ISBN: 978-987-4998-33-0
DOI:https://doi.org/10.33414/AJEA.4.408.2019

Clasificación automática de la calidad vocal

Automatic vocal quality classification

Presentación: 22/11/2019

Doctorando:
Mario Alejandro García
Universidad Tecnológica Nacional Facultad Regional Córdoba
mgarcia@frc.utn.edu.ar

Director/es:
Eduardo A. Destéfanis

Resumen

Se presenta un enfoque para la construcción de un clasificador extremo-a-extremo de la calidad vocal en escala GRBAS basado en redes neuronales profundas. En base a este enfoque se muestran tres redes neuronales. Las redes presentadas calculan la transformada de Fourier de término reducido (STFT), el cepstrum y shimmer de una señal de audio. Las redes neuronales que calculan la STFT y shimmer se logran entrenar correctamente, mientras que la que calcula el cepstrum no. Para este último caso, se plantea una solución alternativa al cepstrum, la *autocovariance*, que sí se puede entrenar. Se concluye que las redes neuronales desarrolladas son compatibles con el enfoque planteado porque permiten que el gradiente del error se propague hacia atrás, condición necesaria para entrenar el modelo completo.

Palabras claves: Aprendizaje Profundo, Redes Neuronales Artificiales, Calidad ocal

Abstract

In order to classify the vocal quality on GRBAS scale, an approach of end-to-end neural network design is presented. Based on this approach, three neural networks are shown. These neural networks calculate the short-term Fourier transform (STFT), cepstrum and shimmer of an audio signal. The training of the networks that calculate STFT and shimmer was successful. The network that calculates the cepstrum could not be trained, but an alternative model that calculates the autocovariance could. It is concluded that the developed neural networks are compatible with the proposed approach. This is because they allow the error gradient backpropagation, a necessary condition for training the complete model.

Keywords: Deep Learning, Artificial Neural Network, Vocal Quality

Introducción

En este documento se presenta el estado del plan de tesis "Valoración de la calidad vocal a través de *deep scattering spectrum* y aprendizaje automático". El objetivo principal del plan de tesis es la clasificación automática de la calidad vocal según la escala GRBAS.

La escala GRBAS es un método de valoración audio-perceptual de la voz. Consiste en la valoración de la fuente glótica a través de 5 parámetros que forman el acrónimo GRBAS, G (*grade* o grado general de disfonía), R (*roughness* o rugosidad), B (*breathiness* o soplosidad), A (*asteny* o astenia) y S (*strain* o tensión). Se valora a través de 4 grados, de 0 a 3, donde el 0 es ausencia de disfonía y el 3 indica disfonía severa. La escala fue mundialmente adoptada y validada en numerosos países. Actualmente se utiliza en la investigación y de manera rutinaria en los consultorios de los profesionales que hacen clínica vocal. Sirve como método simple y al alcance de la mano para valorar la evolución pre-post tratamiento. La debilidad de este método reside en la subjetividad de la valoración de la voz y en la necesidad de que sea realizada por oyentes experimentados en la escucha y la disociación de los parámetros [1, 2].

Para la valoración automática de la calidad vocal, al igual que en cualquier tarea de aprendizaje automático, se debe diseñar un modelo de clasificación que consta de dos etapas principales, extracción de características y clasificación.

En el caso particular del audio, es habitual que la extracción de características se realice sobre una representación espectral, no sobre el audio original (*raw audio*). Las representaciones espectrales más comunes son el espectrograma, el cepstrum, los coeficientes cepstrales de las frecuencias de Mel (MFCC) y los resultados de transformadas *wavelet*. Originalmente, en el plan de tesis se eligió la representación *deep scattering spectrum* (DSS) [3], basada en transformadas *wavelet*. Si bien esta técnica es ventajosa para cierto tipo de clasificación de audio, como el reconocimiento del habla [4], no es la ideal para la valoración de la calidad vocal porque las características que se deben extraer en cada caso son distintas. Mientras que para el reconocimiento del habla es necesario encontrar patrones invariantes al ruido, los cambios de intensidad y los cambios en la frecuencia fundamental de la voz, estos tres factores son fundamentales para el análisis de la calidad. Tal como se verá más adelante, las representaciones espectrales utilizadas en este trabajo son variaciones del espectrograma y del cepstrum.

Materiales y métodos

El estado del arte en técnicas de reconocimiento de patrones sobre imágenes y audio es el aprendizaje profundo (AP) o *deep learning*. Se decidió utilizar este tipo de redes neuronales en la ejecución del plan de tesis.

En los proyectos de AP es habitual reutilizar redes neuronales diseñadas y entrenadas para resolver un problema particular y adaptarlas a un problema similar. En el caso de la clasificación de audio, la mayor parte del esfuerzo de los investigadores se ha enfocado en el reconocimiento del habla (entender qué se dice). No hay existen modelos neuronales profundos de clasificación de la calidad vocal, por lo tanto durante la ejecución de este plan, se diseña desde cero una red neuronal profunda para la clasificación de la calidad de la voz.

Enfoque de diseño de la red neuronal. En un modelo de AP las etapas de extracción de características y clasificación están integradas en una misma red neuronal. En la Figura 1 se muestra un esquema general de clasificación de audio con AP, donde la entrada de la red neuronal es una representación espectral del audio, en este caso un espectrograma calculado con la transformada de Fourier de término reducido (STFT por su sigla en inglés) sobre el audio multiplicado previamente por una ventana (*windowing*).

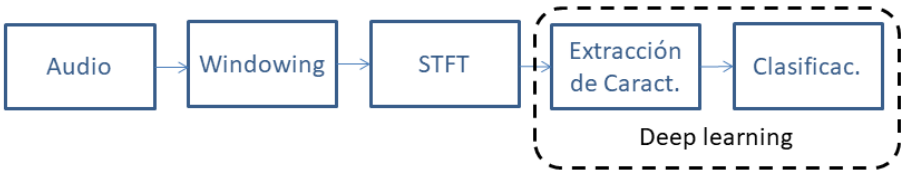

Figura 1. Esquema general de un modelo de aprendizaje automático para clasificación de audio con el espectrograma del audio como entrada.

En el contexto de este proyecto, se busca que la parte de la red que realiza la extracción de características transmita a las capas posteriores información relativa a ciertas medicas acústicas que se saben relacionadas con la calidad vocal, como shimmer, jitter y *harmonics-to-noise ratio* (HNR) [5-7].

Una propuesta novedosa de este trabajo es la integración de las etapas de cálculo de la representación espectral dentro de la red neuronal. En la Figura 2 se muestra el esquema de este enfoque.

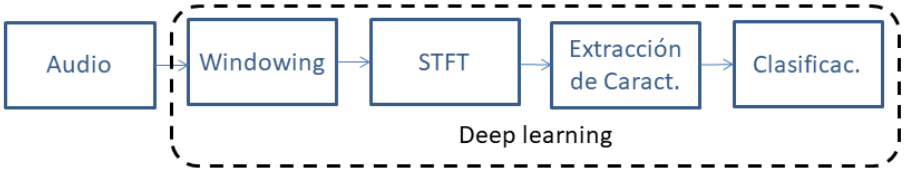

Figura 2. Esquema del modelo de aprendizaje automático propuesto, donde las etapas de cálculo de la representación espectral del audio son parte de la red neuronal.

Con la integración propuesta se espera obtener parámetros óptimos (forma de la ventana y coeficientes de la STFT) para el problema particular que se desea resolver, disminuyendo el error total de clasificación. Este enfoque se podría aplicar en cualquier tarea de clasificación de audio. El resultado es un clasificador extremo a extremo (*end-to-end*), que recibe el audio original y devuelve la categoría a la que pertenece.

Existe un factor crítico para que la integración de estas etapas sea exitosa, el gradiente del error se debe propagar correctamente hacia atrás. La profundidad de la red y las funciones que se utilizan en los cálculos internos afectan la calidad de la información que reciben las primeras capas durante el entrenamiento. Uno de los objetivos principales en esta fase del proyecto es determinar si la información del error se propaga con la calidad suficiente en las etapas de extracción de características y representación espectral.

Implementación. Durante la ejecución del proyecto se ha logrado realizar hasta el momento, con redes neuronales, el cálculo de la STFT y el cepstrum sobre audios reales y el cálculo de shimmer sobre el espectrograma para audios sintetizados. En todos los casos la implementación se llevó a cabo en Python, con la librerías Keras [8] y Theano [9]. El entrenamiento se realizó en una PC con una GPU NVIDIA Titan Xp donada a través del *NVIDIA's GPU Grant Program*. A continuación se explica brevemente la arquitectura de cada una de las redes neuronales desarrolladas.

Cálculo de STFT con redes neuronales. La STFT se calcula como la transformada discreta de Fourier (DFT) por tramos de la señal de entrada. La DFT es una transformación lineal y las redes neuronales realizan una transformación lineal antes de la función de activación, por lo tanto, una red neuronal con función de activación lineal puede calcular la DFT. La extensión a STFT se puede realizar de forma natural con una capa de convolución. En la Figura 3 se muestra un esquema de la red neuronal que calcula la el módulo de la STFT. Publicado en [10].

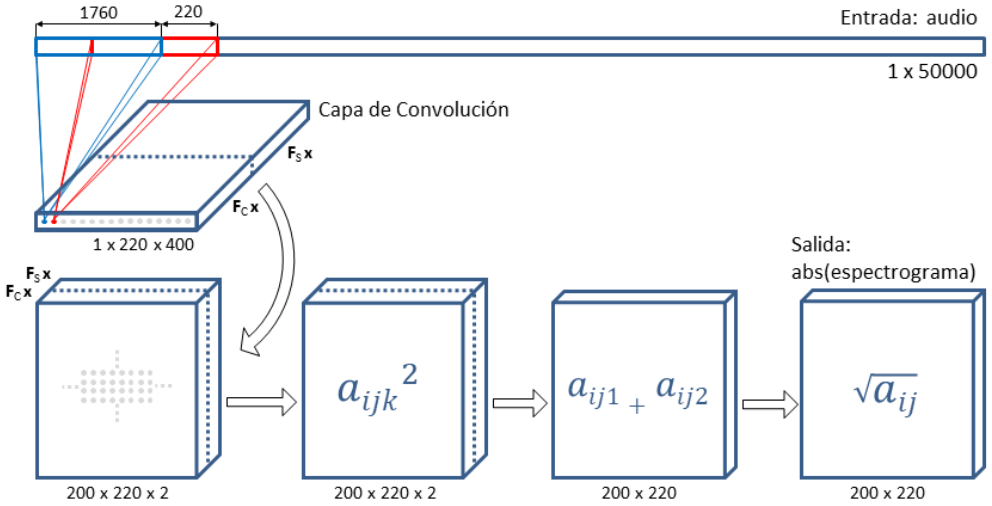

Figura 3. Red neuronal artificial para el cálculo del valor absoluto del espectrograma. La entrada es una secuencia de longitud = 50000. La ventana tiene un ancho de 1760 y una superposición de 220 elementos. Los pesos de la red, $\mathbf{F}_C$ y $\mathbf{F}_S$, son (inicialmente) los coeficientes de la DFT para coseno y seno respectivamente.

Cálculo del cepstrum con redes neuronales. El cepstrum es cuadrado de la DFT del logaritmo del cuadrado del espectro. Su uso en reconocimiento de patrones en audio es muy común. Se utiliza, entre otras cosas, para calcular la frecuencia de vibración (F0) de las cuerdas vocales, para determinar la cantidad de ruido y para separar fácilmente el efecto del filtro según el modelo fuente/filtro de producción de la voz. El cálculo del cepstrum con redes neuronales es una extensión del modelo de la Figura 3. La modificación más significativa es una nueva capa de convolución que calcula la transformada coseno. En la Figura 4 se puede ver el esquema de la red neuronal de cálculo del cepstrum. Publicado en [11].

Cálculo de shimmer con redes neuronales. Shimmer es una medida acústica de las perturbaciones de amplitud de una señal. Este se calcula acumulando la diferencia de amplitud entre ciclos consecutivos de la frecuencia fundamental. En la Figura 5 se muestra la red neuronal desarrollada para calcular shimmer en una señal sintetizada. Esta red tiene una capa convolucional, una de *max pooling* y tres capas densamente conectadas. La entrada al modelo es un espectrograma y, por lo tanto, esta red puede ser conectada con la red de la Figura 3 para calcular shimmer directamente desde el audio original. Publicado en [12].

Datos. Para el entrenamiento y evaluación del modelo de cálculo de shimmer se utilizaron datos sintetizados. El generador de datos crea señales de audio con frecuencia fundamental aleatoria, añade estructura de armónicos, ruido y modela la señal en amplitud con otra señal generada aleatoriamente.

En el caso de los otros dos modelos, se utilizaron datos reales de tres bases de datos de voces:

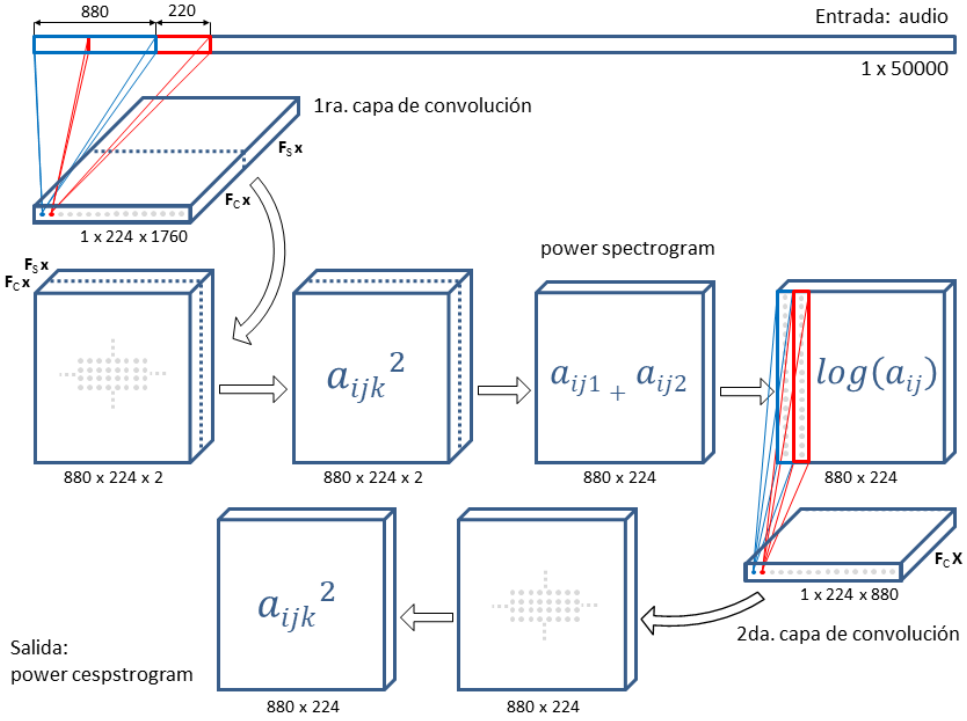

Figura 4. Red neuronal artificial para el cálculo del cepstrum. La entrada es una secuencia de longitud = 50000. La ventana tiene un ancho de 880 y una superposición de 220 elementos. Los pesos de la red, $\mathbf{F}_c$ y $\mathbf{F}_s$, son (inicialmente) los coeficientes de la DFT para coseno y seno respectivamente.

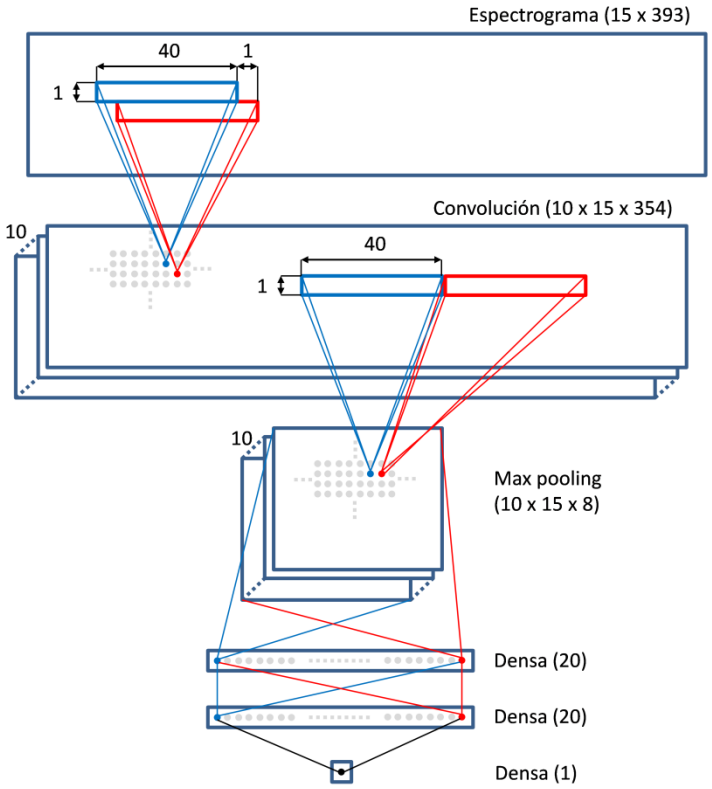

Figura 5. Red neuronal para el cálculo de shimmer de una señal sintetizada.

- Base de datos UTN-UNC. Los audios fueron grabados en colaboración con profesionales de Ciencias Médicas de la Universidad Nacional de Córdoba. Contiene vocales /a/ sostenidas de personas sanas y con patologías vocales.
- *Voice Disorders Database*. Voces grabadas en el Hospital Universitario Príncipe de Asturias (España). Los datos fueron compartidos por la Universidad Politécnica de Madrid. Esta base de datos contiene grabaciones de vocales /a/ sostenidas de personas sanas y con patologías vocales. Además contiene una clasificación en

escala GRBAS de la calidad vocal y otros atributos de interés, como edad, sexo, patología y estimación de F0 [13].

- *Saarbruecken Voice Database*. Audios grabados y compartidos por la *Universität des Saarlandes*. Contiene grabaciones de vocales sostenidas /a/, /i/ y /u/, en tono normal, alto y bajo, sexo, edad y registro de electroglotógrafo [14].

Los audios y datos de las tres bases de datos se integraron en una nueva base de datos. Los archivos de audio se transformaron para unificar las duraciones y se calcularon los espectrogramas, cepstrogramas y algunas variantes de estos últimos para entrenar y evaluar las redes neuronales.

Resultados

Cálculo de STFT con redes neuronales. Si la red neuronal presentada se inicializa con los coeficientes de la DFT, el cálculo es perfecto, ya que los modelos son equivalentes. Para este caso, el error cuadrático medio (MSE) es menor a 10^{-9}.

Incializando los pesos con valores aleatorios y entrenando la red con backpropagation, se obtuvo un MSE de test = 1.41×10^{-6} (9.79×10^{-6}% de la salida esperada media). Los pesos obtenidos durante el entrenamiento forman una base ortogonal, al igual que los coeficientes de la DFT, pero no tienden a los mismos valores. En la Figura 6 se muestra una comparación entre los pesos F_c y F_s de un kernel de convolución obtenidos por entrenamiento y los coeficientes de la DFT (pesos teóricos). Se puede observar que los pesos entrenados (líneas finas) se encuentran desfasados con respecto a los teóricos por una cantidad aleatoria, pero mantienen un desfasaje de 90° entre ellos mismos (F_c y F_s entrenados).

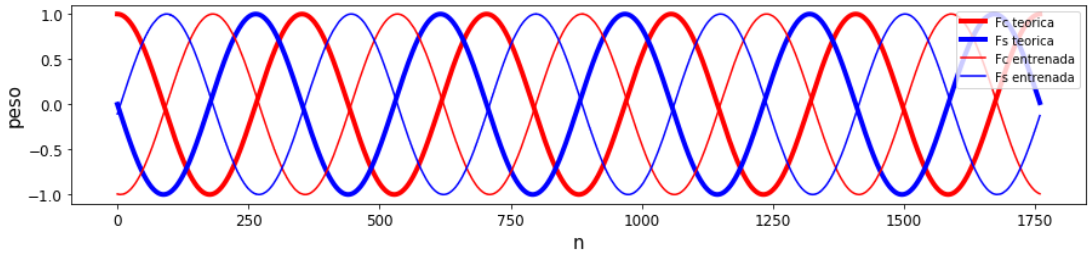

Figura 6. Pesos entrenados vs teóricos del modelo neuronal de cálculo de la STFT.

Cálculo del cepstrum con redes neuronales. Al igual que el modelo anterior, si los pesos son inicializados con los coeficientes de la DFT el resultado es perfecto, pero lo que interesa es la capacidad de entrenamiento/adaptación del modelo. El modelo de la Figura 4 no pudo ser entrenado. La primera capa de convolución no se adapta correctamente debido a que la derivada del logaritmo no permite que el gradiente del error se propague correctamente. Una variante del cepstrum, la *autocovariance*, que no calcula el logaritmo sí pudo ser entrenada. El modelo resultante es similar al de la Figura 4, pero sin la capa de cálculo del logaritmo. Para este caso, inicializando los pesos con los valores teóricos y adicionando ruido, el ratio entre el error absoluto medio y el valor esperado medio es 1.33×10^{-4}. Se eligió este ratio para comparar los resultados porque las salidas del cepstrum y la *autocovariance* no están estandarizadas y no son comparables de forma directa. Para más detalles ver [11].

Cálculo de shimmer con redes neuronales. Con el modelo de la Figura 5 se logró un MSE $= 5.8 \times 10^{-5}$ en la predicción de shimmer. En la Figura 7 se pueden ver las predicciones para 500 valores del set datos de test. Los datos fueron ordenados por el valor de shimmer esperado para facilitar la visualización.

Conclusiones

Los resultados obtenidos para el cálculo de la STFT y shimmer son altamente satisfactorios. El modelo que calcula el cepstrum no se logró entrenar, pero el modelo que calcula la *autocovariance* sí. Como esta última medida mantiene muchas de las características importantes del cepstrum, se concluye, por un lado, que los modelos desarrollados hasta el momento son capaces de calcular la información necesaria para las etapas correspondientes (representación espectral y extracción de características).

Por otro lado, se concluye que los modelos obtenidos son entrenables, es decir, que el gradiente del error se propaga correctamente a través de las capas permitiendo la optimización de los parámetros.

Finalmente, como los modelos calculan correctamente los valores esperados y propagan el error, son compatibles con el enfoque de diseño planteado.

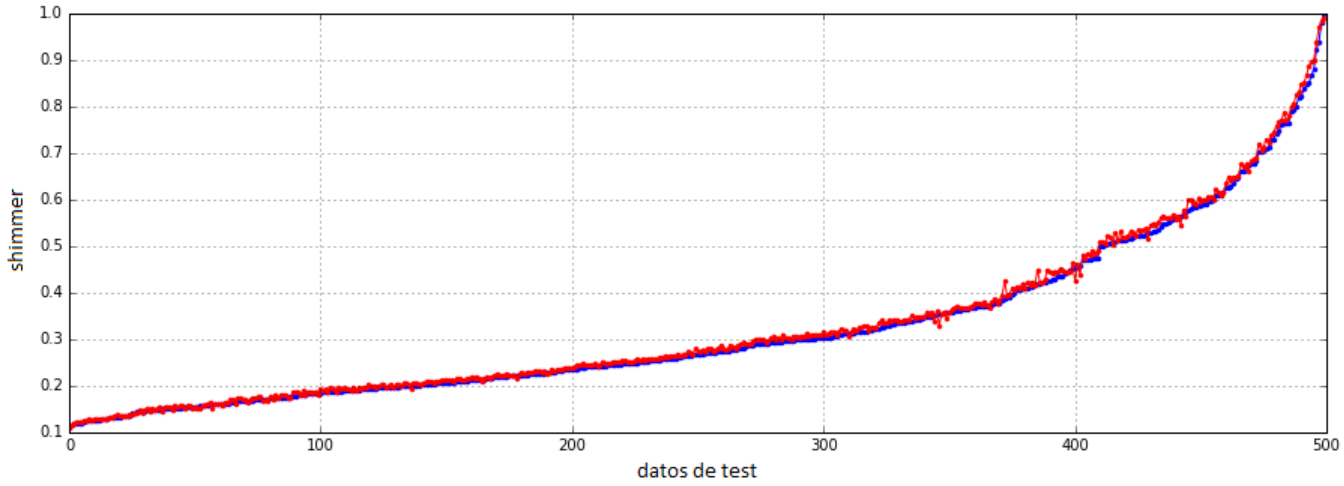

Figura 6. Predicciones de shimmer con la red neuronal propuesta para datos de test. Se muestran los valores esperados de shimmer (azul) ordenados de forma ascendente y las salidas del modelo (rojo) correspondientes.

Referencias

[1] Kreiman, J., Gerratt, B. R. (2010). Perceptual assessment of voice quality: past, present, and future. SIG 3 Perspectives on Voice and Voice Disorders, 20(2), 62-67.

[2] Núñez-Batalla, F., Díaz-Molina, J. P., García-López, I., Moreno-Méndez, A., Costales-Marcos, M., Moreno-Galindo, C., Martínez-Camblor, P. (2012). El espectrograma de banda estrecha como ayuda para el aprendizaje del método GRABS de análisis perceptual de la disfonía. Acta Otorrinolaringológica Española, 63(3), 173-179.

[3] Andén, J., Mallat, S. (2014). Deep scattering spectrum. Signal Processing, IEEE Transactions on, 62(16), 4114-4128.

[4] Sainath, T. N., Peddinti, V., Kingsbury, B., Fousek, P., Ramabhadran, B., & Nahamoo, D. (2014). Deep scattering spectra with deep neural networks for LVCSR tasks. In INTERSPEECH (pp. 900-904).

[5] Nieto, R.G., Marín-Hurtado, J.I., Capacho-Valbuena, L.M., Suarez, A.A., Bolaños, E.A.B. (2014). Pattern recognition of hypernasality in voice of patients with cleft and lip palate. In: 2014 XIX Symposium on Image, Signal Processing and Artificial Vision (STSIVA), pp. 1–5. IEEE.

[6] Holi, M.S., et al. (2015). A hybrid model for neurological disordered voice classification using time and frequency domain features. Artif. Intell. Res. 5(1), 87.

[7] Freitas, S.V., Pestana, P.M., Almeida, V., Ferreira, A. (2015) Integrating voice evaluation: correlation between acoustic and audio-perceptual measures. J. Voice 29(3), 390–e1.

[8] Chollet, F. (2015). Keras . https://github.com/fchollet/keras.

[9] Theano D. Team (2016). Theano: A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688.

[10] García, M. A., Destéfanis, E. A. (2018). Spectogram Prediction with Neural Networks. In XXIV Congreso Argentino de Ciencias de la Computación.

[11] García, M. A., Destéfanis, E. A. (2019). Power Cepstrum Calculation with convolutional Neural Networks. Journal of Computer Science & Technology, vol. 19, no. 2, pp. 132. doi: 10.24215/16666038.19.e13.

[12] García, M. A., Destéfanis, E. A. (2017). Deep neural networks for shimmer approximation in synthesized audio signal. In Argentine Congress of Computer Science (pp. 3-12). Springer, Cham.

[13] Arias-Londoño, J. D., Godino-Llorente, J. I., Markaki, M., Stylianou, Y. (2011). On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices. Logopedics Phoniatrics Vocology, 36(2), 60-69.

[14] W. J. Barry and M. Pützer, “Saarbrücken Voice Database”, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/.