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

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 el campo del reconocimiento de voz, la elección del modelo y la arquitectura 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* [1]. Sin embargo, estos modelos complejos por lo general 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 toma 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.

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.

I-A. Trabajo relacionado

En [2] se aborda la tarea reconocimiento de voz utilizando utilizando el conjunto de datos CommonVoice [3] y tres sistemas de reconocimiento de voz: Google API, Sphinx y Deepspeech. Los autores plantean las explicaciones como un subconjunto de fotogramas de audio que son causas mínimas y suficientes de la transcripción. Para esto, adaptan las técnicas de clasificación de imagen SFL [4] y explicaciones composicionales [5] para luego compararlas con las ofrecidas por LIME [6].

En [7] se aborda la tarea de **reconocimiento de fonemas** utilizando el conjunto de datos TIMIT [8] y el método LIME junto a dos variantes propuestas: LIME-WS y LIME-TS. El conjunto de datos TIMIT segmenta los distintos fonemas en los datos de entrada. Los autores utilizan como representación interpretable un vector que indica la presencia o ausencia de cada segmento de audio. El objetivo es identificar los segmentos más importantes para la predicción del siguiente fonema. Inicialmente, para la generación del vecindario se aplican máscaras aleatorias a los segmentos de TIMIT. Luego, considerando que al analizar un fonema es probable que solo los segmentos cercanos incidan en su identificación, LIME-WS solo aplica máscaras aleatorias a segmentos cercanos al fonema analizado y mantiene los lejanos constantes. Finalmente, en LIME-TS se consideran segmentos creados por un intervalo de tiempo en vez de los definidos por TIMIT y se aplica la misma idea de localidad que en LIME-WS.

En [9] se aborda la tarea de etiquetado de música. Los autores plantean la necesidad de dar **explicaciones audibles** y proponen audioLIME como una extensión de LIME que utiliza explicaciones basadas en separación de fuentes.

II. Método propuesto

II-A. Marco teórico

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 $\mathbf{y}$ [10].

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.

LIME [6] es un modelo agnóstico que explica de manera local las predicciones de un modelo **caja negra** f mediante la aproximación de f con un modelo **explicable** g en una vecindad de la instancia de interés $\boldsymbol{x}$. LIME utiliza una **representación interpretable** $\boldsymbol{x'} \in \{0,1\}^{d'}$ de la instancia $\boldsymbol{x} \in \mathbb{R}^d$. De esta forma, el dominio de la explicación g es $\{0,1\}^{d'}$ y, por lo tanto, g actúa sobre la presencia o ausencia de *componentes interpretables*. Para generar la vecindad se utiliza una función $\pi_x(z)$ que mide la distancia entre dos puntos x y z en el espacio de atributos interpretable. Adicionalmente, dado que la idea es aproximar f mediante un modelo interpretable, se considera una penalización Ω que mide la complejidad de g. Finalmente, y considerando una función de pérdida $\mathcal{L}(f, g, \pi_x)$ que mide la imprecisión del modelo g al aproximar f en la vecindad de $\boldsymbol{x}$ de acuerdo a la función de distancia π, la explicación producida por LIME que asegura **interpretabilidad** y **fidelidad local** se obtiene según la siguiente fórmula:

$$\xi(\boldsymbol{x}) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g). \tag{2}$$

Donde G es el conjunto de todos los modelos interpretables considerados, como por ejemplo modelos lineales o árboles de decisión.

II-B. Conjunto de datos

En la presente investigación se utilizará el conjunto de datos *CommonVoice* [3]. En particular, se utilizarán las grabaciones en inglés asociadas a la versión 11 de CommonVoice disponible en HuggingFace Datasets.

Los conjunto de entrenamiento, validación y prueba consisten en 948736, 16354 y 16354 grabaciones de audio respectivamente. Cada grabación de audio tiene asociada una transcripción. Adicionalmente, cada grabación de audio fue grabada con un *sampling rate* de 48 kHz.

II-C. Descripción propuesta

La propuesta busca evaluar la explicabilidad de modelos estado del arte en la tarea de reconocimiento de voz mediante métodos **post-hoc**.

III. Resultados preliminares

Referencias

[1] 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].

[2] X. Wu, P. Bell y A. Rajan, «Explanations for Automatic Speech Recognition,» en *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, págs. 1-5. DOI: 10.1109/ICASSP49357.2023.10094635.

[3] R. Ardila, M. Branson, K. Davis y col., «Common Voice: A Massively-Multilingual Speech Corpus,» en *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, págs. 4211-4215.

[4] Y. Sun, H. Chockler, X. Huang y D. Kroening, *Explaining Image Classifiers using Statistical Fault Localization*, 2020. arXiv: 1908.02374 [cs.LG].

[5] H. Chockler, D. Kroening e Y. Sun, *Compositional Explanations for Image Classifiers*, mar. de 2021.

[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] X. Wu, P. Bell y A. Rajan, *Can We Trust Explainable AI Methods on ASR? An Evaluation on Phoneme Recognition*, 2023. arXiv: 2305.18011 [cs.CL].

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

[9] V. Haunschmid, E. Manilow y G. Widmer, «audioLIME: Listenable Explanations Using Source Separation,» *CoRR*, vol. abs/2008.00582, 2020. arXiv: 2008.00582. dirección: https://arxiv.org/abs/2008.00582.

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