TECNIACÚSTICA
CUENCA 2023

RECONOCIMIENTO DE FUENTES A PARTIR DE MEDIDAS SONOMÉTRICAS PARA EL ANÁLISIS DE PAISAJES SONOROS

Modan Tailleur [1]
Pierre Aumond [2*]
Mathieu Lagrange [1]
Vincent Tourre [3]

[1]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[2]Univ Gustave Eiffel, CEREMA, UMRAE, F-44344 Bouguenais, France
[3]Nantes Université, École Centrale Nantes, CNRS, AAU, UMR 1563, F-44000 Nantes, France

RESUMEN

La medida sonométrica estándar (IEC 61672-1) proporciona espectrogramas en bandas de tercio de octava con una ventana cada 1 s (lento) y 125 ms (rápido). Esta medida se utiliza a menudo en aplicaciones de monitoreo a largo plazo en la acústica urbana, ya que requiere poca capacidad de almacenamiento y preserva la privacidad. Sin embargo, la composición de la escena sonora se pierde y el análisis del entorno sonoro se dificulta. En los últimos años, los algoritmos de clasificación pre-entrenados como YAMnet o PANN han permitido una calidad de clasificación suficiente para tales análisis. Sin embargo, la mayoría de estos algoritmos requieren representaciones espectrales de Mel con una ventana temporal muy rápida (*e.g.* 10 ms). En este trabajo, proponemos una arquitectura de red neuronal convolucional para la transcodificación de espectrogramas de tercio de octava lentos o rápidos, de modo que puedan ser utilizados como entrada para modelos robustos pre-entrenados. En comparación con la creación de un nuevo modelo que tomaría como entrada espectrogramas de tercio de octava rápidos, esta aproximación es más eficiente y requiere menos entrenamiento. Los experimentos muestran que el modelo propuesto tiene una precisión de clasificación relevante para el análisis del paisaje sonoro.

ABSTRACT

The standard sound lever meter measurement (IEC 61672-1) provides spectrograms in 1/3 octave bands with a window every 1 s (slow) and 125 ms (fast). This measurement is often used in long-term monitoring applications in urban acoustics, as it requires little storage capacity and preserves privacy. However, the composition of the sound scene is lost and the analysis of the sound environment is made difficult. In recent years, pre-trained classification algorithms such as YAMnet or PANN have allowed a sufficient classification quality for such analyses. However, most of these algorithms require Mel spectral representations with a very fast temporal window (e.g. 10 ms). In this work, we propose a convolutional neural network architecture for transcoding slow or fast 1/3 octave spectrograms, so that they can be used as input for pre-trained robust models. Compared to the creation of a new model that would take fast 1/3 octave spectrograms as input, this approach is more efficient and requires less training. The experiments show that the proposed model has a relevant classification accuracy for soundscape analysis.

***Palabras Clave*—** Paisaje sonoro, clasificación, fuentes sonoras.

1. INTRODUCCIÓN

En los últimos años, varios modelos de clasificación de fuentes de sonido han obtenido reconocimiento por su robustez y precisión. Entre ellos, los modelos pre-entrenados YAMNet y PANNs han surgido como modelos potentes capaces de predecir la presencia de más de 500 fuentes de sonido, gracias a su entrenamiento en la extensa base de datos Audioset [1]. Estos modelos utilizan representaciones espectrales muy finas (e.g. 63 bandas Mel, 10 ms) en datos de entrada (muchas veces pre-procesada internamente desde una señal audio).

* ***Autor de contacto***: *pierre.aumond@univ-eiffel.fr*

La medida sonométrica estándar (IEC 61672-1) proporciona espectrogramas en bandas de tercio de octava con una ventana cada 1 s (lento) y 125 ms (rápido). Los espectrogramas de tercio de octava rápidos ofrecen varias ventajas sobre los espectrogramas Mel/10ms o el audio para aplicaciones de monitoreo a largo plazo. En primer lugar, hacen que las grabaciones sean incomprensibles [2]. Además, son más livianos, con una velocidad de bits aproximadamente 138 veces menor que la de las grabaciones de formas de onda mono de 16 bits, 32 kHz, y alrededor de 30 veces menor que la de las grabaciones Mel, 10ms. Finalmente, es importante destacar que numerosos ingenieros acústicos, investigadores y observatorios de ruido han acumulado en sus bases de datos extensas colecciones de estos espectrogramas capturados en los últimos años. Estos recursos representan horas y días de datos que presentan una valiosa oportunidad para un análisis exhaustivo.

Gontier et al. abordaron tareas de clasificación de múltiples etiquetas en entornos urbanos utilizando una Red Neuronal Convolucional (CNN) entrenada directamente en espectrogramas de tercio de octava [3]. Si bien su modelo mostró un buen rendimiento en el conjunto de datos Cense Lorient [5], carece de robustez en otros conjuntos de datos grabados en tercios de octava. Esta limitación surge en parte del entrenamiento del modelo en conjuntos de datos altamente homogéneos.

Para usar modelos pre-entrenados, se puede entonces imaginar una tarea de super resolución, generando señales completas a partir de tercio de octava, 125ms. Hasta donde sabemos, no hay otros trabajos disponibles específicamente para la tarea en cuestión en el procesamiento de audio. En el campo de la visión por computadora, se han propuesto varios métodos para abordar la tarea de convertir un conjunto de características en otro conjunto de características (feature translación) [4]. Se puede emplear una pseudoinversa para recuperar un espectrograma Mel, 10 ms a partir de un espectrograma de tercio de octava, 125ms y la información temporal puede ser interpolada. Esto resultaría en un espectrograma Mel difuminado, que podría considerarse análogo a una imagen ruidosa en un paradigma de eliminación de ruido. Se han utilizado métodos de codificación automática, métodos adversarios y métodos de difusión en tareas de súper resolución y eliminación de ruido. La idea es, entonces, de aplicar tareas de súper resolución en el campo de la espectrogrametría y verificar el potencial de tal metodología.

2. METODO

2.1. Transcodificador o *Transcoder*

En este estudio, empleamos el método de cálculo de tercio de octava con una ventana de 125 ms, como se describe en el proyecto Lorient, para generar el espectrograma utilizado. Este enfoque implica la subdivisión del rango de frecuencia de 20 Hz a 12,5 kHz en 29 bandas de tercio de octava y utiliza una ventana temporal rectangular de 125 ms.

Es relevante destacar que, aunque tanto el modelo YAMNet como los clasificadores PANNs requieren espectrogramas Mel de 10 ms como entrada, los espectrogramas utilizados por estos modelos presentan diferencias sutiles entre sí. Por lo tanto, es necesario realizar ajustes en el transcodificador propuesto para PANNs en este contexto a fin de adaptarlo para su uso con YAMNet.

La arquitectura del modelo de transcodificación propuesto consta de dos componentes principales: un transcodificador basado en la pseudoinversa (PINV) y una Red Neuronal Convolucional (CNN). El primer componente, el transcodificador PINV, realiza la reconstrucción del espectrograma de banda fina (63 Mel bins, 10 ms) a partir del espectrograma de tercio de octava de manera eficiente mediante un proceso de pseudoinversa. El segundo componente, la red neuronal convolucional (CNN), asume la tarea de refinar el espectrograma Mel, 10 ms mediante la incorporación de información residual.

Para el entrenamiento de nuestro modelo de CNN, seguimos un enfoque similar al de profesor-alumno. En este enfoque, aprovechamos las salidas generadas por clasificadores pre-entrenados, como YAMNet o PANNs, para guiar la generación del espectrograma Mel. Es decir, utilizamos la información proporcionada por estos clasificadores como una guía para mejorar la calidad y la precisión del espectrograma generado por nuestra red neuronal convolucional.

El conjunto de datos utilizado para entrenar y evaluar el transcodificador es el TAU Urban Acoustic Scenes 2020 Mobile dataset [5]. Este conjunto contiene clips de audio de 10 segundos de duración de 10 escenarios acústicos diferentes, como aeropuertos, centros comerciales, estaciones de metro y más. Utilizamos solo el subconjunto de desarrollo y datos del dispositivo A, con 29 horas y 20 minutos de audio. Dividimos aleatoriamente este subconjunto en conjuntos de entrenamiento (70%), validación (15%) y evaluación (15%). Los archivos de audio se normalizaron usando el valor absoluto máximo. Es relevante destacar que el enfoque profesor-alumno permite aprovechar conjuntos de datos no anotados, ya que PANN (el modelo guía) proporciona las anotaciones.

La metodología completa y la arquitectura detallada de las redes se encuentran disponibles en Tailleur et al. 2023 [6].

2.2. Dataset

En el marco del proyecto ADEME GRAFIC, se llevaron a cabo cuatro rondas de evaluaciones perceptivas en un recorrido de 2.1 km en el 13er distrito de París. Los detalles completos del conjunto de datos se pueden encontrar en Aumond et al. [7], pero aquí presentamos los aspectos esenciales.

Se realizaron evaluaciones perceptivas simultáneas durante pausas de alrededor de 3 a 5 minutos en 19 puntos a lo largo del recorrido. Se recopilaron grabaciones de audio y mediciones físicas utilizando un sonómetro móvil durante las caminatas de reconocimiento en el terreno y en los días de evaluación, manteniendo cierta distancia del grupo para evitar influencias en las mediciones. El tiempo total del recorrido fue de aproximadamente 45 minutos, con un promedio de 115 metros entre cada punto. El recorrido abarcaba desde lugares muy silenciosos hasta muy ruidosos en la escala de la ciudad de París.

Los recorridos se realizaron en dos días diferentes: el 23 de marzo de 2015 entre las 11:00 y las 12:00, y entre las 15:00 y las 16:00, así como el 30 de marzo de 2015 en los mismos horarios. Participaron 8, 8, 11 y 9 sujetos respectivamente, sumando un total de 37 participantes. En cada punto de evaluación, el grupo se detenía para completar un cuestionario y calificar 16 variables en una escala del 1 al 11. En este contexto, nos centramos en tres variables: el tiempo percibido de presencia de voces, canto de aves y tráfico vehicular. Estas tres fuentes están correlacionadas con los entornos sonoros sociales, mecánicos y naturales, componentes esenciales de los entornos urbanos que influyen en la percepción [8].

Aunque se evaluaron 19 lugares diferentes, se obtuvieron evaluaciones de alrededor de diez personas para un total de 74 entornos sonoros. En resumen, el conjunto de datos utilizado en este estudio corresponde a los 74 entornos sonoros con calificaciones promedio de aproximadamente 10 participantes en las tres categorías: tiempo percibido de presencia de voces, canto de aves y tráfico vehicular.

2.2. Métrica de rendimiento

Las métricas globales de rendimiento de nuestro transcodificador se encuentran detalladas en el artículo de referencia [6]. En este trabajo, aplicamos el transcodificador en el contexto de un paisaje sonoro. El algoritmo PANN puede ser empleado directamente en los 74 audios recolectados. Para cada segmento de audio de 10 segundos, el algoritmo genera un índice en el rango de 0 a 1 para cada una de las 527 clases, indicando la presencia o ausencia de la clase en los 10 segundos previos. Dado que nuestros clips de audio tienen una duración aproximada de un minuto, obtenemos alrededor de 6 valores por cada clase y por cada segmento de audio.

Basándonos en esto, evaluamos la correlación entre la evaluación perceptiva promedio proporcionada por los peatones y el promedio de estos índices para tres clases específicas: "Vehicle", "Speech" y "Chirp, Tweet". Luego, comparamos estos resultados con el rendimiento de los modelos PANNs, pero aplicados no directamente al audio, sino a las señales sonométricas en tercios de octava, 125 ms, gracias a la transcodificación.

3. RESULTADOS

La Tabla 1 presenta la correlación de Pearson entre el tiempo de presencia percibido promedio para las fuentes sonoras de Voz, Cantos de Aves y Tráfico Véhicular, y el promedio de las evaluaciones proporcionadas por PANN y el método propuesto, para las categorías de "Vehicle", "Speech" y "Chirp, Tweet".

Tabla 1. Correlación de Pearson entre el tiempo de presencia percibido promedio para las fuentes sonoras de Voz, Cantos de Aves y Tráfico Véhicular, y el promedio de las evaluaciones proporcionadas por PANN y el método propuesto, para las categorías de "Vehicle", "Speech" y "Chirp, Tweet".

	PANN	**Transcodificador + PANN**
Voz	0,51	0,47
Cantos de aves	0,72	0,70
Trafico rodado	0,72	0,77

Se puede observar que los resultados de correlación obtenidos entre PANN por sí solo y a través del transcodificador son muy similares. En otras palabras, a partir de una señal de baja resolución, gracias al transcodificador entrenado previamente, las actuaciones solo se degradan ligeramente. Es importante destacar que la parte de varianza no explicada por el indicador puede deberse tanto a un error de clasificación como a la incertidumbre inherente a las mediciones perceptivas, ya que nuestra línea base se fundamenta en evaluaciones de percepción que están sujetas a todos los sesgos propios de este tipo de medidas.

4. DISCUSION Y CONCLUSION

En este estudio, proponemos un enfoque de profesor-alumno para aprender un transcodificador que convierte representaciones espectrales de tercio de octava con una ventana de 125 ms en espectrogramas Mel de 10 ms. Estos espectrogramas se utilizan como entrada para clasificadores pre-entrenados como PANN y YAMNet. Nuestra técnica

muestra correlaciones muy similares a las obtenidas directamente con PANN en términos de tiempo de presencia percibido para cantos de aves, voz y tráfico vehicular, a pesar de las limitaciones de resolución temporal y frecuencial de los espectrogramas de tercio de octava. Además, la precisión obtenida ($r > 0,5$, $p < 0,001$) podría ser de gran relevancia para el análisis del entorno sonoro.

En términos generales, los resultados de precisión del modelo para clasificar muestras sonoras o sonométricas se pueden encontrar en Tailleur et al. [6].

No obstante, una limitación de nuestro enfoque es que se requiere entrenar un nuevo transcodificador para cada representación espectral-temporal, con el fin de adaptarse a diferentes parámetros como el número de bins Mel, la cantidad de frecuencias, el tamaño del salto, la frecuencia de muestreo, entre otros. Para abordar esta limitación, futuras investigaciones podrían explorar la reconstrucción completa del audio a partir de una representación espectral de tercio de octava rápida o lenta, lo que permitiría utilizar clasificadores pre-entrenados como el avanzado modelo PANN Wavegram-Logmel-CNN (que toma audio como entrada). Algunas pruebas preliminares también se presentan en el "companion page" siguiente:
github.com/modantailleur/paperSpectralTranscoder

5. AGRADECIMIENTOS

Este trabajo fue financiado en parte por el proyecto ANR-20-THIA-0011 "AiBy4".

6. REFERENCIAS

[1] J. F. Gemmeke *et al.*, « Audio Set: An ontology and human-labeled dataset for audio events », in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, mars 2017, p. 776-780. doi: 10.1109/ICASSP.2017.7952261.

[2] F. Gontier, M. Lagrange, P. Aumond, A. Can, et C. Lavandier, « An Efficient Audio Coding Scheme for Quantitative and Qualitative Large Scale Acoustic Monitoring Using the Sensor Grid Approach », *Sensors*, vol. 17, nº 12, p. 2758, nov. 2017, doi: 10.3390/s17122758.

[3] F. Gontier, C. Lavandier, P. Aumond, M. Lagrange, et J.-F. Petiot, « Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques », *Acta Acust. United Acust.*, vol. 105, nº 6, p. 1053-1066, 2019.

[4] W. Kuang, Y.-L. Chan, S.-H. Tsang, et W.-C. Siu, « Fast HEVC to SCC Transcoder by Early CU Partitioning Termination and Decision Tree-Based Flexible Mode Decision for Intra-Frame Coding », *IEEE Access*, vol. 7, p. 8773-8788, 2019, doi: 10.1109/ACCESS.2018.2890720.

[5] A. Mesaros, T. Heittola, et T. Virtanen, « A multi-device dataset for urban acoustic scene classification ». arXiv, 11 octobre 2018. doi: 10.48550/arXiv.1807.09840.

[6] M. Tailleur, M. Lagrange, P. Aumond, et V. Tourre, « Spectral trancoder: using pretrained urban sound classifiers on undersampled spectral representations », in *8th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2023.

[7] P. Aumond, A. Can, B. De Coensel, C. Ribeiro, D. Botteldooren, et C. Lavandier, « Global and continuous pleasantness estimation of the soundscape perceived during walking trips through urban environments », *Appl. Sci.*, vol. 7, nº 2, p. 144, 2017.

[8] F. Aletta, J. Kang, et Ö. Axelsson, « Soundscape descriptors and a conceptual framework for developing predictive soundscape models », *Landsc. Urban Plan.*, vol. 149, p. 65-74, mai 2016, doi: 10.1016/j.landurbplan.2016.02.001.