ISBN: 978-950-42-0200-4
DOI: https://doi.org/10.33414/ajea.5.748.2020

Data Augmentation para la Clasificación Automática de la Calidad Vocal

Data Augmentation in Automatic Classification of Voice Quality

Presentación: 06/10/2020

Doctorando:
Mario Alejandro García
Grupo de Inteligencia Artificial (GIA), Facultad Regional Córdoba, Universidad Tecnológica Nacional - Argentina.
mgarcia@frc.utn.edu.ar

Director:
Eduardo Atilio Destéfanis

Resumen

Se presenta el estado del plan de tesis "Valoración de la calidad vocal a través de *deep scattering spectrum* y aprendizaje automático" y se plantean tres transformaciones para incrementar la cantidad de datos de entrenamiento y reducir el sobreajuste. Estas transformaciones realizan un desplazamiento en frecuencia de los datos (audios), una segmentación por tiempo y la inversión del orden temporal (*flipping*). Como resultado, se obtiene un juego de datos 18 veces mayor al original. Se ejecuta un experimento que consta del el entrenamiento de una red neuronal profunda para evaluar el rendimiento con los datos aumentados. Se concluye que las transformaciones propuestas disminuyen el sobreajuste, mejoran el error de clasificación y se pueden utilizar en el ámbito de este plan de tesis, clasificación de la calidad vocal a partir de audios de vocales sostenidas.

Palabras clave: Calidad vocal, Aprendizaje profundo, Data augmentation.

Abstract

The status of the thesis plan "Vocal quality assessment through deep scattering spectrum and machine learning" is presented. Three transformations are proposed in order to increase the amount of training data and reduce overfitting. These transformations perform a frequency shift, time segmentation and flipping. It results in a dataset 18 times larger than the original dataset. An experiment consisting of training a deep neural network is run to evaluate performance with the augmented data. It is concluded that the proposed transformations reduce the overfitting, improve the classification error and it could be useful for the thesis plan scope, classification of vocal quality from sustained vowels.

Keywords: Vocal quality, Deep learning, Data augmentation.

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" y se presenta un desarrollo destinado a aumentar la cantidad de datos de entrenamiento.

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*), R (*roughness*), B (*breathiness*), A (*asteny*) y S (*strain*). 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 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 (Kreiman & Gerratt, 2010; Núñez-Batalla *et al*, 2012).

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*). Originalmente, en el plan de tesis se eligió la representación *deep scattering spectrum* (DSS) (Andén & Mallat, 2014), basada en transformadas *wavelet*. Si bien esta representación es adecuada para cierto tipo de clasificación de audio, como el reconocimiento del habla (Sainath *et al*, 2014), 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. Las representaciones espectrales utilizadas en este trabajo son variaciones del espectrograma y del cepstrum. Para más detalles ver (García & Destéfanis, 2019).

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.

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) sobre el audio multiplicado previamente por una ventana (*windowing*).

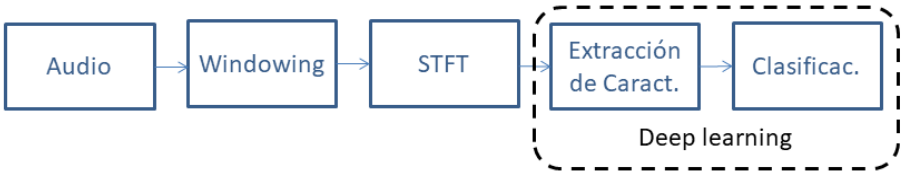

Figura 1. Esquema general de un modelo de aprendizaje automático para clasificación de audio con el espectrograma del audio como entrada.

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 medidas acústicas que se saben relacionadas con la calidad vocal, como shimmer, jitter y *harmonics-to-noise ratio* (HNR) (Nieto, Marín-Hurtado, Capacho-Valbuena, Suarez & Bolaños, 2014; Uma Rani & Holi, 2016; Freitas, Pestana, Almeida & Ferreira, 2015).

Una propuesta novedosa desarrollada durante la ejecución del plan de tesis 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.

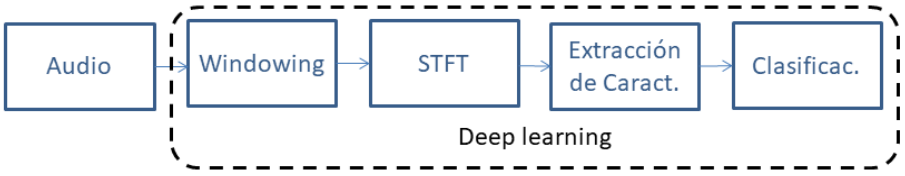

Figura 2. Esquema del modelo de aprendizaje automático propuesto, donde las etapas de cálculo de la representación espectral del audio son parte de la red neuronal.

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. Durante la ejecución del proyecto se ha probado que la STFT, una variante del cepstrum y shimmer pueden ser calculados con redes neuronales y que es posible ajustar los pesos a través del método de retropopagación (García & Destéfanis, 2017, 2018, 2019b). También se logró adaptar los pesos de una red que realiza el cálculo de la etapa de *windowing* (en proceso de publicación).

Data Augmentation.

Las redes neuronales profundas tienen una gran cantidad de parámetros internos que se ajustan durante el aprendizaje. Esta cantidad de parámetros hace que los modelos de AP sean propensos al sobreajuste. Para evitar el sobreajuste es necesaria una gran cantidad de datos de entrenamiento. Los datos utilizados para clasificar la calidad vocal son difíciles de conseguir porque implican la grabación de personas con patologías de la voz y la calificación de expertos, por lo tanto, las bases de datos disponibles no tienen el tamaño necesario para entrenar redes neuronales profundas. Una técnica frecuentemente utilizada para enfrentar este tipo de problemas es *data augmentation*, la cual consiste en la creación de nuevos datos mediante transformaciones sobre los datos originales. Estas transformaciones deben mantener la información sobre los invariantes, propios del dominio del problema, que se utilizan para clasificar los datos (Shorten & Khoshgoftaar, 2019).

En el resto de este artículo se presentan los detalles de la técnica utilizada para aumentar la cantidad de datos.

Desarrollo

Se cuenta con una base de datos de 440 muestras formadas por audios de personas sanas y con patologías vocales pronunciando una vocal /a/ sostenida. Las 440 voces han sido clasificadas en la escala GRBAS por profesionales. Los audios fueron muestreados con una frecuencia de 25 KHz y tienen duración entre 1.5 y 4 segundos. Para aumentar la cantidad de datos se realizaron tres transformaciones:

Desplazamiento en frecuencia. Para cada archivo de audio se realizan dos desplazamientos en frecuencia, incrementándola un 20% en el primero y decrementándola en la misma proporción para el segundo. Cada desplazamiento genera un nuevo archivo de audio, multiplicando por 3 la cantidad de datos. El proceso implica un remuestreado (*resample*) a 20.000 muestras por segundo en el primer caso y 31250 en el segundo caso. Después del remuestreado los datos se guardan con frecuencia de muestreo de 25 KHz en los metadatos del archivo de audio. El resultado de estos cambios se muestra en la Figura 3, donde se puede ver el efecto en los dominios del tiempo y la frecuencia de la transformación sobre un segmento de audio. Desde arriba hacia abajo, se puede ver que la señal en el dominio del tiempo se acorta, es decir, aumenta la frecuencia, mientras que en el dominio de la frecuencia se hace evidente en la ubicación de los picos y la separación entre los armónicos.

Segmentación por tiempo. De cada archivo de audio se extraen tres segmentos de 1 segundo de duración. La ubicación de los segmentos izquierdo (I), central (C) y derecho (D) depende de la longitud L del audio original. Si la L > 2 segundos, se toma el segmento central S de dos segundos, en otro caso, el segmento S será el audio original completo. El segmento I es el primer segundo de S, C es el segundo central de S y D el último segundo de S. En la Figura 4 se muestra gráficamente la segmentación.

Flipping. En esta transformación se genera un nuevo audio a partir de cada uno de los audios existentes invirtiendo el orden la señal en el tiempo. En la Figura 5 se muestra la transformación de un segmento pequeño de audio. Esta técnica de *data augmentation* no es aplicable al reconocimiento del habla porque cambia el orden de los fonemas, pero sí resulta útil para la clasificación de la calidad vocal si se utiliza, como en este caso, audios de vocales sostenidas. El espectro de frecuencias de la señal original es el mismo que el de la señal invertida, pero su espectrograma resulta en un espejado horizontal, tal como se puede ver en la Figura 6.

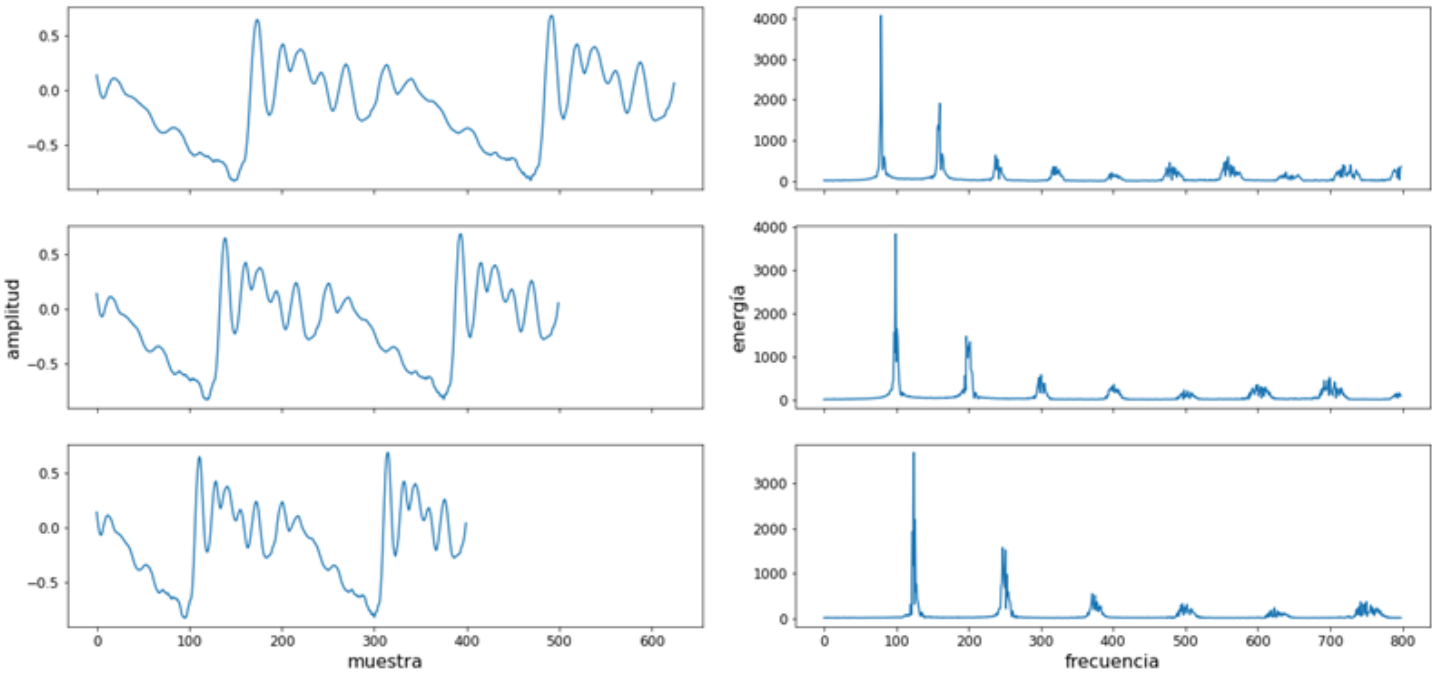

Figura 3. Desplazamiento en frecuencia. En la fila central se ubica la señal original, en la fila superior la señal con un decremento del 20% en la frecuencia y en la fila inferior la misma señal con un 20% de incremento. En la columna de la izquierda se muestra el efecto en el dominio del tiempo (25000 muestras por segundo) y a la derecha el espectro de frecuencias.

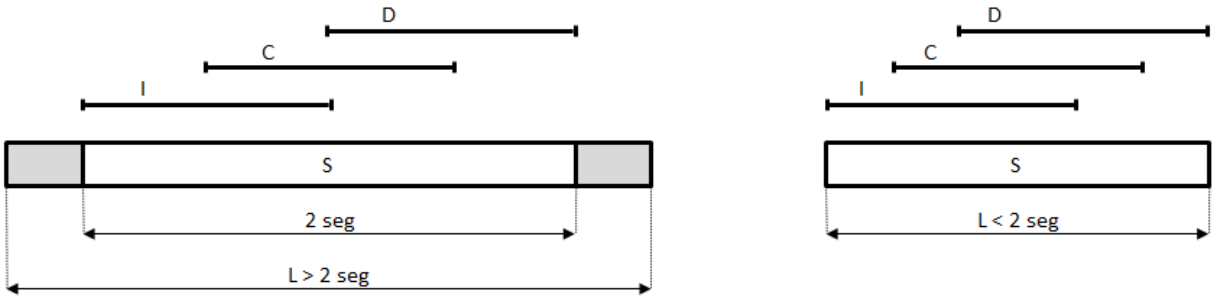

Figura 4. Extracción de tres segmentos de audio para un archivo original de longitud L > 2 segundos (izquierda) y para un archivo original de L < 2 segundos (derecha).

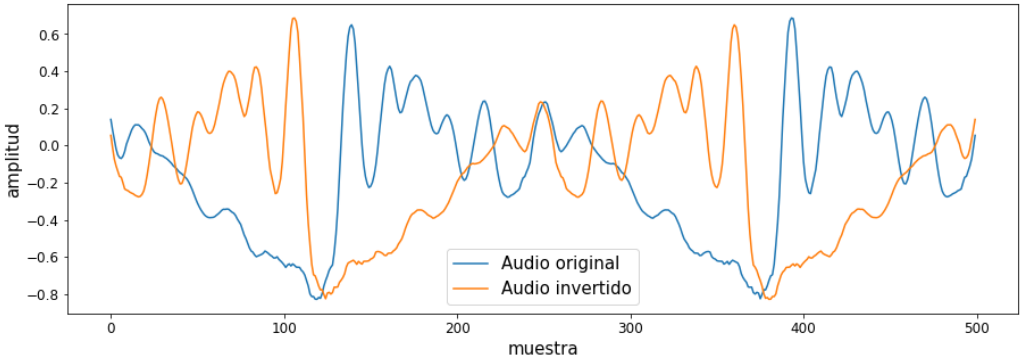

Figura 5. Segmentos de audio original y transformado en la etapa de *flipping*.

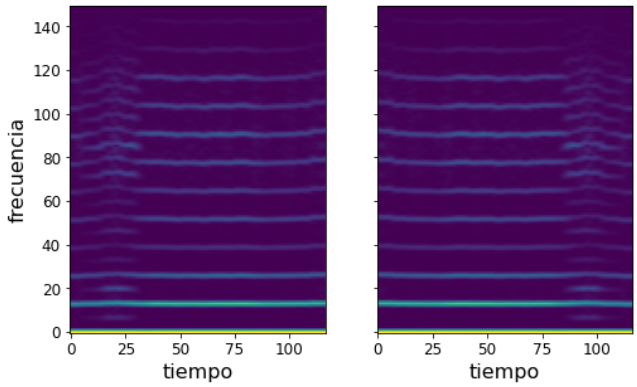

Figura 6. Espectrogramas de audio original (izquierda) y del audio transformado en la etapa de *flipping* (derecha).

Resultados

Se comparó el rendimiento de un modelo de clasificación de la calidad vocal utilizando los datos originales y los datos aumentados. El modelo consiste en una red neuronal, desarrollada en Keras, con dos capas de convolución y dos capas densas que recibe como entrada el cepstrum calculado para cada uno de las columnas del espectrograma. El juego de datos original tiene 440 audios, mientras que el aumentado tiene 7920. En cada caso, se utilizó el 70% de los datos para entrenar el modelo y el 30% para validación. El modelo fue inicializado y entrenado 50 veces con semillas de números aleatorios distintas. Para cada semilla se realizó el entrenamiento de los dos juegos de datos. Comparando el mejor resultado para cada juego de datos en cada entrenamiento, el entrenamiento con el juego de datos aumentado mejora el error con una media del 5,12%. En la figura 7 se muestra la evolución del error durante uno de los entrenamientos. Se puede ver con facilidad que la brecha entre el error de entrenamiento y el error de validación es mucho menor para los datos con *data augmentation* que para los datos originales.

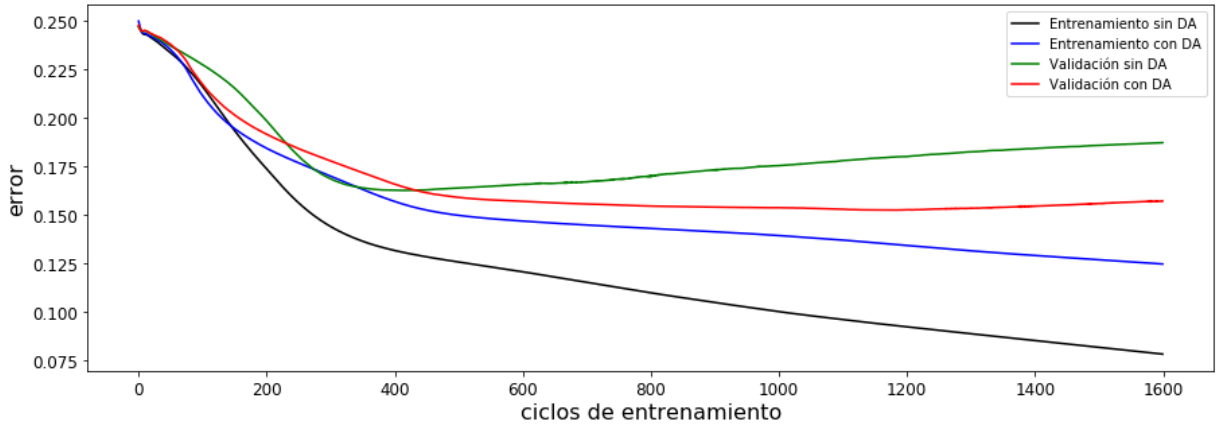

Figura 7. Evolución del error durante el entrenamiento de una red neuronal para la clasificación de la calidad vocal con datos sin *data augmentation* (DA) y datos con DA.

Conclusiones

Los resultados del experimento muestran que el uso de los datos generados con las tres transformaciones propuestas disminuye la diferencia entre el error de entrenamiento y el error de validación y también se logra una mejora absoluta del error de validación. Se concluye que las transformaciones propuestas para aumentar la cantidad de datos ayudan a reducir el sobreajuste del modelo y es posible utilizarlas para entrenar redes neuronales profundas orientadas a la clasificación de la calidad vocal a partir de audios de vocales sostenidas.

Referencias

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

Freitas, S. V., Pestana, P. M., Almeida, V., & Ferreira, A. (2015). Integrating voice evaluation: correlation between acoustic and audio-perceptual measures. *Journal of Voice*, 29(3), 390-e1.

García, M. A., & Destéfanis, E. A. (2017, October). Deep neural networks for shimmer approximation in synthesized audio signal. In *Argentine congress of computer science* (pp. 3-12). Springer, Cham.

García, M. A., & Destéfanis, E. A. (2018). Spectrogram Prediction with Neural Networks. In *XXIV Congreso Argentino de Ciencias de la Computación* (La Plata, 2018).

García, M. A., & Destéfanis, E. A. (2019a). Clasificación Automática de la Calidad Vocal. *AJEA*, (4).

García, M. A., & Destéfanis, E. A. (2019b). Power Cepstrum Calculation with Convolutional Neural Networks. *Journal of Computer Science & Technology*, 19.

Kreiman, J., & Gerratt, B. R. (2010). Perceptual assessment of voice quality: Past, present, and future. *Perspectives on Voice and Voice Disorders*, 20(2), 62-67.

Nieto, R. G., Marín-Hurtado, J. I., Capacho-Valbuena, L. M., Suarez, A. A., & Bolaños, E. A. B. (2014, September). Pattern recognition of hypernasality in voice of patients with Cleft and Lip Palate. In *2014 XIX Symposium on Image, Signal Processing and Artificial Vision* (pp. 1-5). IEEE.

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.

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 *Fifteenth Annual Conference of the International Speech Communication Association*.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60.

Uma Rani, K., & Holi, M. S. (2016). A hybrid model for neurological disordered voice classification using time and frequency domain features. *Artif. Intell. Research*, 5(1), 87-94.