BATS: Bridging Acoustic Transparency in Speech

Diego Quezada
Departamento de Informática
Universidad Técnica Federico Santa María
Valparaíso, Chile
diego.quezadac@sansano.usm.cl

Felipe Cisternas
Departamento de Informática
Universidad Técnica Federico Santa María
Valparaíso, Chile
felipe.cisternasal@sansano.usm.cl

***Resumen*—El reconocimiento de voz se basa en representaciones de señales acústicas, como espectrogramas y MFCCs. Sin embargo, los modelos actuales son en gran medida opacos en cuanto a cómo toman decisiones en este proceso. La naturaleza física de los datos de entrada en el reconocimiento de voz agrega una capa adicional de complejidad, lo que plantea el desafío de mejorar la transparencia y la comprensión de estos modelos para garantizar un reconocimiento de voz más preciso y confiable.**

***Index Terms*—ASR, XAI, CNN, RNN, Transformers**

I. Introducción

El reconocimiento de voz, o más conocido en inglés como *speech recognition*, es la tarea de asignar una secuencia de palabras a señales acústicas que contienen lenguaje hablado. Implica reconocer las palabras pronunciadas en una grabación de audio y transcribirlas a un formato escrito. El objetivo es transcribir con precisión el discurso en tiempo real o a partir de audio grabado, teniendo en cuenta factores como el acento, la velocidad del habla y el ruido de fondo. Cuando la transcripción se realiza en tiempo real se habla de reconocimiento automático de voz o *Automatic Speech Recognition* (ASR) en inglés. Considerando $\mathbf{X} = (x^{(1)}, x^{(2)}, \ldots, x^{(T)})$ como una secuencia de audio de largo T e $y = (y_1, y_2, \ldots, y_N)$ como una secuencia de palabras de largo N podemos definir la tarea de reconocimiento de voz de manera precisa mediante el siguiente problema de optimización:

$$f^*(\mathbf{X}) = \arg\max_{\mathbf{y}} P^*(\mathbf{y}|\mathbf{X} = X) \tag{1}$$

Donde P^* es la verdadera distribución de probabilidad condicional que relaciona las entradas $\mathbf{X}$ con las salidas $\mathbf{y}$ [1].

La representación utilizada para $\mathbf{X}$ es de vital importancia para el desempeño de un modelo de reconocimiento de voz. La representación más simple es mediante una serie temporal univariada que modela la amplitud de la **señal de audio** en el tiempo. Al dividir la señal de audio en pequeñas ventanas de tiempo y calculando el espectro de frecuencia para cada ventana se obtiene un **espectrograma**: una representación visual de la señal de audio en el tiempo y en el dominio de la frecuencia. A partir del espectrograma se pueden extraer características de audio tales como los **coeficientes cepstrales de Mel** (MFCCs) que son ampliamente utilizados en la literatura para el reconocimiento de voz.

Como es de esperarse, los modelos del estado del arte para el reconocimiento de voz son cajas negras, es decir, no es posible interpretar el proceso de decisión que realizan para asignar una secuencia de palabras a una señal de audio. Por lo tanto, es necesario utilizar técnicas de *eXplainable Artificial Intelligence* (XAI) para poder entender las decisiones de un modelo de reconocimiento de voz.

En la presente investigación se trabajará con conjuntos de datos estándar para el reconocimiento de voz tales como *TIMIT* [2] y *CommonVoice* [3]. Ambos distribuídos de manera gratuita y con licencia abierta.

II. Problema

Para lograr una mejor cohesión entre las dos ideas presentadas, es importante establecer una conexión clara entre los modelos específicos mencionados y la necesidad de explicabilidad en el contexto del reconocimiento de voz. Aquí está mi sugerencia:

En el campo del reconocimiento de voz, la elección del modelo y la arquitectura adecuados es crucial para abordar eficazmente los desafíos únicos que presenta esta área. Existen múltiples opciones disponibles, incluyendo modelos destacados como Whisper, basado en la arquitectura de Transformers de OpenAI, que ha demostrado ser una herramienta poderosa en la tarea de *speech recognition* [4]. Sin embargo, estos modelos avanzados a menudo carecen de explicabilidad, un aspecto que puede obstaculizar su robustez y la confianza del usuario en ellos.

La explicabilidad no solo facilita una mayor comprensión del proceso de decisión del modelo, sino que también es vital para mejorar su robustez, especialmente cuando estos modelos se integran como componentes centrales en un software. En este contexto, la explicabilidad se convierte en una herramienta indispensable para ganar la confianza de los usuarios en el producto final. A pesar de la importancia crítica de la explicabilidad, el reconocimiento de voz presenta desafíos adicionales debido a la naturaleza física de los datos de entrada, lo que complica la tarea de proporcionar explicaciones de alto nivel basadas en estos datos. Por lo tanto, mientras se busca avanzar en la precisión y eficiencia de estos modelos, es igualmente imperativo trabajar hacia soluciones que ofrezcan una mayor explicabilidad, equilibrando así la balanza entre el rendimiento y la comprensibilidad del modelo.

La pregunta de investigación que se aborda en la presente investigación es: ¿Cómo lograr una integración efectiva de explicabilidad en reconocimiento de voz?

Esta investigación tiene como objetivo mejorar la comprensión de los modelos de reconocimiento de voz para que los usuarios, especialmente aquellos con discapacidades auditivas, puedan confiar en su funcionamiento.

III. Métodos

Se utilizarán los siguientes métodos de ML:

1. CNNs para aprender patrones en los espectrogramas de audio.
2. RNNs para modelar la dependencia temporal de los espectrogramas de audio.
3. Transformers y mecanismos de atención [5] para aprender las características más importantes de los espectrogramas de audio.

Respecto a los métodos de XAI, se utilizarán los siguientes:

1. LIME [6] para identificar los atributos más importantes.
2. Grad-CAM [7] para visualizar las regiones más importantes de los espectrogramas.
3. Métodos de prominencia para visualizar cuánto contribuye una característica de entrada al resultado del modelo.
4. Análisis de la matriz de atención para entender a qué características el modelo les asigna mayor importancia.

Referencias

[1] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[2] J. Garofolo, L. Lamel, W. Fisher et al., «TIMIT Acoustic-phonetic Continuous Speech Corpus,» *Linguistic Data Consortium*, nov. de 1992.

[3] R. Ardila, M. Branson, K. Davis et al., «Common Voice: A Massively-Multilingual Speech Corpus,» *CoRR*, vol. abs/1912.06670, 2019. arXiv: 1912.06670. dirección: http://arxiv.org/abs/1912.06670.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey e I. Sutskever, *Robust Speech Recognition via Large-Scale Weak Supervision*, 2022. arXiv: 2212.04356 [eess.AS].

[5] A. Vaswani, N. Shazeer, N. Parmar et al., *Attention Is All You Need*, 2023. arXiv: 1706.03762 [cs.CL].

[6] M. T. Ribeiro, S. Singh y C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, 2016. arXiv: 1602.04938 [cs.LG].

[7] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh y D. Batra, «Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization,» *CoRR*, vol. abs/1610.02391, 2016. arXiv: 1610.02391. dirección: http://arxiv.org/abs/1610.02391.