Red neuronal multiescala para clasificación de la calidad vocal

García Mario Alejandro ✉[1], Rosset Ana Lorena[2], Eduardo Destéfanis[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, mediante la aplicación de técnicas de aprendizaje profundo sobre voces grabadas. Particularmente, en este trabajo se muestran los resultados del diseño de una red neuronal multiescala para la clasificación de la calidad vocal.

Palabras clave: ***red neuronal multiescala***, ***calidad vocal, aprendizaje profundo***

CONTEXTO

Este trabajo de investigación se desarrolla en el marco del proyecto "*Deep learning* para clasificación de señales vocales" (UTN5274) 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 modelos de aprendizaje profundo (*deep learning)*, un subgrupo de técnicas del campo de aprendizaje automático (*machine learning*).

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].

Estado del arte: La aplicación de técnicas de aprendizaje profundo 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 [9-15], pero también utilizadas en detección de emociones, edad y género entre otros. Durante la ejecución de este proyecto se aplican técnicas de aprendizaje a la clasificación de la calidad de la voz.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de investigación que se presenta en este trabajo se enfoca en el desarrollo de un clasificador neuronal multiescala para la clasificación de la calidad vocal.

La teoría del espacio escalar propone que el ojo humano es capaz de reconocer patrones en distintos niveles de suavizado de la imagen original y que los patrones reconocidos en diferentes niveles mejoran la clasificación general. Esta idea aplica con frecuencia en el reconocimiento de imágenes. Existen trabajos de reconocimiento de patrones en audio que utilizan la misma técnica [16-19]. En estos se muestra que representando el audio en distintas escalas se puede mejorar la capacidad de clasificación.

El suavizado se logra convolucionando la señal con una ventana gaussiana. El efecto obtenido es la pérdida de detalles. Cuanto más grande sea la ventana, mayor será el efecto.

La incorporación de la clasificación multiescala en el proyecto se lleva a cabo mediante la adaptación de un modelo de red neuronal profunda (M_0) desarrollado en etapas anteriores.

El modelo M_0 realiza la clasificación de la calidad vocal partiendo desde el audio crudo (*raw* audio). El audio se divide en segmentos y se multiplica cada segmento por una ventana de pesos adaptables mediante una capa STHadamard [20]. Para cada segmento de *windowed* audio se calcula el power cepstrum con otras capas de la red neuronal, tal como se explica en [21]. Las salidas anteriores son las entradas de la última parte de la red, donde se realiza la extracción de características con capas de convolución y la clasificación con dos capas de neuronas densamente conectadas.

Los distintos niveles de suavizado se introducen entre la representación cepstral del audio y la extracción de características (Figura 1). En el modelo presentado en este trabajo se utilizan cuatro niveles o escalas. El suavizado se logra con una capa de convolución de tamaño n y desplazamiento 1, donde los pesos de cada kernel de convolución forman una campana gaussiana y no se modifican durante el entrenamiento. La salida de las cuatro capas de convolución se concatenan formando un vector de tres dimensiones.

 RedUNCI - UNdeC. ISBN: 978-987-24611-3-3

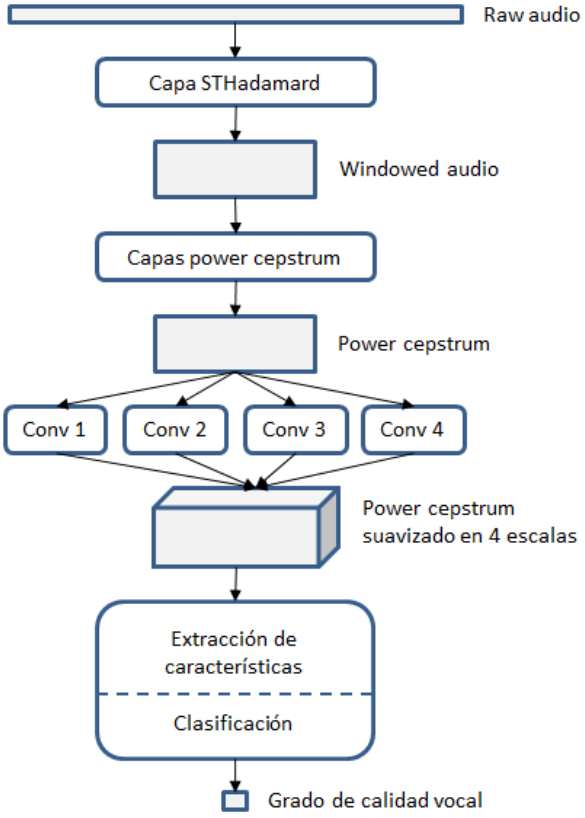

Figura 1. Modelo M_0 modificado con escalado múltiple de cuatro niveles.

Experimento:

Se utilizaron audios de la Voice Disorders Database (VDD) [22], grabados por la Universidad Politécnica de Madrid en colaboración con el Hospital Universitario Príncipe de Asturias. Estas grabaciones se realizaron sobre personas sanas y personas con patologías vocales pronunciando una vocal /a/ sostenida durante aproximadamente dos segundos. Para el experimento se extrajeron segmentos de 1 segundo y se aplicaron las técnicas de *data augmentation* explicadas en [23].

Se realizó una clasificación binaria del grado general de disfonía (G). Debido a la distribución de los valores de G, se crearon las categorías 0 para G = 0 y 1 para G = 1, 2 y 3.

Los tamaños de los kernels de convolución para suavizado fueron $n = 3, 11, 21\ y\ 31$. En la Figura 2 se muestra la salida de las capas de suavizado.

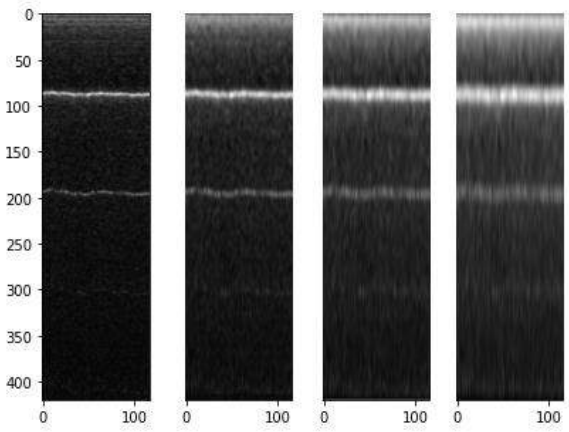

Figura 2. Salidas de las capas de suavizado para $n = 3, 11, 21, 31$ de izquierda a derecha.

3. RESULTADOS OBTENIDOS

A continuación se exponen los resultados de 50 entrenamientos del modelo propuesto y 50 entrenamientos del modelo M_0. Los pesos se inicializaron con valores aleatorios entre -10^{-6} y 10^{-6} uniformemente distribuidos. Se utlizó el método de optimización Adam [24] con los parámetros provistos por los autores y las actualizaciones de los pesos se realizaron en batchs de tamaño 5244 (la totalidad de datos de entrenamiento).

Los cálculos se realizaron sobre una GPU NVIDIA Titan Xp, donada a través del GPU Grant Program de NVIDIA.

El *accuracy* medio de validación alcanzado para el modelo M_0 fue 0.7698, mientras que para el modelo M_0 con suavizado multiescala se obtuvo 0.8069, lo que significa una mejora media del 4.8%.

Estos resultados indican que el modelo propuesto es capaz de mejorar la clasificación de la calidad vocal y que se puede integrar en una red neuronal profunda.

 RedUNCI - UNdeC. ISBN: 978-987-24611-3-3

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] 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.

[10] 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.

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

[12] 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.

[13] Palaz, D., Collobert, R. (2015) Analysis of cnn-based speech recognition system

using raw speech as input (No. EPFL-REPORT-210039). Idiap.

[14] 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.

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

[16] Chi, T., Shamma, S. A. (2006). Spectrum restoration from multiscale auditory phase singularities by generalized projections. IEEE transactions on audio, speech, and language processing, 14(4), 1179-1192.

[17] Zhu, Z., Engel, J. H., Hannun, A. (2016). Learning Multiscale Features Directly from Waveforms. Interspeech 2016, 1305-1309.

[18] von Platen, P., Zhang, C., Woodland, P. C. (2019). Multi-Span Acoustic Modelling Using Raw Waveform Signals. Interspeech 2019, 1393-1397.

[19] Mesgarani, N., Slaney, M., & Shamma, S. A. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. IEEE Transactions on Audio, Speech, and Language Processing, 14(3), 920-930.

[20] García, M. A., Destéfanis, E. A., Rosset, A. L. (2020, September). Trainable Windowing Coefficients in DNN for Raw Audio Classification. In Conference on Cloud Computing, Big Data & Emerging Topics (pp. 153-166). Springer, Cham.

[21] García, M. A., Destéfanis, E. A. (2019). Power cepstrum calculation with convolutional neural networks. Journal of Computer Science & Technology, 19.

[22] Arias-Londoño, J. D., Godino-Llorente, J. I., Markaki, M., Stylianou, Y.: On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices. Logopedics Phoniatrics Vocology 36.2. 2011: 60-69

[23] García, M. A., Destéfanis, E. A. (2020). Data Augmentation para la Clasificación Automática de la Calidad Vocal. AJEA, (5).

[24] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)