TÉCNICAS PARA LA VISUALIZACIÓN DEL HABLA

Rodríguez A. Luis Miguel
cronos987@gmail.com

León V. Rubén D.
rleon@fie-espe.edu.ec

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA
ESCUELA POLITECNICA DEL EJÉRCITO

Abstract

The present document describes a new system for speech visualization creating readable patterns by the integration of different characteristics of the speech in a single picture. The system generates the picture starting from the extraction of the acoustic characteristics of the signs of the speech discriminating against them through four neural nets (Backpropagation). The training of the neuronal nets carries out it entering 4 types of representative parameters of the speech signals. The outputs of the neural nets control each pattern's visual intensity used for the vowels and selected consonants, that is, each consonant possesses a distinctive pattern and each vowel has its own color. All this is possible thanks to the design of the image synthesis unit.

This proposal constitutes an important beginning the field of the communication with human beings with auditory deficiencies and the process for their integration to the community.

Resumen

Este artículo plantea y describe un nuevo sistema para la visualización del habla creando patrones legibles gracias a la integración de diferentes características de la misma en una sola imagen. El sistema genera la imagen a partir de la extracción de las características acústicas de las señales del habla discriminándolas a través de cuatro redes neuronales (Backpropagation).

El entrenamiento de las redes neuronales se lo llevo a cabo a partir de cuatro tipos de parámetros representativos de las señales del habla. Las salidas de salida de las redes neuronales controlan la intensidad visual de cada patrón usado para las vocales y consonantes seleccionadas, es decir, cada consonante posee un patrón distintivo, así como a cada vocal su propio color. Todo esto es posible gracias al diseño de la unidad sintetizadora de imagen.

Esta propuesta constituye un inicio importante para el campo de la comunicación con seres humanos con deficiencias auditivas y el proceso de su integración a la comunidad.

1. Introducción

Las señales del habla han sido estudiadas por varias razones y aplicaciones por muchos investigadores por muchos años. Algunos estudios han segmentado la señal del habla en pequeñas porciones llamadas fonemas [1]. Es en este punto donde se empezara a describir la señal del habla en términos de sus características generales a través de la visualización de la misma incluyendo sus características acústicas y fonéticas en una sola imagen.

A lo largo del siglo pasado varias propuestas han sido presentadas para la visualización del habla, una de ellas es la denominada “visible speech”, la cual es utilizada actualmente para el análisis de los espectrogramas de sonido. Además, han sido propuestos el “correlatogram” que utiliza la función de autocorrelación, el “intervalgram” que utiliza la rata de cruce por cero y la “wave collation visual speech display” que se basa en métodos de pitch-sincrónicos. La legibilidad de los espectrogramas del habla han generado resultados, en los cuales el usuario adquirió una habilidad excelente para “visualizar el habla”, después de que él había invertido entre 2000–2500 horas leyendo espectrogramas; su habilidad de lectura era excelente a pesar de estar leyendo materiales casi desconocidos. Esto mostró que el cerebro humano puede extraer características fonéticas de los espectrogramas del habla a través de vastas horas de entrenamiento [2].

Históricamente, se ha utilizado el periodograma como el método más común para la visualización de las señales del habla. Este se forma tomando la Discrete Fourier Transform

(DFT) de un segmento ventaneado del habla, y encontrando el modulo al cuadrado de cada valor complejo de la salida. De esta manera, el periodograma proporciona una estimación de la Power Spectral Density (PSD), la cual es solo degradada por los efectos espectrales de la ventana temporal con la que se segmento. La resolución en frecuencia del periodograma es inversamente proporcional a la longitud de la trama de entrada [3].

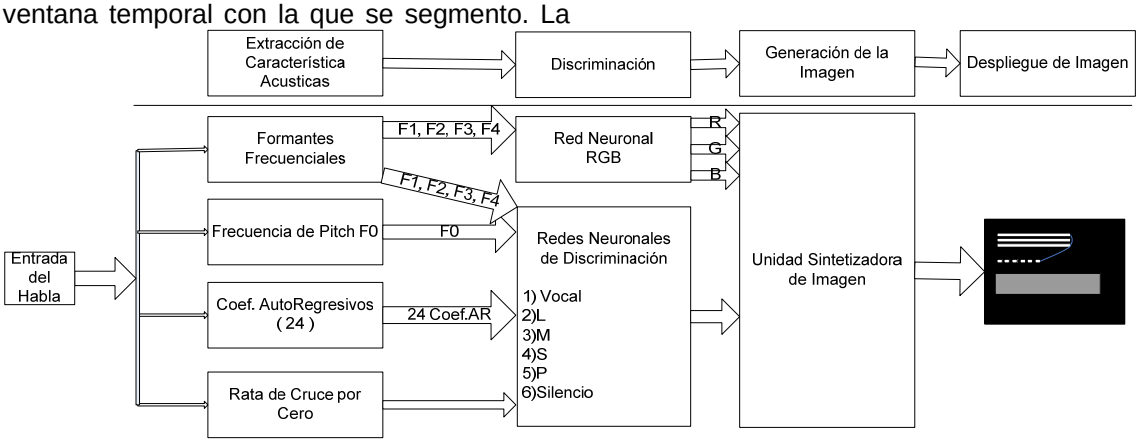


Figura 1 Diagrama de bloques del sistema para la visualización del habla

Otro método muy utilizado usa los coeficientes de predicción lineal ya que el modelamiento Auto Regresivo (AR) de las señales del habla suministra una descripción muy concisa de la función de transferencia del tracto vocal [3].

Con esto en mente, primero se desarrollará un sistema que permita extraer características acústicas del habla como el pitch y sus formantes frecuenciales, para que de esta manera se pueda centrar todo el esfuerzo hacia la creación de la unidad sintetizadora de imagen cuyo objetivo es obtener un verdadero sentido visual orientado al entendimiento del habla.

Con esta idea, el presente trabajo tiene como objetivo proporcionar un sistema que permita la visualización del habla y que haga posible leer secuencias de fonemas después de un periodo relativamente corto de entrenamiento.

2. Características del Sistema de Visualización del Habla

En la Figura 1 se muestra el diagrama de bloques del sistema propuesto para la visualización del habla. Aunque este sistema esta basado en conceptos similares a sistemas previos [2], [4], [5], el presente sistema cuenta con una red neuronal que controla y transforma las componentes de frecuencias de vocales, en idioma castellano, a colores (RGB), así como también con la unidad sintetizadora de imagen orientada a proporcionar un verdadero entendimiento visual del habla en el idioma castellano. La operación del sistema se describe a continuación:

Las cuatro primeras formantes frecuenciales se extraen trama a trama de la señal del habla y con una red neuronal son convertidas en tres señales que representan colores primarios (RGB). En la unidad sintetizadora de imagen y con la ayuda de una de las salidas de las redes neuronales de discriminación (1 en la Figura 1) se controla que solo las vocales sean representadas con color en la imagen desplegada al final. Los coeficientes autoregresivos son extraídos de cada trama de entrada además de la rata de cruce por cero. Las salidas de las redes neuronales de discriminación (1, 2, 3, 4, 5, 6 en la Figura 1) junto con las señales de color (RGB) ingresan a la unidad sintetizadora de imagen para que esta se encargue de crear una imagen que represente en forma visual los valores respectivos a cada categoría ya sea vocal o consonante.

Para extraer las cuatro primeras componentes frecuenciales de cada trama se emplea el estimador espectral por el método de covarianza modificada, este método ajusta un modelo autoregresivo (AR) a la señal, minimizando los errores de predicción directo y reverso que a su vez minimiza el error cuadrático medio total. Una vez hecho esto se ubico la posición de los picos correspondientes a las cuatro primeras componentes frecuenciales. Por otro lado, el método para lograr identificar las consonantes es uno de los mayores inconvenientes a ser superados. Las

consonantes en general al carecer de características tan distintivas como las que tienen las vocales no aportan suficiente información de si mismas para ser identificadas de manera tradicional, lamentablemente tampoco en la

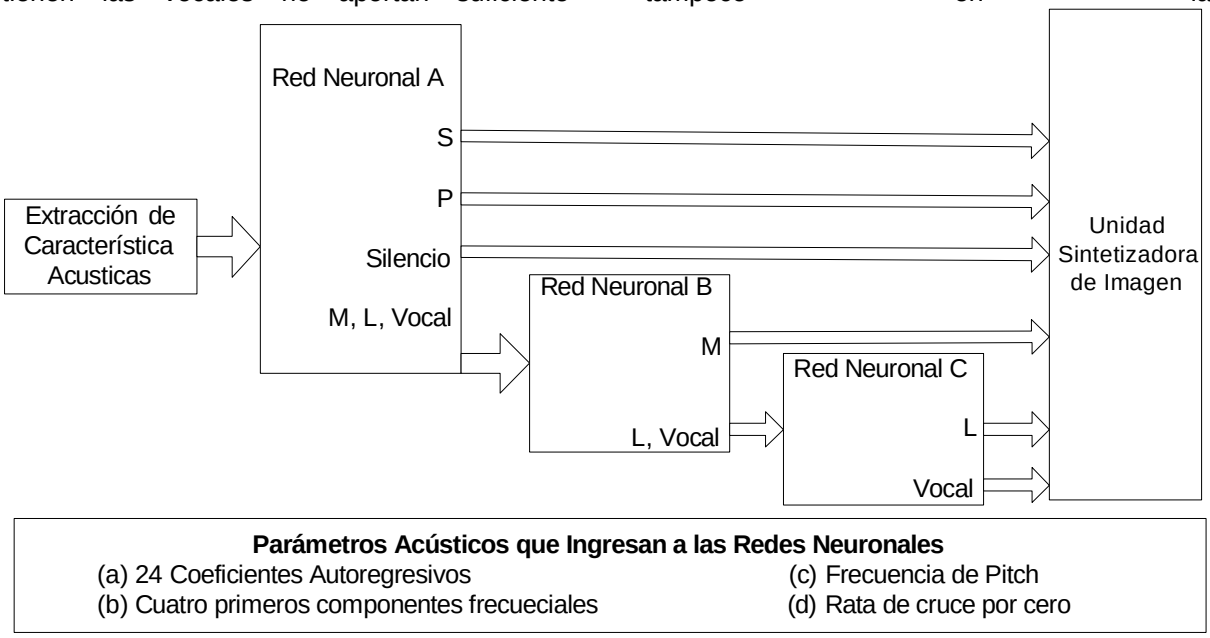


Figura 2 Redes Neuronales de discriminación

presente investigación se posee un método o algoritmo que las logre discriminar con alta confiabilidad. Para este fin se ha utilizado redes neuronales que son de gran utilidad gracias a su característica de aprender las distintas y sutiles diferencias que existen entre consonantes.

Por ultimo, la unidad sintetizadora de imagen despliega los patrones y colores predefinidos para cada vocal o consonante dependiendo del grado de reconocimiento que esta ha obtenido en las redes neuronales previas, es decir, que se asignara menor brillo a los patrones con un índice de reconocimiento bajo y gran brillo a los patrones con un índice de reconocimiento alto.

3. Diseño de las Redes Neuronales

Los fonemas preseleccionados en conjunto contienen tres parámetros de distinta naturaleza, que son: vocales, consonantes y silencio. Una vez que se determina estas categorías se procede a añadir subdivisiones en las vocales y consonantes. Finalmente teniendo ya todas las categorías, las redes neuronales ya pueden ser entrenadas para poder discriminar entre una categoría de otra.

Para este fin, se crearon cuatro redes neuronales Backpropagation de tres capas, ya que estas tienen la habilidad de aprender funciones no lineales complejas, para la toma de decisión [6]. A continuación se enumeran las redes neuronales creadas las que se ilustran en la figura 2:

1) Red Neuronal RGB, esta red neuronal es entrenada para convertir las cuatro señales de las componentes frecuenciales en tres colores primarios, rojo, verde y azul. De esta manera a cada vocal le corresponde un determinado color que puede ir variando en su tonalidad dependiendo de las dimensiones y la forma del tracto vocal del locutor.

2) Red Neuronal A, esta red es entrenada para clasificar las características acústicas de su entrada en cuatro categorías.
La S, la P, el silencio y la última clasifica en conjunto a la M, L y vocales.

3) Red Neuronal B, esta red es entrenada para clasificar las características acústicas que pertenecen a la cuarta categoría de la Red Neuronal A (M, L, y Vocal) en dos.
La M y la última clasifica en conjunto a la L y vocales.

4) Red Neuronal C, esta red es entrenada para clasificar las características acústicas de la segunda categoría de la Red Neuronal B (L y Vocal) en dos, lógicamente las categorías son:
La L y vocales.

La Figura 2 muestra las tres redes neuronales utilizadas para discriminar vocales, consonantes (S, P, M y L) y el silencio. Los parámetros que ingresan como entradas a las redes neuronales son de cuatro tipos, los

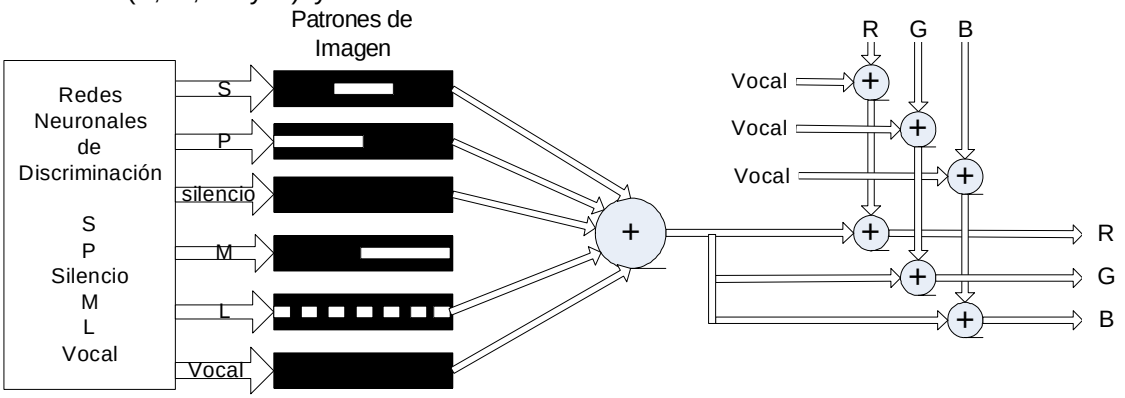


Figura 3 Esquema de la unidad sintetizadora de imagen

cuales contienen 30 características acústicas en total. El habla ha sido muestreada a 12KHz y las características acústicas extraídas de cada trama. La longitud de la trama es de 256 muestras. Cabe recalcar que las salidas de las redes neuronales como se muestra en la figura 2 son las entradas de la unidad sintetizadora de imagen.

Para determinar las apropiadas condiciones para la creación de las redes neuronales se realiza experimentos preliminares y comparándolos con sistemas previos [2], [6], se concluyo que cada red neuronal necesita 30 unidades en la primera capa oculta y quince en la segunda. Además el método escogido para acelerar la convergencia fue de back-propagation [7], definiendo un valor de 0.8 para el momentum y una rata de aprendizaje igual a 0.2. Para la función de activación para cada perceptron se utiliza la usual función de Log-sigmoid debido a sus propiedades [6] y por último el error de convergencia máximo permitido es de 0.001.

4. Unidad Sintetizadora de Imagen

La unidad sintetizadora de imagen representa las consonantes, vocales y el silencio a través de colores y patrones de imagen predefinidos. Lo primero fue el diseño y creación de patrones de imagen (bitmap) para cada categoría perteneciente a las consonantes. Por ejemplo para la consonante M se creo una barra horizontal posicionada a la derecha, para la L una línea entrecortada, todos los patrones correspondientes a cada categoría se muestran en la figura 3.

Las proporciones que cada patrón de imagen va a tener en la imagen final desplegada es controlado por las salidas de las redes neuronales de discriminación, es decir, el patrón que se mostrará con más claridad en la imagen final será el que haya obtenido el mayor índice de reconocimiento en las redes neuronales, así de esta manera, los patrones opacos y con menos brillo indicaran lo contrario.

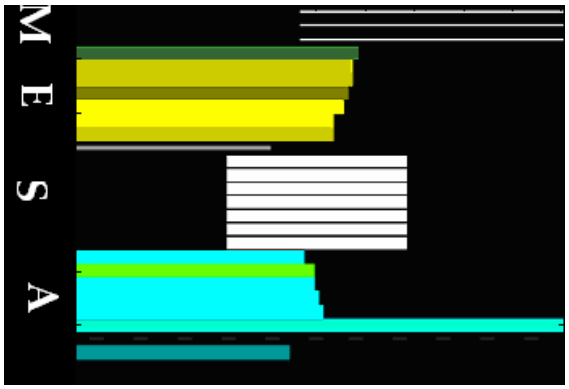
En la figura 3 los patrones de imagen correspondientes a las vocales y al silencio son exactamente iguales, estos patrones se van a diferenciar uno de otro en la imagen final ya que al patrón de las vocales se le agrega color (RGB) y en el caso del silencio no. Es debido a este hecho que los únicos patrones en poseer color en la imagen desplegada al final del sistema son los correspondientes a las vocales. El color que se agrega a cada patrón de imagen de vocal esta especificado en las tres salidas (R G B) de la Red Neuronal RGB.

Por último, a cada patrón de imagen que se va obteniendo en el sistema se lo acopla con los anteriores para que de esta manera se cree la imagen completa representativa del fonema ingresado. Una vez terminado todo este proceso la imagen final es desplegada en pantalla conteniendo tantos patrones de imagen como tantas tramas de 256 muestras puedan generarse con la señal analizada.

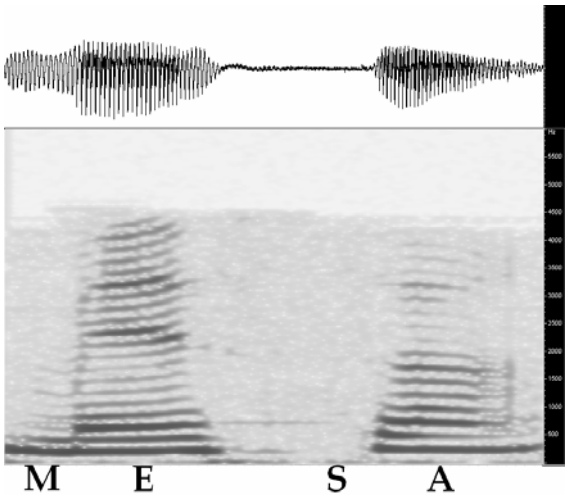
La imagen del habla desplegada fluye desde el tope de la pantalla hacia la parte inferior de la misma, es decir, comienza en la parte superior de la pantalla y termina en su parte inferior.

5. Despliegue de Ejemplos de Palabras Castizas y Comparación con los Espectrogramas del Habla

Con el fin de ilustrar el presente sistema se han agregado cuatro ejemplos que despliegan los resultados para cuatro diferentes fonemas ingresados, además de sus respectivas funciones temporales y sus espectrogramas

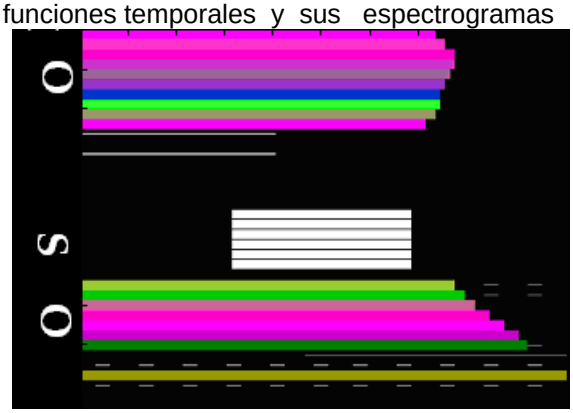


(a) Visualización del habla propuesta

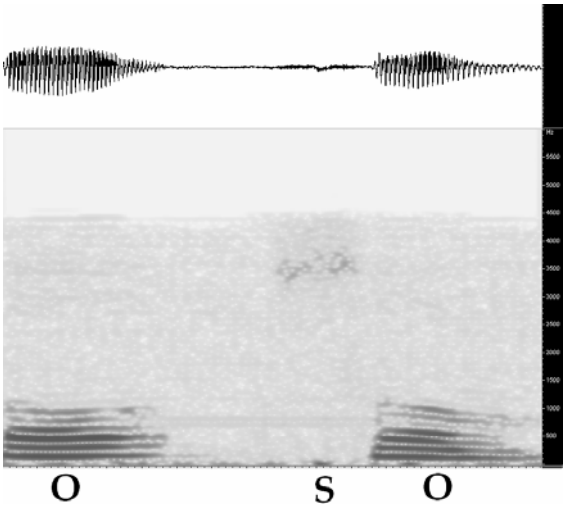


(b) Forma de Onda y Espectrograma

Figura 4 Ejemplo de la visualización del habla del fonema “mesa” en comparación con el espectrograma



(a) Visualización del habla propuesta



(b) Forma de Onda y Espectrograma

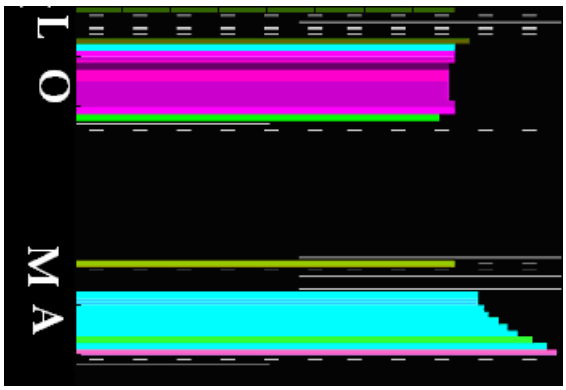
Figura 5 Ejemplo de la visualización del habla del fonema “oso” en comparación con el espectrograma

En las figuras 4-7, cabe indicar que uno de los cuatro fonemas mostrado fue producido por una mujer y los otros tres producidos por un hombre.

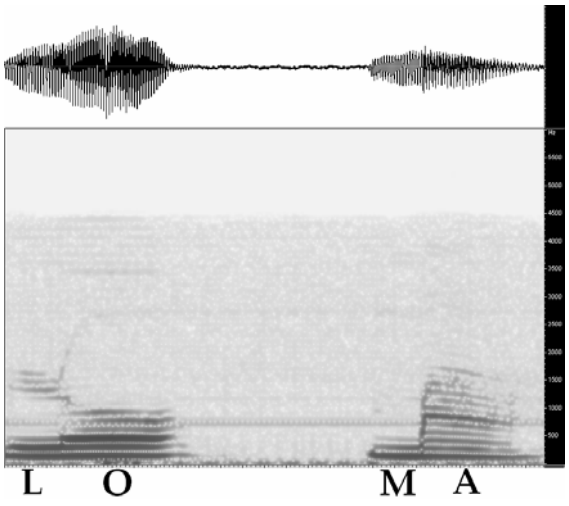
En las figuras 4-7 las imágenes correspondientes a la visualización del habla propuesta, el eje vertical representa tiempo y la longitud horizontal de los patrones de colores indica el pitch. Como se puede observar en las figuras el pitch varia con el tiempo. El color y su tonalidad representan las características y cualidades de la vocal, por otro lado, los patrones de imagen junto con su posición indican que consonante es. El brillo de los patrones de imagen correspondientes a las consonantes indican el grado de certeza que tiene el sistema al momento de discriminar entre una consonante y otra, es decir, a mayor brillo es más probable que el patrón de imagen que se muestra este indicando lo correcto.

La figura 4 fue creada a partir de un fonema (“mesa”) producido por una mujer mientras que las figuras 5-7 se crearon a partir de fonemas producidos por un hombre. A través de la comparación entre la figura 4 y las figuras 5-7 se puede observar que las longitudes de los patrones de color en la figura 4 son apreciablemente más cortas que los que se encuentran en las figuras 5-7, esto nos indica que el pitch de la figura 4 es menor que los

otros, esto concuerda perfectamente al saber que provino de una mujer.



(a) Visualización del habla propuesta

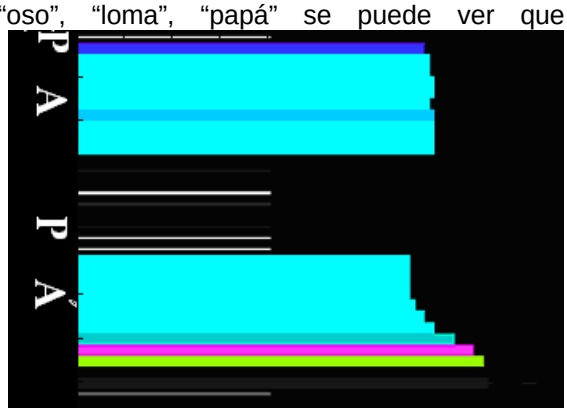


(b) Forma de Onda y Espectrograma

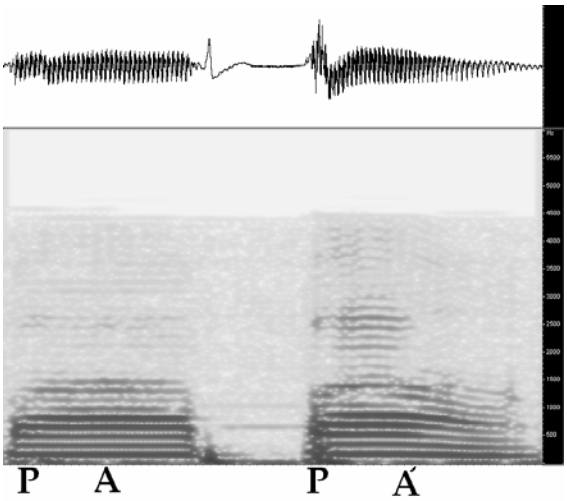
Figura 6 Ejemplo de la visualización del habla del fonema “loma” en comparación con el espectrograma

existe cierta confusión por parte del sistema al momento de catalogar las ultimas tramas de los fonemas, esto se debe a que en estas secciones la forma onda siempre tiende a tener menor energía. En las figuras 4-7 se observa que los patrones de color presentan buena uniformidad en su color, cualidad primordial para lograr un verdadero sentido de identificación de la vocal representada, a pesar de que tienen ciertos errores especialmente al iniciar y finalizar la vocal. Cabe indicar que el sistema propuesto no contiene ningún tipo de simbología como letras o número para lograr la visualización del habla. En las figuras 4-7 se ha agregado las palabras junto a la visualización del habla propuesta solo con fines didácticos.

Se observa en las figuras 4-7 las siguientes características. Casi no existe ningún error en la palabra “mesa”, pero en las otras palabras “oso”, “loma”, “papá” se puede ver que



(a) Visualización del habla propuesta



(b) Forma de Onda y Espectrograma

Figura 7 Ejemplo de la visualización del habla del fonema “papá” en comparación con el espectrograma

En el caso de los espectrogramas del habla, es a menudo difícil leer los fonemas correctamente aparte del contexto porque el movimiento de las formantes juega un papel importante en la segmentación visual. Será difícil para los lectores ordinarios leer las vocales y algunas consonantes que son influenciadas por las diferencias individuales de los portavoces, a menos que se representen las pronunciaciones en una hoja o una pantalla. [2]

El sistema propuesto tiene una menor dependencia del contexto y la pronunciación individual, que los espectrogramas del habla

tienen. Las diferencias anteriores entre la visualización propuesta y los espectrogramas del habla influirán en la legibilidad intuitiva [2].

6. Conclusiones

Se ha creado un sistema para la visualización del habla integrando redes neuronales (Backpropagation), de las cuales tres sirvieron para la discriminación de las vocales, el silencio y las consonantes seleccionadas, y una red neuronal para la creación de las señales de color (RGB) en base de las cuatro primeras componentes frecuenciales de la señal extraídas trama a trama.

A continuación, se diseñó y creó la unidad sintetizadora de imagen, unidad que agrupa todas las características entregadas por las redes neuronales y las convierte en una imagen gracias a la integración de patrones de imagen predefinidos. Así, de esta manera, fue posible desplegar una imagen representativa del fonema ingresado, en pantalla. Varias pruebas realizadas muestran que el color de cada vocal en la imagen final se muestra de forma clara y uniforme además de existir una buena distinción entre patrones de vocales y consonantes.
De acuerdo a una inspección visual y algunas pruebas de lectura, la presente forma de visualización entrega una imagen que puede ser entendida fácilmente.

El presente sistema puede ser de gran utilidad en el campo de la comunicación con seres humanos con deficiencias auditivas, ya que puede proporcionar algunas claves para ayudar a entender los sonidos producidos del habla, además, será de utilidad en nuevas aplicaciones así como para las convencionales como es el caso del entrenamiento y la transmisión del habla. Por ejemplo, se puede emplear discos compactos en los cuales se graben las señales del habla con sus respectivas imágenes, siendo de esta manera una herramienta para el entrenamiento de personas, requiriendo únicamente una computadora personal y audífonos.

La presente propuesta constituye un inicio importante dentro del campo de la comunicación con seres humanos con deficiencias auditivas y el proceso de su integración a la comunidad.

Referencias Bibliográficas

[1] Kondoz A. M (2004). “Digital Speech Coding for Low Bit Rate Communication Systems” Segunda Edición, Wiley.

[2] A. Watanave, Tomishige S. y Nakatake M., “Speech Visualization by Integring Features for the Hearing Impared” *IEEE Trans. On Speech and Audio Processing.,* vol. 8, 454-466, 2000

[3] Baghai-Ravary L. y Beet S., “Multistep Coding of Speech Parameter for Compression” *IEEE Trans. On Speech and Audio Processing,* vol. 6 No. 5, 1998.

[4] A.Watanabe, Y. Ueda, y A. Shigenaga, “Color displaysystem for connected speech to be used for the hearing impaired,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 164–173, 1985.

[5] A.Watanabe y Y. Ueda, “Speech visualization and its application for the hearing impaired,” *J. Acoust. Soc. Amer.*, vol. 84, p. 42, 1988.

[6] Sada Siva Sarma A. y Strube H.W, “Hindi Phoneme Recognition Using Time Delay Neural Network”.

[7] R. P. Lippmann, “An introduction to computing with neural nets,” *IEEE Acoust., Speech, Signal Process. Mag.*, pp. 4–22, Apr. 1987.

Biografías

Rodríguez A. Luis Miguel
Nació en Quito en 1983. Realizo sus estudios secundarios en el Colegio Cardenal Spellman de Varones. Obtuvo el titulo de bachiller con especialidad Físico-Matemáticas en 2001.Entre 2001 y 2006 estudió en la Escuela Politécnica del Ejército, en este momento esta realizando la tesis profesional sobre “Técnicas para la Visualización del habla”.

León Vásquez Rubén Darío
Nació el 30 de Abril de 1962 y obtuvo su título de Ing. Electrónico en la ESPE en 1985, su grado de Magíster en Ciencias en Brasil en 1992 y sus áreas de interés son el Procesamiento Digital de Señales, Análisis Espectral Digital y su aplicación en los Sistemas de Telecomunicaciones.