CLASIFICACIÓN DE SONIDOS EN EL EXTERIOR DE VEHÍCULOS MEDIANTE INTELIGENCIA ARTIFICIAL

Lucas Banchero Martínez[1*]
José Javier López Monfort[1]

[1]Universidad Politécnica de Valencia - ITEAM - GTAC, Valencia, España

RESUMEN

En este estudio, se utiliza la capacidad de las redes neuronales para comprender, caracterizar y clasificar distintos patrones de sonidos existentes en el entorno de un vehículo en funcionamiento. Para este cometido, se llevan a cabo mediciones acústicas en diferentes puntos del exterior del vehículo cuando éste está en marcha. Una vez obtenidas las medidas, se extraen los descriptores necesarios para llevar a cabo dicho entrenamiento del modelo de inteligencia artificial.

Como descriptor principal, se hace uso de los Espectrogramas de Mel. Este descriptor es elegido porque permite mantener la información en tiempo y en frecuencia de los audios (y por tanto, la información secuencial), y por otro lado, gracias a la naturaleza de los filtros de Mel, se puede obtener más resolución de la señal en baja frecuencia, donde se centra la información relevante.

Una vez entrenado el sistema con los descriptores elegidos, se procede a usar el modelo en la actuación del algoritmo de clasificación. Dicho algoritmo permite que, a raíz de segmentar una señal de audio en tramos, y posteriormente extraer su espectrograma, se pueda obtener la información acústica que predomina en dicho segmento, clasificarla con una clase concreta y poder actuar en consecuencia.

ABSTRACT

In this study, the capacity of neural networks is harnessed to comprehend, characterize, and classify various sound patterns present in the environment of a functioning vehicle. To accomplish this task, acoustic measurements are conducted at different points on the exterior of the vehicle while it is in operation. Once the measurements are obtained, the necessary descriptors are extracted for training the artificial intelligence model.

The primary descriptor utilized in this context is Mel Spectrograms. This choice is made because it enables the preservation of temporal and frequency information in the audio data (and, consequently, sequential information). Additionally, due to the nature of Mel filters, it provides higher signal resolution in the lower frequency range, where relevant information is concentrated.

Upon training the system with the selected descriptors, the model is employed in the classification algorithm's performance. This algorithm, when applied to audio signal segments and subsequent extraction of their spectrograms, allows for the identification of predominant acoustic characteristics in the segment, classification into specific categories, and subsequent actions accordingly.

Palabras Clave— ast, srp-phat, clasificación de audio, doa, inteligencia artificial

1. INTRODUCCIÓN

La seguridad vial y el desarrollo de sistemas de conducción autónoma han sido temas centrales en la industria automotriz y la investigación tecnológica en los últimos años. Uno de los desafíos clave en este contexto es la detección temprana y precisa de eventos sonoros relevantes en el entorno del vehículo. La identificación de estos eventos es fundamental para mejorar la seguridad vial y permitir una conducción autónoma más segura y eficiente.

Sin embargo, en ciertas circunstancias, la disponibilidad de información visual puede ser limitada o insuficiente. Factores como la niebla, la lluvia intensa o simplemente la distancia con el objetivo pueden dificultar la comprensión completa de nuestro entorno a través de la vista. Es precisamente en estas situaciones cuando el sonido se convierte en una fuente valiosa de información, capaz de

[*] ***Autor de contacto***: *lbanmar@upv.edu.es*

proporcionar detalles cruciales que escapan a la percepción visual en esos momentos.

En este contexto, el presente trabajo se enfoca en el desarrollo de un sistema que hace uso de la inteligencia artificial (IA) y técnicas avanzadas de procesamiento de señales acústicas para detectar y localizar eventos sonoros relacionados con la conducción automovilística, en concreto bocinas y sirenas. El objetivo principal de este sistema es proporcionar información crítica en situaciones en las que la información visual puede resultar incompleta o insuficiente, mejorando así la seguridad vial y facilitando una conducción autónoma más segura y eficaz. A lo largo de este trabajo, se describirán con detalle los descriptores utilizados, la arquitectura de la IA, el algoritmo de estimación de dirección de llegada (DOA) y los resultados de los experimentos prácticos que respaldan la eficacia de esta aproximación.

2. DESCRIPTORES DE AUDIO

En este estudio, los espectrogramas de Mel han sido seleccionados como descriptores principales en base a su tratamiento logarítmico en frecuencia a lo largo de todo el espectro audible. Esta característica es resultado de la naturaleza de los filtros utilizados en la transformación de Mel y se revela como un atributo valioso para el análisis de entornos automovilísticos, concretamente, en sonidos de bocinas y sirenas, siendo los objetos de este estudio.

Como se puede apreciar en el Figura 1 y la Figura 2, las principales componentes frecuenciales de los sonidos a analizar tienen su energía concentrada entre 500Hz y 4000Hz. Dado que el ruido de rodadura automovilística se centra en 500 Hz–2.5 kHz, como se demuestra en estudio anteriores [1][2], la mayor resolución en estas bandas resulta un elemento crítico a la hora de discernir entre ruidos y señales de interés.

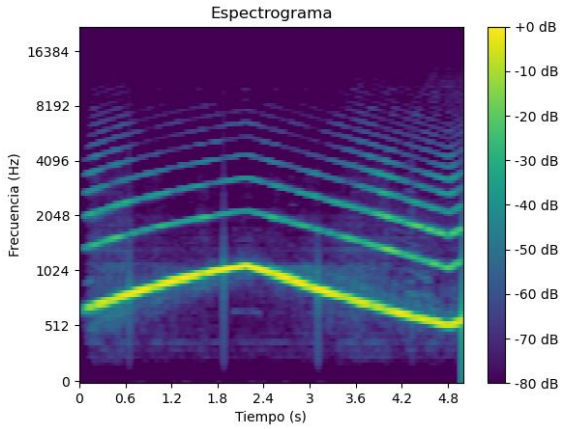

Figura 1. Espectrograma de Mel de Sirena.

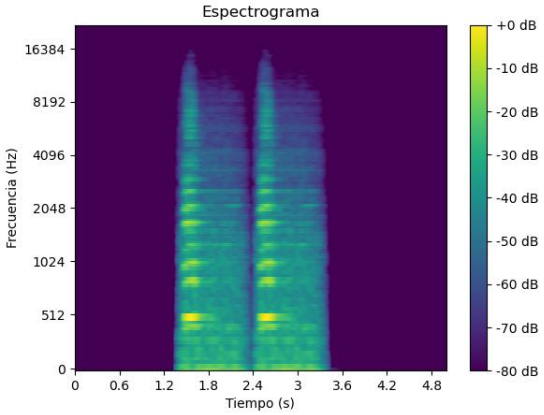

Figura 2. Espectrograma de Mel de Bocina.

3. RECONOCIMIENTO DE SONIDO

En este estudio, se ha desarrollado un sistema de reconocimiento de sonido que funciona como un disparador en tiempo real. Su función principal es detectar la presencia de sirenas y bocinas en el entorno acústico del vehículo, lo que activa el algoritmo de localización basado en TDOA (*Time Difference of Arrival*) que se describe en la sección siguiente.

La arquitectura utilizada para el reconocimiento de sonido se basa en el Audio Spectrogram Transformer (AST) [3], que es una adaptación de las redes *Transformers* [4] diseñada especialmente para analizar espectrogramas. Esta elección de arquitectura se respalda por su eficacia en la clasificación de eventos sonoros, así como su capacidad para aprender y reconocer patrones en representaciones espectrales de señales de audio.

Para entrenar el modelo de reconocimiento de sonido, se utilizó la base de datos Audioset de Google [5], que contiene una amplia variedad de grabaciones de audio etiquetadas con 527 etiquetas diferentes. Una vez que el sistema aprendió a diferenciar entre estas numerosas etiquetas, se modificó la estructura del *transformer* para que solo detectara tres clases: Sirena, Bocina o Nada. De esta manera, cualquier sonido que no perteneciera a las dos primeras categorías se consideraba ruido y no activaba el algoritmo de localización basado en TDOA.

Este segundo entrenamiento se llevó a cabo mediante *transfer learning* [6], desconectando las últimas capas de neuronas del modelo, para introducir unas nuevas que aprendieran solo las etiquetas comentadas anteriormente. Para este cometido, se utilizó la base de datos de UrbanSound8K, aplicando una validación cruzada, usando el método de 10 *k-folds*, tal como se indica en [7]. Esta base de

datos permitió refinar aún más el modelo, asegurando su capacidad para distinguir de manera efectiva entre sirenas, bocinas y otros sonidos ambientales.

El proceso de entrenamiento y validación del sistema se basa en la recopilación de muestras de audio de las bases de datos mencionadas. A partir de estas muestras, se calculan los espectrogramas de Mel como descriptores, que posteriormente se introducen en la red neuronal AST. El proceso de entrenamiento permite al sistema aprender y ajustar sus pesos y parámetros internos para reconocer de manera efectiva los sonidos de interés en el entorno automovilístico, cumpliendo así su función de disparador para la localización de eventos sonoros. Los resultados de la validación cruzada del sistema se pueden observar en Tabla 1.

Tabla 1. Porcentaje de acierto de detección en la validación cruzada de la base de datos UrbanSound8K

K-fold	Precisión (%)
1	100
2	87.55
3	91.98
4	98.15
5	97.03
6	97.52
7	97.56
8	99.05
9	93.92
10	98.61
Promedio	96.14

Dado la fiabilidad del modelo, y al funcionar como disparador, se determina un valor umbral de predicción mayor del 90% para que el sistema active el algoritmo de localizador de fuente.

4. LOCALIZADOR DE SONIDO

Para abordar la localización de fuentes en entornos automovilísticos ruidosos, se implementó un sistema de localización de sonido basado en el algoritmo SRP-PHAT (*Steered Response Power with Phase Transform*). Este algoritmo se eligió debido a su destacada capacidad para mantener una alta relación señal-ruido en presencia de interferencias acústicas, lo que resulta fundamental para la localización precisa de fuentes sonoras en condiciones ruidosas.

4.1. Principio de Funcionamiento de SRP-PHAT

El algoritmo SRP-PHAT calcula una función de distribución espacial, que representa la probabilidad de que la fuente de sonido esté ubicada en cada dirección angular posible. Esta función se calcula mediante una transformación de fase llamada "*Phase Transform*" (PHAT) que ajusta las señales de los micrófonos para tener en cuenta las diferencias de tiempo. La función resultante se llama "*Steered Response Power*" (SRP), que muestra la potencia relativa de la señal en diferentes direcciones angulares. [8][9]

Esta última característica es esencial para determinar la procedencia de las fuentes de sonido en el entorno y es especialmente valiosa en condiciones adversas, como entornos ruidosos, donde SRP-PHAT ha demostrado una notable robustez y precisión en la detección de sonidos [10].

4.2. Localización de fuentes

El algoritmo de localización se activa automáticamente una vez que el sistema de reconocimiento de sonido, alimentado por el modelo AST, detecta la presencia de una sirena o una bocina. En ese momento, el algoritmo calcula la dirección de llegada del sonido utilizando SRP-PHAT y utiliza esta información para tomar decisiones apropiadas, como alertar al conductor o permitir que un vehículo de conducción autónoma ajuste su trayectoria en consecuencia.

4.3. Resolución Angular

Dado que este sistema se adapta tanto a conductores manuales como a vehículos de conducción autónoma, la resolución angular se estableció en intervalos de 30 grados, cubriendo un rango de 360 grados. Esto significa que el sistema determina la dirección de llegada del sonido dentro de uno de los doce ángulos predefinidos (por ejemplo, 30 grados, 60 grados, etc.), lo que proporciona información suficiente para tomar decisiones de dirección relevantes sin necesidad de una resolución angular muy alta. La resolución se puede aumentar sin problemas, pero se ha creído suficiente esta resolución para entornos automovilísticos.

5. DESCRIPCIÓN DEL EXPERIMENTO

Para evaluar exhaustivamente la eficacia y el rendimiento de nuestro sistema de localización de sonido en entornos automovilísticos, se realizaron pruebas utilizando una configuración de simulación meticulosamente diseñada. El objetivo principal de esta fase de experimentación fue recrear condiciones realistas y desafiantes que se asemejaran al entorno de un vehículo en movimiento.

La experimentación se desarrolló en dos fases fundamentales. En una primera etapa, con el propósito de adquirir datos de un entorno automovilístico genuino, se dispuso un conjunto de cuatro micrófonos en el exterior de un vehículo de prueba, tal como se visualiza en la Figura 3. Estos micrófonos fueron estratégicamente ubicados en el vehículo para lograr una captura precisa del entorno sonoro mientras

éste estaba en movimiento. Esta disposición nos permitió recolectar datos reales de ruido de la rodadura del entorno. Posteriormente, en la segunda fase de nuestro experimento, que se explica a continuación, pudimos evaluar el rendimiento de nuestro sistema en situaciones que replicaban las condiciones de conducción real en carreteras.

Figura 3. Micrófono exterior instalado en vehículo

5.1. Montaje del experimento

En el marco de nuestras pruebas de simulación, implementamos una configuración específica en una sala especialmente diseñada para emular condiciones realistas de un entorno vehicular en movimiento. La sala está aislada acústicamente para evitar ruido del exterior, a la vez que está acondicionada acústicamente para minimizar las reflexiones desde las paredes y el techo, creando así un ambiente de prueba, que, aunque no sea anecoico, está muy controlado.

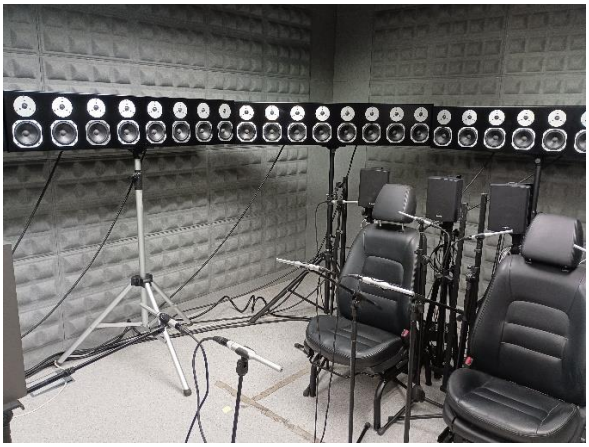

Figura 4. Sala de entorno de simulación automovilística

Para simular los sonidos provenientes de diferentes direcciones, se aprovechó la instalación de un sistema de Wave-Field Synthesis que consta de un array de 96 altavoces dispuestos en una forma octogonal estirada, de 6.75 metros de largo y 3.45 metros de ancho. Esta configuración nos permitió evaluar el rendimiento de nuestro sistema en un entorno controlado y ajustado a las dimensiones de la sala.

En cuanto a la disposición de los micrófonos, cuyo objetivo es detectar y localizar las fuentes de interés mencionadas, colocamos cuatro micrófonos dentro de la sala, estratégicamente ubicados en el interior del octágono formado por los altavoces, como se ilustra en la Figura 5. Esta disposición se ajustó a las dimensiones de la sala y se escaló para emular un entorno real de carretera.

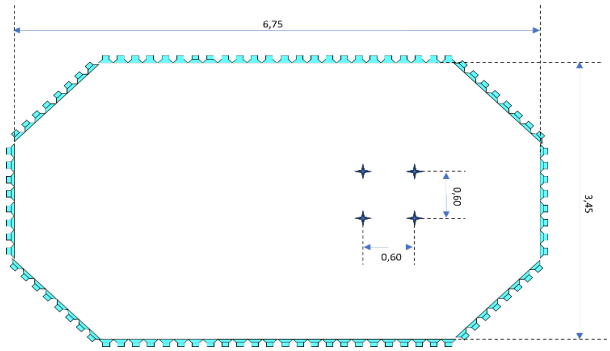

Figura 5. Distribución de altavoces y micrófonos.

5.2 Procedimiento de evaluación

Para llevar a cabo las mediciones, emitimos sonidos de sirenas o bocinas a través de uno de los altavoces durante 20 segundos, mientras que los otros altavoces reproducían el sonido de tráfico real captado del entorno de rodadura de los vehículos que se había grabado previamente. La configuración se repetía en incrementos de 30 grados, cubriendo así todo el diagrama polar. Cabe destacar que, para lograr una mayor fidelidad con el entorno real, dividimos el octógono en cuatro secciones, de modo que cada sección emitía el sonido captado por uno de los cuatro diferentes micrófonos que se instalaron en el vehículo de prueba. Esto aseguró que el sonido de la rodadura tuviera la suficiente especialidad y realismo para el experimento.

5.3 Activación del Sistema Disparador de IA

En el escenario de simulación, implementamos el sistema de inteligencia artificial (IA). Este sistema está conectado a uno de los cuatro micrófonos del conjunto de simulación de micrófonos y opera en tiempo real, analizando de forma continua tramas de audio de 2 segundos. Si el micrófono detecta la presencia de una clase relevante, activa un disparador y clasifica el sonido en una de las clases definidas.

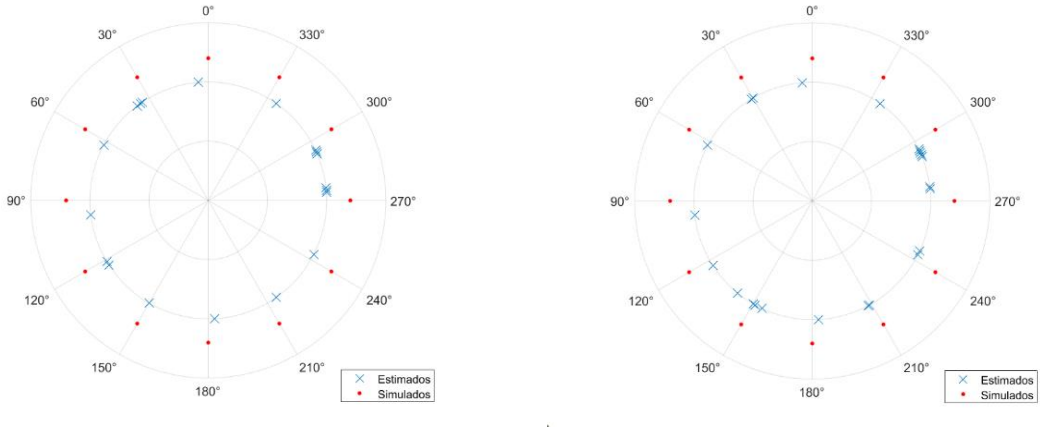

Figura 6. Comparación entre direcciones simuladas y estimadas mediante SRP-PHAT para audio de Sirena (Izquierda) y para audio de Bocina (Derecha).

A continuación, para el tramo de sonido donde está el evento, se analiza la señal de los cuatros micrófonos mediante el algoritmo SRP-PHAT comentado en la sección 4, proporcionando la dirección estimada de llegada.

Posteriormente, se muestra en pantalla tanto la clase detectada como la dirección de llegada. Esta información se mantiene visible en pantalla mientras se continúa detectando la clase relevante. En el momento en que no se detecta ninguna clase relevante, también se informa de este hecho.

6. ANÁLISIS DE LOS RESULTADOS

Para evaluar la eficacia y el rendimiento de nuestro sistema de detección y localización de sonido en entornos automovilísticos, se realizaron pruebas utilizando una configuración de simulación cuidadosamente diseñada. En primer lugar, sometimos nuestro sistema de detección basado en inteligencia artificial (IA) a una serie de pruebas para evaluar su capacidad de identificar con precisión sirenas y bocinas en condiciones simuladas. Se utilizó la base de datos ESC-50[11], que consta de 40 sonidos de sirenas y 40 sonidos de bocinas, y el sistema logró una tasa de acierto del 100%.

El sistema IA demostró una capacidad sobresaliente para detectar de manera precisa y confiable las clases de sirenas y bocinas en todo momento. Cuando no se detectaba ninguna de estas clases de sonidos, el sistema informaba de manera adecuada sobre la ausencia de eventos relevantes. Una vez confirmada su eficacia, se procedió a la siguiente fase de nuestro experimento.

En la segunda etapa, se puso a prueba el sistema de localización de fuentes sonoras utilizando el algoritmo SRP-PHAT. Se seleccionó al azar una sirena y una bocina de la base de datos ESC-50 y se simularon situaciones realistas haciendo sonar estas fuentes sonoras a través de uno de los altavoces durante 20 segundos. Dado que las tramas de audio se tomaron en intervalos de 2 segundos, se generaron 10 predicciones de ubicación para cada fuente simulada en incrementos de 30 grados. Como resultado, el sistema proporcionó una representación precisa y útil de la procedencia de los sonidos en el entorno vehicular, como se ilustra en la Figura 6.

Es importante destacar que, si bien se observó un ligero aumento en el error de predicción en alrededor del grado 270, esto se alinea con la disposición de los micrófonos en relación con el entorno sonoro. Como se ilustra en la Figura 5, los micrófonos están ubicados más cerca de la zona comprendida entre 0 y 180 grados que de la zona entre 180 y 360 grados. Sin embargo, es fundamental remarcar que el error se mantuvo dentro del margen de resolución angular establecido, lo que garantiza una representación precisa y útil de la localización de fuentes sonoras en condiciones realistas.

7. CONCLUSIONES

En este estudio, hemos desarrollado un sistema basado en inteligencia artificial (IA) que demuestra una capacidad sobresaliente tanto en la detección de sonidos cruciales, como sirenas y bocinas, como en la localización precisa de su procedencia en entornos vehiculares. Nuestro enfoque se apoya en dos componentes clave: el algoritmo SRP-PHAT, que ofrece resultados con una resolución angular de 30 grados, y un sistema de IA que opera como detector/disparador para activar este algoritmo.

La eficiencia de nuestro sistema es notoria en la detección temprana de eventos sonoros relevantes. El sistema de IA, que funciona como disparador, se conecta a uno (o dos para mejorar la detección) de los cuatro micrófonos en el exterior del vehículo de prueba y opera en tiempo real,

analizando tramas de audio de 2 segundos de manera continua. Este componente demuestra una capacidad excepcional para detectar la presencia de sirenas y bocinas en medio del entorno sonoro, lo que desencadena automáticamente la activación del algoritmo SRP-PHAT para la localización precisa de la fuente sonora.

Durante las pruebas de simulación, nuestro sistema sobresale en su capacidad para predecir con precisión la dirección de llegada de las fuentes de sonido. Esta capacidad de localización precisa es particularmente relevante, ya que, en el entorno automovilístico típico, las direcciones de interés suelen estar separadas por intervalos de 30 grados. El rendimiento de nuestro sistema, al superar este umbral de precisión, proporciona una comprensión sólida del entorno sonoro, lo que se traduce en decisiones más informadas y, en última instancia, en una conducción más segura.

En resumen, nuestro sistema representa un avance en la intersección entre la inteligencia artificial y la seguridad vial, ofreciendo una solución efectiva y confiable para la detección y localización de eventos sonoros esenciales en el contexto automovilístico. La eficiencia tanto en la detección de sonidos como en la localización de su origen contribuye a la seguridad en las carreteras y al desarrollo de vehículos más seguros y autónomos.

8. LINEAS FUTURAS

Este estudio sienta una base para futuras investigaciones y aplicaciones en el campo de la seguridad vial y la conducción autónoma. Varias áreas de desarrollo y mejora emergen como oportunidades prometedoras. En primer lugar, consideramos la ampliación de la variedad de sonidos detectables. Aunque nuestro estudio se centró en sirenas y bocinas, la capacidad de detectar y clasificar otros sonidos pertinentes en el entorno automovilístico, como frenazos o neumáticos chirriantes, podría enriquecer aún más la seguridad en las carreteras.

Además, explorar la posibilidad de utilizar modelos basados en Transformers para la detección y localización de sonidos en tiempo real podría representar un avance significativo en la eficiencia computacional y la precisión de nuestro sistema. Esta técnica podría aplicarse tanto al disparador como al localizador, mejorando la capacidad de nuestro sistema para identificar sonidos relevantes y determinar sus ubicaciones de manera más precisa.

La integración de nuestra tecnología con vehículos autónomos también se presenta como un área de investigación prometedora. Incorporar sistemas como el nuestro en vehículos autónomos podría mejorar su capacidad para interactuar con el entorno y tomar decisiones más fundamentadas. Investigar la integración de esta tecnología en el contexto de vehículos autónomos es un paso natural hacia un futuro más seguro y eficiente en la conducción autónoma.

Estas direcciones de investigación prometen avanzar aún más en la intersección de la inteligencia artificial y la seguridad vial, con la visión de hacer que las carreteras sean más seguras y eficientes para todos los usuarios.

9. REFERENCIAS

[1] Cai M., Zhong S., Wang H., Chen Y., Zeng W. Study of the traffic noise source intensity emission model and the frequency characteristics for a wet asphalt road. Appl. Acoust. 2017; 123:55–63. doi: 10.1016/j.apacoust.2017.03.006

[2] Wang H., Luo P., Cai M. Calculation of Noise Barrier Insertion Loss Based on Varied Vehicle Frequencies. Appl. Sci. 2018; 8:100. doi: 10.3390/app8010100.

[3] Gong, Y., Chung, Y.-A., Glass, J. AST: Audio Spectrogram Transformer. Proc. Interspeech 2021, 571-575, doi: 10.21437/Interspeech.2021-698

[4] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, Polosukhin, Illia. (2017). "Attention is All You Need." Neural Information Processing Systems (NeurIPS), páginas 30-38

[5] J. F. Gemmeke et al., "Audio Set: An ontology and human-labeled dataset for audio events," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 776-780, doi: 10.1109/ICASSP.2017.7952261.

[6] F. Zhuang et al., "A Comprehensive Survey on Transfer Learning," in Proceedings of the IEEE, vol. 109, no. 1, pp. 43-76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[7] Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 1041–1044. ACM. 2014

[8] J. Dibiase, A High-Accuracy Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays, 2000.

[9] M. Cobos, A. Marti and J. J. Lopez, "A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling," in IEEE Signal Processing Letters, vol. 18, no. 1, pp. 71-74, Jan. 2011, doi: 10.1109/LSP.2010.2091502.

[10] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson III, “Performance of real-time source-location estimators for a large-aperture microphone array,” IEEE Transactions on Speech and Audio Processing, vol. 13, no. 4, pp. 593–606, 2005.

[11] K. J. Piczak. ESC: Dataset for Environmental Sound Classification. In Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015-1018, ACM, 2015.