad. Se comprueba de esta manera que el nivel de significacion es mayor a 0,05 en todos los coeficientes, valor que determina que la distribucion es normal. Los resultados se obtuvieron en funcion de 10 repeticiones con el mismo usuario (reproduciendo la misma palabra, en distintas condiciones. De esta manera se resuelven dos cuestiones: en primer lugar, se estandariza un procedimiento que garantiza tener una cantidad constante y fija de coeficientes para alimentar la ANN, sin importa la longitud temporal. En segundo lugar, se disminuye considerablemente la cantidad de valores por cada coeficiente, por ejemplo de una matriz de 69x12 (como el anterior ejemplo) se pasa a un vector columna de 12x1, logrando no solo disminuir la cantidad de computos y de esta manera tener mayor versatilidad, sino que ademas, es posible agregar datos de la dinamica de coeficientes, realizando el mismo procedimiento anteriormente citado, pero con la primera y segunda derivada de los coeficientes. De esta manera se obtiene 3 vectores columnas de 12x1, que proporcionan informacion estatica y dinamica de los coeficientes de la seal. Es producto final de esta etapa un vector columna de 36x1 (la cantidad de filas depende de la cantidad de coeficientes x 3).

Tabla 1. Prueba de Kolmogorov-Smirnov para una muestra.

Coeficiente s	Coef_ 1	Coef_ 2	Coef_ 3	Coef_ 4	Coef_ 5	Coef_ 6	Coef_ 7	Coef_ 8	Coef_ 9	Coef_1 0	Coef_1 1	Coef_1 2
N	55,00	55,00	55,00	55,00	55,00	55,00	55,00	55,00	55,00	55,00	55,00	55,00
Parametros normales												
Media	-0,81	0,61	-1,17	-1,06	-0,48	-1,40	-0,70	-1,00	0,17	-0,96	-0,54	-1,08
Desviacion tipica	3,84	1,78	1,89	1,14	1,13	0,87	1,04	0,55	0,58	0,81	0,50	0,53
Z de Kolmogorov-Smirnov												
Sig. asintot. (bilateral)	0,07	0,25	0,28	0,48	0,13	0,73	0,82	0,51	0,97	0,39	0,94	0,96

3.2.5 ANN

Esta investigación propone utilizar una ANN con topología perceptrón multicapa con un aprendizaje supervisado del tipo Backpropagation. En este sentido, los parámetros fundamentales de la red son: el número de capas, el número de neuronas por capa, función de activación, cantidad de ciclos, el error mínimo, entrenamiento. A continuación se describen cada uno de los parámetros:

- **Número de capas:** Se definen 3 capas; entrada, oculta y salida (Varela Rincón, J., & Loaiza Pulgarín, J. E., 2008).
- **Número de neuronas por capa:** Cada capa posee un número particular de neuronas:
 - **Capa de entrada:** en principio, la cantidad de neuronas en la capa de entrada está definido por el vector característico, pero es posible modificar esta cantidad. En la figura 16, se muestra una representación de este proceso.
 - **Capa oculta:** El número de neuronas en la capa oculta tiene un alto impacto en el rendimiento de la ANN. Entre más neuronas tenga esta capa, más complejas son las superficies de decisión que se pueda formar y por ende mejor precisión se puede obtener al momento de la clasificación (Varela Rincón, J., & Loaiza Pulgarín, J. E., 2008).
 - **Capa de salida:** El número de neuronas en la capa de salida lo definen la cantidad de usuarios con los que se entrenó la ANN. En particular en esta investigación se utilizó una salida del tipo +1 (positivo) y -1 (negativo), por ejemplo el usuario 1 tiene una salida [+1 -1 -1 -1] (considerando solo 4 usuarios).
- **Función de activación:** La selección de estas funciones está basada en el hecho que las redes aprenden mucho más eficientemente cuando se usan funciones de transferencia simétricas (por ejemplo en rangos de [-1, 1]) en todas las neuronas que no sean de salida (incluyendo las neuronas de la capa de entrada) (Varela Rincón, J., & Loaiza Pulgarín, J. E., 2008).
- **Cantidad de ciclos:** En esta investigación se utiliza 500 épocas como cota superior de análisis.
- **Error mínimo:** En esta investigación se utiliza un error = 0,00001 como condición mínima.
- **Entrenamiento:** En esta investigación se utiliza un aprendizaje supervisado, basado en minimizar la función de error mediante un método de descenso de gradiente (Backpropagation), De la base de muestra adquirida por los usuarios. Se utiliza 50% de muestras para el entrenamiento y 50% para la validación.

La ANN fue implementada en el toolbox de MatLab (“Neural Network”), en esta investigación se utiliza Levenberg-Marquardt Backpropagation, un algoritmo de aprendizaje basado en un método de descenso de gradiente (Cruz Beltrán, L.; Acevedo Mosqueda, M., 2008).

4 Resultados obtenidos

4.1 Definición de la etapa de prueba

En la tabla 2, se presentan los parámetros constantes, detallando a que proceso corresponde cada uno.

Tabla 2. Parámetros fijos del sistema, durante las pruebas.

Proceso	Parámetro	Valor	Proceso	Parámetro	Valor
Adquisición	Fm	11025 [Hz]	Ventaneo	Tamaño de la ventana	~ 20ms
Adquisición	Resolución	16 [Bits]	Ventaneo	Overlap	%50
Adquisición	Duración	2 [s]	MFCC	Fm	11025 [Hz]
Adquisición	Cantidad de PIN por usuario	1	MFCC	Tipo de ventana (filtrado)	Triangular
Adquisición	Cantidad de repeticiones de PIN por usuario	10	MFCC	Cantidad de filtros	24
Filtro de ruido	Wavelet madre	dB 12	MFCC	Límite superior del filtro	5071,5 [Hz]
Filtro de ruido	Nivel de descomposición	6	ANN	Número de capas	3
Filtro de ruido	Umbral	Hard	ANN	Cantidad de neuronas en capa de salida	4
Acotamiento o silencio	Tamaño de la ventana	~ 20ms	ANN	Cantidad de ciclos (épocas) limite	500
Acotamiento o silencio	Coefficiente del umbral	0.09 (%9)	ANN	Error mínimo limite	1.00e-05
Pre-énfasis	Coefficiente del filtro	0.9375	ANN	Set de entrenamiento (porcentaje de muestras utilizadas del total)	%50 ³

En la tabla 3, se presentan los parámetros variables, detallando a que proceso corresponde cada uno.

Tabla 3. Parámetros variables del sistema, durante las pruebas.

Proceso	Parámetro	Valor
MFCC	Cantidad de coeficientes	12-20-25
ANN	Cantidad de neuronas en capa de entrada	2-12-36-72
ANN	Cantidad de neuronas en capa oculta	12-32-64
ANN	Funciones de activación (por capa)	Tansig – Logsig - Purelin

4.2 Etapa de prueba de topologías

El entrenamiento se llevó a cabo con 4 usuarios 2 hombres y 2 mujeres, cada uno repitió 10 veces su nombre⁴, para conformar la base de datos de entrenamiento. En la tabla 4 se

³ 50% entrenamiento y 50% validación.

⁴ En esta investigación no se consideró el grado de seguridad que tiene el nombre de una persona, la elección de este PIN solo fue para poner a prueba y evaluar el rendimiento del sistema.

muestran las distintas prueba realizadas, los valores de MSE (Mean Square Error) en el proceso de validación⁵ y la cantidad de épocas, son arrojadas por el Toolbox de Matlab. Las distintas pruebas se realizan, bajo el análisis planteado en (Varela Rincón, J., & Loaiza Pulgarín, J. E., 2008). Para lo cual en primer lugar se busca una configuración óptima, la misma es testada con dos conjuntos diferentes de muestras. El test CC se refiere al porcentaje de aciertos promedio en un conjunto cerrado (la validación se realiza con los mismos datos que se realiza el entrenamiento) y la medida CA indica el porcentaje de aciertos promedio en un conjunto abierto (la validación se realiza con muestras diferentes con las que se realiza el entrenamiento, situación que mejor se aproxima a la respuesta real del sistema, dado que las muestras utilizadas en este Test son totalmente distintas a las utilizadas para entrenar la ANN), en ambos casos se utilizan 10 muestras por usuario, dado un total de 40 muestras para el test CC y 40 muestras distintas para el test CA. La elección de estas seis pruebas se basó en comenzar con una configuración típica y propuesta en (Varela Rincón, J., & Loaiza Pulgarín, J. E., 2008), luego modificar los parámetros variables hasta logra disminuir el valor de MSE de la etapa de validación, a este valor se accede desde la pantalla de entrenamiento básico de la ANN en la opción de *Performance*.

Tabla 4. Pruebas de las distintas topologías.

Prueba #	Proceso	Parámetro	Valor	MSE	Épocas
Prueba 1	MFCC	Coefficientes	12	0.23244	7
	ANN	Neuronas en capa de entrada	36		
	ANN	Neuronas en capa oculta	12		
	ANN	Funciones de activación	Tansig - Tansig - Tansig		
Prueba 2	MFCC	Coefficientes	12	0.30325	9
	ANN	Neuronas en capa de entrada	36		
	ANN	Neuronas en capa oculta	32		
	ANN	Funciones de activación	Tansig - Tansig - Tansig		
Prueba 3	MFCC	Coefficientes	12	0.248	13
	ANN	Neuronas en capa de entrada	36		
	ANN	Neuronas en capa oculta	64		
	ANN	Funciones de activación	Tansig - Tansig - Tansig		
Prueba 4	MFCC	Coefficientes	12	0.11718	11
	ANN	Neuronas en capa de entrada	36		
	ANN	Neuronas en capa oculta	12		
	ANN	Funciones de activación	Purelin - Tansig - Tansig		
Prueba 5	MFCC	Cantidad de coeficientes	12	0.078937	7
	ANN	Neuronas en capa de entrada	36		
	ANN	Neuronas en capa oculta	64		
	ANN	Funciones de activación	Purelin - Tansig - Tansig		
Prueba 6	MFCC	Cantidad de coeficientes	30	0.036863	8
	ANN	Neuronas en capa de entrada	90		
	ANN	Neuronas en capa oculta	12		
	ANN	Funciones de activación	Purelin - Tansig - Tansig		

⁵ Los valores de MSE presentados en la tabla 11 son los obtenidos en el proceso de validación, que difieren sustancialmente del entrenamiento. Este valor de MSE representa la performance de la ANN.

Como se presenta en la tabla 4, el primer parámetro que se modifica para mejorar el rendimiento son la cantidad de neuronas en la capa oculta, sin lograr obtener grandes cambios. Si se observa detenidamente entre las primeras 3 pruebas, a medida que se aumenta la cantidad de neuronas en la capa oculta, fluctúa el valor de MSE sin obtenerse una clara tendencia. Es a partir de la prueba 4, que se retoma la configuración base (prueba 1) y se modifica solamente la función de activación de la capa de entrada por una función lineal (Purelin), a expensas que se sabe de antemano que los valores de entrada están acotados entre -1 y 1. Sorprendentemente el valor de MSE disminuye considerablemente. Para la prueba 5, se modifican la cantidad de neuronas en la capa oculta obteniendo aún mejores resultados. Por último, la prueba 6 presenta los mejores valores de MSE. En esta instancia se espera que la prueba 6 sea la que obtenga mejores resultados en el Test CC. En la tabla 5 se muestran los resultados de los porcentajes de acierto para ambos test. En dichos test se utiliza una base de 80 muestras distintas de voz⁶, 10 muestras por usuario, a un total de 4 usuarios, dando un total de 40 muestras distintas para cada Test.

Tabla 5. Pruebas de las distintas topologías.

Prueba #	Test CC	Test CA	Prueba #	Test CC	Test CA
Prueba 1	37/40 (92,5%)	27/40 (67,5%)	Prueba 4	39/40 (97,5%)	27/40 (67,5%)
Prueba 2	37/40 (92,5%)	27/40 (67,5%)	Prueba 5	39/40 (97,5%)	22/40 (55,0%)
Prueba 3	38/40 (95,0%)	25/40 (62,5%)	Prueba 6	40/40 (100%)	27/40 (67,5%)

En el caso del Test CC, el porcentaje de aciertos aumenta con cada modificación que se realiza en cada una de las pruebas, describiendo una tendencia que relaciona el valor de MSE con el porcentaje de acierto en el Test CC⁷. Los resultados del Test CA, no muestran una relación tan definida como fue el caso del anterior Test, en este Test se establece que el sistema tiene un porcentaje de acierto promedio de 64,58 %.

A continuación se establece un orden de mérito de cada prueba: Prueba 6, Prueba 3, Prueba 1, Prueba 2, Prueba 4, Prueba 5.

Por otra parte, los usuarios que presentaron más dificultad para su identificación son Florencia y Omar, con un porcentaje de acierto promedio de todas las pruebas de 51,6% y 55%, respectivamente. El resto de los usuarios presentaron valores por encima del 70%. En un sistema perfectamente entrenado y con una configuración precisa de la ANN, el vector de salida tendría que ser igual al de la etapa de entrenamiento. Pero en esta investigación no se logró dicha precisión es por este motivo que el sistema toma el máximo valor del vector y su posición para identificar a quien pertenece la muestra de voz. Esto se realiza como se muestra a continuación.

5 Conclusiones

Las conclusiones obtenidas de esta investigación son las siguientes:

⁶ La obtención de estas muestras, en primer lugar para el Test CC, se realiza con el mismo sistema de SR en la instancia que el usuario se registra. En segundo lugar para el Test CA, las muestras son obtenidas por un software externo respetando los parámetros de la tabla 2.

⁷ Para confirmar esta hipótesis, es necesario realizar un trabajo estadístico más exhaustivo, con el objetivo de comprobar que existe una correlación entre estos dos parámetros.

- ✓ De esta investigación se concluye indiscutiblemente que es posible construir un sistema de control de acceso, utilizando la voz como elemento de identificación, basado en los coeficientes cepstrales de frecuencia Mel para la caracterización de la voz y redes neuronales como sistema de clasificación.
- ✓ El porcentaje de acierto logrado con la mejor configuración (prueba 6), cumple con el objetivo planteado y verifica la hipótesis planteada en la sección
- ✓ En esta investigación se infirió que los coeficientes Mel de una señal de voz poseen una distribución normal, permitiendo de esta manera utilizar los valores medios como valores representativos de cada coeficiente, lo que se comprobó experimentalmente.
- ✓ La velocidad y el desempeño del sistema permite tomar este prototipo como base para futuras investigaciones. La estructura paralela de la ANN permite conseguir tiempos de ejecución del algoritmo muy cortos, a expensas de tiempos elevados en el entrenamiento de la ANN.
- ✓ Se demostró que las técnicas de filtrado de ruido con Wavelet responde correctamente frente a señales de voz, disminuyendo considerablemente las componentes de baja y alta frecuencia fuera del rango útil de la voz. Además se demostró que la técnica Denoising mantiene la forma original de la señal, sin alterar el dominio temporal.
- ✓ Se encontró que el rendimiento de la ANN está íntimamente relacionado con el entrenamiento de la misma. Además que uno de los parámetros que más repercute en el rendimiento son las funciones de activación de la capa de entrada y la capa oculta, junto a la cantidad de neuronas en la capa oculta.

6 Trabajos futuros

En esta investigación se pudo abordar diversos tópicos, dejando de lado otros, que a continuación se detallan:

- ✓ Si bien los valores de MSE obtenidos en la instancia de validación son bajos, queda como futura investigación configurar de forma más precisa las ANN, por medio de la ampliación de la cantidad de muestras de entrenamiento.
- ✓ Esta investigación no se orientó a optimizar el algoritmo para implementar en hardware dedicado, punto fijo o flotante. La inclusión de esta complejidad permite diseñar un producto comercial y de capacidad superior al presentado en esta investigación.
- ✓ En vista que la señal de voz es descompuesta por técnica de wavelet, estos mismos elementos podrían alimentar la ANN, de manera separada o conjuntamente con los coeficientes Mel, utilizados en esta investigación.
- ✓ En esta investigación no se realizaron grandes consideraciones del ruido que podría afectar al sistema. Por este motivo se propone agregar un grado de complejidad al sistema de filtros, no solo para mitigar estas condiciones, sino además para contemplar una situación real de uso, obteniendo de esta manera un sistema mucho más robusto que el presentado en esta investigación.
- ✓ Incorporar al sistema la posibilidad de falsos usuarios o impostores.
- ✓ Implementar el uso de otros PINs de identificación, así como complementarlo junto a otro medio de ingreso de información, por ejemplo: teclado o detector de huella.
- ✓ A la hora de incrementar el porcentaje de aciertos del sistema, es necesario contemplar aspectos morfológicos de la voz de cada usuario.

Referencias

- Alcon Choque, E. F. (2010). *Filtrado de señales de voz a través de Wavelet*. Facultad de Ciencias Puras y Naturales, Universidad Mayor de San Andrés, La Paz, Bolivia.
- Basso, G. (2006). *Percepción auditiva*. Editorial de la Universidad Nacional de Quilmes.
- Borrás, O. (2006). *Reductor de ruido mediante resta espectral en entorno Matlab*. Universidad Politécnica de Madrid, Proyecto Fin de Carrera, EUIT Telecomunicación Universidad, Madrid, España. Obtenido de http://oa.upm.es/954/1/PFC_ORIOL_BORRAS_GENE.pdf.
- Campbell Jr, J., P. (1997). Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9), 1437-1462.
- Cruz Beltrán, L.; Acevedo Mosqueda, M. (2008). Reconocimiento de voz usando redes neuronales artificiales backpropagation y coeficientes lpc. In *6to Congreso Internacional de Cómputo en Optimización y Software* (págs. 89-99). CiCos.
- Floyd, T., L. (2006). *Fundamentos de sistemas digitales*. Madrid, España.
- Kouro, S., & Musalem, R. (2002). *Tutorial introductorio a la Teoría de Wavelet. Artículo presentado como trabajo de la asignatura Técnicas Modernas en Automática*. Departamento de Electrónica, Universidad Técnica Federico Santa María, Valparaíso, Chile.
- Mammone, R. J., Zhang, X., & Ramachandran, R. P. (1996). Robust speaker recognition: A feature-based approach. *Signal Processing Magazine*, 13(5), 58.
- Markel, J. D., & Gray, A. J. (2013). *Linear Prediction of Speech* (Vol. 12). Springer Science & Business Media.
- Marquina, A. Á. (2001). *Fundamentos del Reconocimiento Automático de la Voz, Algoritmos de Extracción de Características*. Universidad Politécnica de Madrid, Facultad de Informática, España.
- McLoughlin, I. (2009). *Applied speech and audio processing: with Matlab examples* (Vol. 86). Cambridge University Press.
- Misiti, M.; Misiti, Y.; Oppenheim, G.; Poggi, J. M. (1996). *Wavelet toolbox user's guide*. 2° ed., The Math Works, The Math Works Ins.
- Miyara, F. (1999). *La voz humana*. Laboratorio de Acústica y Electroacústica, Escuela de ingeniería, Electrónica, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario, Rosario, Santa Fe, Argentina. Obtenido de <http://www.fceia.unr.edu.ar/prodivoz/fonatorio.pdf>
- Moreno, P. J.; Stern, R. M. (1994). Sources of degradation of speech recognition in the telephone network. *CASSP-94. 1*, pág. 109. IEEE International Conference.
- Reynolds, D., A. (1996). M.I.T. Lincoln Laboratory site presentation. *NIST Speaker Recognition Workshop*, 27-28.
- Rufiner, H. L. (2009). *Análisis y modelado digital de la voz: técnicas recientes y aplicaciones* (1 ed.). Santa Fe, Argentina: Ediciones UNL, Colección Ciencia y Técnica.
- Sanchez Marin, G., D. (2004). *Segmentación y realce de señales de voz usando la transformada Wavelet y DSP's*. Tesis de grado, Facultad de Ingeniería, Universidad del Quindío, Armenia, Quindío, Colombia.
- Schaefer, R., W., Rabiner, L., R. (1975). Digital representation of speech signals. *Proc. IEEE*, 63(4), 662-677.
- Varela Rincón, J., & Loaiza Pulgarín, J. E. (2008). *Reconocimiento de palabras aisladas mediante redes neuronales sobre FPGA*. Tesis de grado, Ingeniería Eléctrica, Universidad Tecnológica de Pereira, Pereira, Colombia.
- Xiaoguo, X. (2013). *Joint speech and speaker recognition using neural networks*. Bachelor's thesis, Electrical Engineering, Novia university of applied sciences, Vaasa, Finland.