Deep Neural Network para Análisis Acústico

García Mario Alejandro ✉[1], Rosset Ana Lorena[2], Eduardo Destéfanis[1], Cerruti Santiago[1], Moyano Miguel[1]

[1]Universidad Tecnológica Nacional Facultad Regional Córdoba (UTN FRC)
[2]Universidad Nacional de Córdoba (UNC)
mgarcia@frc.utn.edu.ar

RESUMEN

La valoración de la calidad vocal mediante el análisis audio-perceptual es parte de la rutina clínica de evaluación de pacientes con trastornos de la voz. La debilidad de este método reside en la subjetividad y en la necesidad de que sea realizada por oyentes experimentados. Este proyecto tiene como objetivo la realización de una clasificación automática de la calidad vocal, valuada en la escala GRBAS, a través de características extraídas del análisis acústico de la señal y técnicas de aprendizaje automático. Particularmente, en este trabajo se muestran los resultados del cálculo de *shimmer* con un modelo de *deep learning.*

Palabras clave: ***machine learning, deep learning, acoustic analysis***

CONTEXTO

Este trabajo de investigación se desarrolla en el marco del proyecto "Análisis acústico de la voz con técnicas de aprendizaje automático" (UTN3947) de la Universidad Tecnológica Nacional, Facultad Regional Córdoba y cuenta con la colaboración del Departamento de Investigación Científica, Extensión y Capacitación "Raquel Maurette", Escuela de Fonoaudiología, Facultad de Ciencias Médicas, Universidad Nacional de Córdoba.

1. INTRODUCCIÓN

Se intenta reconocer, de forma automática, características del análisis acústico de la voz que permitan clasificar muestras de audio. El estudio se enfoca en la medición de la calidad vocal según la escala GRBAS. La clasificación se realiza aplicando principalmente modelos de *deep learning*, un subgrupo de técnicas del campo de aprendizaje automático (*machine learning*). Las grabaciones de la voz, la clasificación de los ejemplos y la validación de los resultados se realizan por especialistas en análisis de la voz de la Escuela de Fonoaudiología de la Universidad Nacional de Córdoba. El análisis acústico se lleva a cabo en conjunto (especialistas vocales e integrantes de UTN) y el modelado y desarrollo de los clasificadores por los integrantes de UTN.

GRBAS: La escala GRBAS es un método de valoración perceptivo-auditivo de la voz. Surge de la necesidad de estandarizar la valoración subjetiva y de interrelacionar los aspectos auditivos y fisiológicos de la producción vocal. Está basada en estudios del año 1966 de la *Japan Society of Logopedics and Phoniatrics* [1] y posteriormente divulgada y descripta por Minoru Hirano en el año 1981 [2]. Consiste en la valoración de la

fuente glótica a través de 5 parámetros que forman el acrónimo GRBAS:

G: (*Grade*) Grado general de disfonía.

R: (*Roughness*) Rugosidad, irregularidad de la onda glótica.

B: (*Breathiness*) Soplosidad, sensación de escape de aire en la voz.

A: (*Asteny*) Astenia, pérdida de potencia.

S: (*Strain*) Tensión, sensación de hiperfunción vocal.

Puede valorarse de dos maneras: a través de 4 grados, desde el 0 al 3 o mediante un valor en un rango continuo de 0 a 100. En ambas el 0 es ausencia de disfonía y el 3 o 100 implican disfonía severa. La escala fue mundialmente adoptada y validada en numerosos países [3-6]. Actualmente se utiliza en la investigación y de manera rutinaria en los consultorios de los profesionales que hacen clínica vocal. Sirve como metodología 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 [7,8].

Análisis acústico: Existen otras formas de analizar la voz de manera más objetiva a través del análisis acústico. Éste consiste en la digitalización de la señal vocal y su análisis mediante gráficos como el Espectrograma, el espectro FFT (*Fast Fourier Transform*) o LPC (*Linear Predictive Coding*) y medidas numéricas de perturbación de la señal, como *Jitter, Shimmer* y HNR (*Harmonics to Noise Ratio*).

Para lograr una integración de la valoración subjetiva (GRBAS u otras escalas) con el análisis acústico, se han realizado numerosos trabajos de correlación [9,10], algunos relacionados a la voz normal y otros a diferentes patologías. Por ejemplo, el trabajo de Nuñez Batalla, F. et al [11] es un referente y establece una relación entre el parámetro de Astenia del GRBAS y el Espectrograma de banda angosta.

Aprendizaje automático: El aprendizaje automático o *machine learning* es un campo de las ciencias de la computación que abarca el estudio y la construcción de algoritmos capaces de aprender y hacer predicciones. Estas predicciones se pueden tomar como una clasificación de los datos de entrada a partir del reconocimiento de patrones existentes en los mismos

Estado del arte: La aplicación de técnicas de *deep learning* es el estado del arte en el análisis automático de audio, con la detección de los fonemas pronunciados y la identificación de la persona que habla como objetivos principales [12-18], pero también utilizadas en detección de emociones, edad, género, etc.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de investigación que se presenta en este trabajo tiene como objetivo predecir con técnicas de aprendizaje automático la medida acústica *shimmer*, asociada a la calidad vocal.

Shimmer:

Shimmer es una medida acústica relativa a las perturbaciones de amplitud de una señal. Las variaciones de este tipo en la voz humana son perceptibles al oído y permiten caracterizar ciertas propiedades, tanto de la voz, como de las personas que la emiten [19]. El valor de *shimmer* está asociado a la calidad vocal, estado de ánimo, edad y género de las personas. Existen numerosos trabajos de

investigación que utilizan, entre otras, la medida *shimmer* con objetivos que van desde la detección de patologías [19-21] hasta la mejora de interfaces humano máquina a través de la estimación de la intensionalidad de una frase hablada [22]. Con respecto a las voces sintetizadas, en [23] se determina que cierto nivel se *shimmer* aumenta el grado de naturalidad.

Hay variantes en el cálculo de *shimmer* . La versión elegida para este trabajo es la de Klingholz y Martin [23], también conocida como *Relative Shimmer*.

Metodología:

Los datos de entrada para el modelo de predicción son espectrogramas calculados sobre archivos de audio sintetizado.

Se generan datos de audio sin armónicos. Al igual que en [19] la modulación en amplitud de la voz se aproxima con una onda senoidal. La expresión para generar cada señal de audio $y(t)$ es:

$$y(t) = \frac{1}{1+k}\sin(\alpha + 2t\pi f_0)\,(1 + k\sin(\beta + 2t\pi f_{mod}))$$

donde t es tiempo $[seg]$, f_0 es la frecuencia de vibración glótica $[Hz]$, f_{mod} es la frecuencia de modulación $[Hz]$, k es la constante de sensibilidad en amplitud del modulador, α y β son constantes para manejar la fase de la señal a modular y de la señal moduladora respectivamente.

Para la generación de los datos de entrenamiento y test, se toman valores aleatorios con distribución uniforme. f_0 toma valores en el intervalo $[200; 1000]Hz$, f_{mod} en $[5; 10]Hz$, k en $[0; 0{,}4]$, α y β en $[0; 2\pi]$.

El espectrograma se calcula sobre 2 segundos de audio generado con 44100 muestras/seg. Para el cálculo se utiliza una ventana tipo *Tukey*(0.25) de ancho = 256, lo que determina una estructura de forma 129 x 393 (frecuencia/tiempo) que contiene la densidad espectral de la señal.

El modelo de predicción es una *deep neural network*. Esta red se diseña en un proceso que genera modelos de complejidad ascendente, aproximando en primer lugar el valor de f_0, k y f_{mod} individualmente a partir de los datos espectrales, luego *shimmer* en función de f_0, k y f_{mod}, y por último, *shimmer* en función de los datos espectrales.

Se utilizan juegos de datos de 3000 muestras para entrenamiento, 500 para test y 500 para validación.

3. RESULTADOS OBTENIDOS

Se obtuvo un modelo neuronal con una capa de convolución y tres capas densamente conectadas (figura 1) capaz de aproximar *shimmer* con datos espectrales como entrada. El error cuadrático medio obtenido sobre los datos de test fue $MSE = 5.8 \times 10^{-5}$ [25].

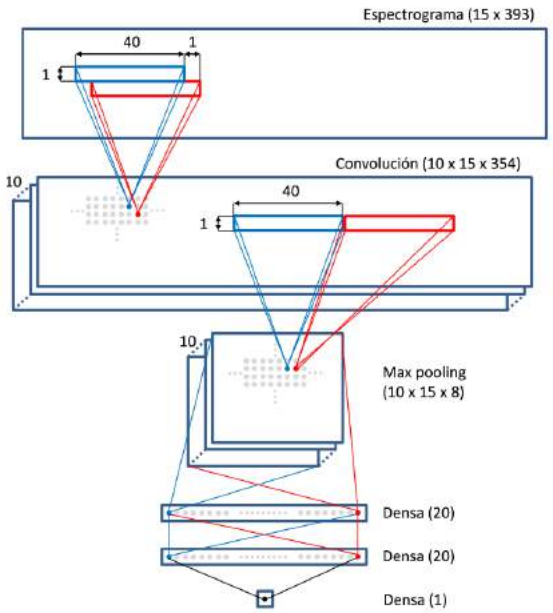

Figura 1. Modelo de *deep learning* para predicción de *shimmer*.

 RedUNCI - UNNE - ISBN 978-987-3619-27-4

Se logró comprobar que para datos simples de audio modulados en amplitud por una onda senoidal, con parámetros de frecuencia fundamental, frecuencia moduladora y sensibilidad de modulación variables, es posible obtener un modelo neuronal capaz de aproximar el valor de *shimmer*. La importancia de este resultado radica en que, bajo las condiciones planteadas en el trabajo, se puede afirmar que un modelo de *deep learning* que respete la estructura del modelo presentado en sus primeras capas es capaz de utilizar el valor de *shimmer*, internamente calculado, para realizar clasificaciones de otro tipo, como la calidad vocal.

4. FORMACIÓN DE RECURSOS HUMANOS

El equipo del proyecto está formado por un docente/investigador de la UTN FRC, dos docentes/investigadores de la UNC y cuatro alumnos de la carrera de grado de la UTN FRC.

Además de formación de los alumnos participantes, el conocimiento generado por el proyecto se incorporará a las cátedras de los docentes de la UTN y UNC.

5. REFERENCIAS

[1] Isshiki, N., Yanagihara, N., & Morimoto, M. (1966). *Approach to the objective diagnosis of hoarseness*. Folia Phoniatrica et Logopaedica, 18(6), 393-400.

[2] Hirano, M. (1981). *Clinical examination of voice* (Vol. 5). Springer.

[3] Yun, Y. S., Lee, E. K., Baek, C. H., & Son, Y. I. (2005). *The correlation of GRBAS scales and laryngeal stroboscopic findings for the assessment of voice therapy outcome in the patients with vocal nodules*. Korean Journal of Otolaryngology-Head and Neck Surgery, 48(12), 1501-1505.

[4] Hui, H., Weijia, K., & Shusheng, G. (2007). *The Validation of Acoustic Analysis and Subjective Judgment Scales of Several Voice Disorders* [J]. Journal of Audiology and Speech Pathology, 3, 010.

[5] Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). *Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders*. Journal of Voice, 21(5), 576-590.

[6] Jesus, L. M., Barney, A., Couto, P. S., Vilarinho, H., & Correia, A. (2009, December). *Voice quality evaluation using cape-v and GRBAS in european Portuguese*. In MAVEBA (pp. 61-64).

[7] 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.

[8] Núñez-Batalla et al (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.

[9] 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.

[10] ELISEI, N. G. (2013). Percepción auditiva de voces patológicas. In XIV Reunión Nacional y III Encuentro Internacional de la Asociación Argentina de Ciencias del Comportamiento.

[11] Nuñez Batalla, F., Corte Santos, P., Señaris Gonzalez, B., Rodriguez Prado, N., Suárez Nieto, C. (2004) Evaluación espectral de la hipofunción vocal. Acta Otorrinolaringol. Esp. 55:327-333.

[12] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Kingsbury, B. (2012): Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, vol. 29.6, 82-97. IEEE.

[13] Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., Saltzman, E., Tiede, M. (2017) Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition. Speech Communication, vol. 89. pp 103-112.

[14] Collobert, R., Puhrsch, C., Synnaeve, G. (2016) Wav2letter: an end-to-end convnet-based speech recognition system.arXiv preprint arXiv:1609.03193.

[15] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Chen, J. (2016) Deep speech 2: End-to-end speech recognition in english and mandarin. International Conference on Machine Learning. pp. 173-182.

[16] Palaz, D., Collobert, R. (2015) Analysis of cnn-based speech recognition system using raw speech as input (No. EPFL-REPORT-210039). Idiap.

[17] Sainath, T. N., Kingsbury, B., Mohamed, A. R., Ramabhadran, B. (2013) Learning filter banks within a deep neural network framework. IEEE Workshop onASRU. pp 297-302. IEEE.

[18] Farrús, M. (2007) Jitter and shimmer measurements for speaker recognition. 8th Annual Conference of ISCA. pp. 778-781. (2007)

[19] Jafari, M., Till, J. A., Law-Till, C. B. (1993) Interactive effects of local smoothing window size and fundamental frequency on shimmer calculation. Journal of Voice, vol. 7.3. pp. 235-241.

[20] Tsanas, A., Little, M. A., Fox, C., Ramig, L. O. (2014) Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease.Neural Systems and Rehabilitation Engineering, IEEE Transactions, vol. 22.1. pp 181-190.

[21] Gómez-Coello, A. et al. (2017) Voice Alterations in Patients With Spinocerebellar Ataxia Type 7 (SCA7): Clinical Genetic Correlations. Journal of Voice,vol. 31.1. pp. 123-e1.

[22] Kotti, M., Patern, F. (2012) Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema.International journal of speech technology,vol. 15.2. pp. 131-150.

[23] Yamasaki, R. et al (2017) Perturbation Measurements on the Degree of Naturalness of Synthesized Vowels.Journal of Voice, vol 31.3, 389-e1.

[24] Klingholz, F., Martin, F. (1985) Quantitative spectral evaluation of shimmer and jitter. J Speech Hear Res, vol 28.2. pp 169-174.

[25] García M.A., Destéfanis E.A. (2018) Deep Neural Networks for Shimmer Approximation in Synthesized Audio Signal. In: De Giusti A. (eds) Computer Science – CACIC 2017. CACIC 2017. Communications in Computer and Information Science, vol 790. Springer, Cham