DOI: 10.61782/fa.2025.0089

FORUM ACUSTICUM
EURONOISE 2025

ACOUSTIC SCENE DESCRIPTION USING ACOUSTIC IMAGING AND MACHINE LEARNING

Enzo Tiffeneau[1,2] **Quentin Leclère**[1] **Simon Bouley**[2]
[1] Univ Lyon, INSA Lyon, LVA, 25 bis av. Jean Capelle F-69621, Villeurbanne Cedex, France.
[2] MicrodB, 28 Chemin du Petit Bois, F-69134, Ecully Cedex, France.

ABSTRACT

The growing concern about acoustic monitoring in domains such as construction site supervision, drone tracking or wildlife sighting calls for the enhancement of acoustic scene description. Combining source localization and statistical learning, the position, the level, and the identity of each source present in the scene can be predicted. Source localization brings together microphone array and acoustic imaging techniques to draw acoustic maps. Moreover, time-domain techniques such as Delay-and-Sum beamforming or CLEAN-T allow to extract audio signals of each source present in the scene. Once each source is localized and isolated, deep-learning models based on Transformer architectures are used to identify the collected sound sources. These models mainly rely on neural networks fed with time-frequency spectrograms. However, based on the source energy, the mentioned phased-array techniques may fail at localizing impulsive or tonal sources in intricate acoustic scene, which tend to vanish in the background noise. Therefore, a novel time-deconvolution technique denoted as CLEAN-STFT and based on CLEAN-T algorithm, is proposed to reveal low-energy that would not emerge previously. Taking advantage of both time and frequency dimensions of targeted source spectrograms, this method allows a refined description of acoustic scene and can seamlessly feed deep-learning algorithms.

Keywords: *source separation, acoustic imaging, acoustic scene description).*

1. INTRODUCTION

The description of acoustic scenes is based on the localization, level estimation and separation of the noise sources present, enabling their nature to be identified using sound recognition models. This problem arises in various contexts, particularly for acoustic monitoring solutions [1]. The main challenge lies in the ability to handle sources of diverse nature (broadband, tonal, impulse, stationary, etc.), some of which can overlap in time-frequency domain, complicating the analysis of acoustic scenes.

Existing approaches that address sound scene description come into several categories. They include single-microphone methods such as non-negative matrix factorization (NMF), based on mathematical and signal-processing techniques [2, 3], and deep-learning source separation models such as SUDO [4]. However, these methods have limitations in terms of source separation performance and are unable to localize sources. Other approaches exploit a small number of microphones while combining deep learning, localization and source identification. This strategy is receiving growing interest, particularly in the DCASE (Detection and Classification of Acoustic Scenes and Events) community. However, the performance of these methods remains limited at this stage, mainly due to the interdependence between the localization and identification stages [5, 6].

Acoustic imaging methods based on microphone arrays offer several advantages for the description of sound scenes. They can be used to locate sources in a scene, estimate their sound level and separate them from background noise, while improving the performance of sound recognition models. In particular, the spatial filtering of antennas enhances the signal-to-noise ration (SNR) of targeted source signals [7]. In addition, the spatial separation of sources provides a polyphonic

description of acoustic scenes.

Among acoustic imaging methods, the frequency-domain approach is effective at locating sources and quantifying sound levels, but it does not directly reconstruct the temporal signals required by sound recognition algorithms. An alternative is to combine these methods with a temporal approach: first accurately locate sources in the frequency domain, then, in a second step, extract signals at the locations identified using a temporal approach [8–10]. This approach allows the use of algorithms such as conventional beamforming [11], high-resolution methods such as MUSIC [12] or CAPON [13], as well as deconvolution techniques such as DAMAS [14] or CLEAN-SC [15]. Each method has its own advantages and limitations. Conventional beamforming, while robust, has poor low-frequency resolution and method-related artifacts. MUSIC offers better resolution, but requires prior knowledge of the number of sources, making it unsuitable for automatic monitoring applications. Other methods, such as CLEAN-SC, DAMAS or CAPON, can also be used.

However, a stringent temporal approach can be used, enabling signals from different sources to be located and extracted in a single step, making the method process much simpler. Even though it is slower and more costly than frequency-based methods, the temporal approach offers a number of advantages. As mentioned by Jaeckel [16], this approach enables all frequencies to be processed simultaneously, making it particularly suitable for broadband sources. It does not require long signals, and facilitates the correction of the Doppler effect, a much more complex task in the frequency domain. CLEAN-T [17], based on a temporal deconvolution principle, offers a robust solution for separating the individual contributions of sources in a complex environment, enabling identification of their nature, as well as description of their location and level. The method is not automatic, however, and requires a pre-filtering step. Indeed, CLEAN-T localizes sources based on the entirety of the measured signals, and requires frequency or time filtering to highlight tonal or impulse sources.

To improve source separation and identification, CLEAN-STFT method is proposed and will be presented in section 2. By simultaneously exploiting the temporal and frequency dimensions of spectrograms, this approach uses an adaptive filtering to highlight low-intensity sources, whether impulsive, tonal or masked by ambient noise. By combining this method with deep learning models detailed in section 3, it is then possible to obtain a fine description of sound scenes, opening the way to advanced applications in acoustic monitoring, environmental surveillance and automatic identification of sound sources. Results of the method will be presented in section 4, before concluding with performance and a discussion of the method in section 5.

2. CLEAN-STFT

The proposed method for describing acoustic scenes is based on spatial filtering and source separation using CLEAN-STFT, followed by identification using deep learning models.

2.1 Beamforming delay-and-sum

The first step is to locate and quantify sound sources in a given scene using acoustic imaging methods. CLEAN-STFT is based on the beamforming delay-and-sum (DAS) technique, which exploits signals measured by a microphone array. It is used to estimate the acoustic level of equivalent sources placed at the nodes of a scan grid, representative of the acoustic scene. In this study, the configuration proposed to illustrate the method is shown in Fig. 1.

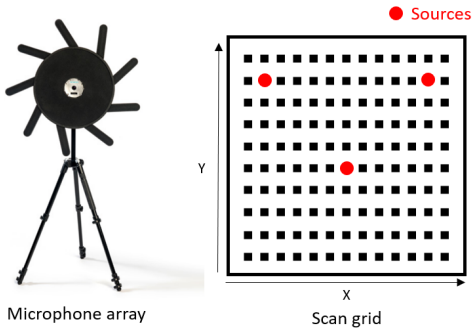

Figure 1. Microphone array (Simcenter Sound Camera) with a scan grid (here a 2D computational plane representing a portion of space), with three stationary noise sources

.

DAS beamforming consists of the following steps:

1. The sound field is measured by a microphone array.
2. Knowing the distances between each microphone and each point on the calculation grid, a correction for sound wave propagation delays is made for each grid node/microphone pair.
3. The phased-shifted signals are summed and averaged to obtain a temporal signal focused on each point of the calculation grid.

This technique reconstructs the temporal signal at each point on the calculation grid. The acoustic level is averaged, and the location of sound sources is estimated by identifying the nodes with the highest levels, taking secondary lobes into account. The latter correspond to artifacts caused by the limited resolution of the microphone antenna.

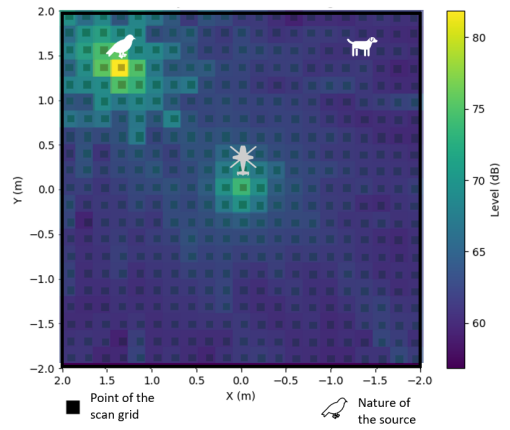

Figure 2. Acoustic map obtained with delay-and-sum beamforming.

The acoustic scene studied in this paper includes three sound sources: a crow cawing at top left, a helicopter in the center and a dog barking at top right. However, due to its impulsive nature and low energy, the barking is masked, blended into a side lobe of the main sources. The acoustic map obtained after beamforming in Fig. 2 shows the main and side lobes associated with the various sources.

2.2 Temporal deconvolution

Temporal deconvolution improves the resolution of acoustic maps by removing artifacts associated with the antenna response, notably the side lobes and most of the main lobe, except for its maximum. As a result, real sources can be more accurately extracted, including for the delay-and-sum beamforming seen in part 2.1.

In the frequency domain, this corresponds to the CLEAN-SC algorithm, while in the time domain, deconvolution is performed by the CLEAN-T algorithm. The latter identifies and successively subtracts dominant sources, while cleaning up their contribution to the residual acoustic map. The method is based on the following steps :

1. **Initialization** : Two maps are initialized. The first one, called dirty map, corresponds to the raw results of DAS beamforming and contains the residual contributions of the sources. The second, the clean map, is initially empty and stores the sources deconvoluted during iterations.
2. **Identification of dominant source**: In the CLEAN-T algorithm, at each iteration, the source with the highest average integrated level is identified as dominant. Another approach is adopted for CLEAN-STFT, developed in section 2.3.
3. **Repropagation and subtraction**: The signal from the dominant source is obtained via DAS beamforming, then repropagated to the microphones to estimate its exact contribution. This contribution is then subtracted from the signals measured by the antenna, eliminating the artifacts associated with this source on the map. The deconvoluted source is then added to the clean map.
4. **Successive iterations**: The process is repeated for the next dominant source, until a predefined stopping criterion is reached. This criterion can be a maximum number of sources, a minimum residual energy threshold or a given display dynamic.

Using this iterative process, temporal deconvolution improves the accuracy of acoustic maps by removing sidelobes and keeping only the location of real sources, as shown in Fig. 3.

2.3 CLEAN-STFT

2.3.1 Passing in time-frequency domain

In an automated monitoring context, CLEAN-T, based on the total energy of the temporal signal, does not

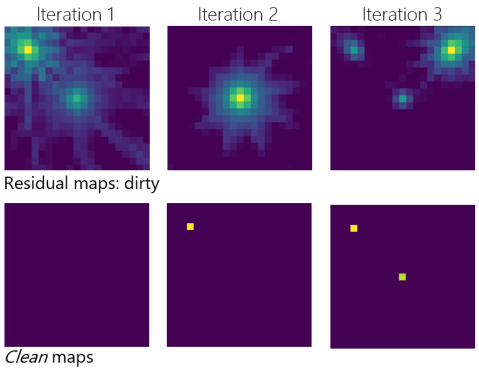

Figure 3. Results by iteration of the CLEAN-T temporal deconvolution method.

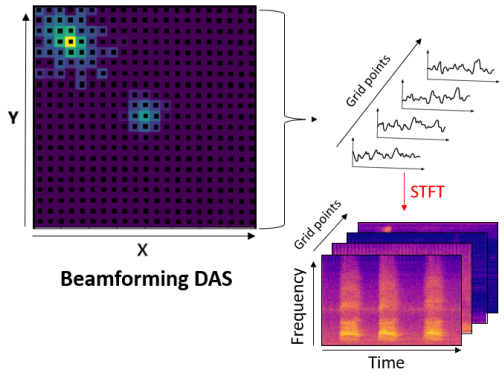

Figure 4. Estimation of focused signals using DAS beamforming in the time-frequency domain with Short-Time Fourier Transform.

directly discriminate between different categories of sound sources. Thus, to detect specific sources, such as tonal or impulsive sources in a noisy environment, signal pre-filtering is required to isolate the relevant frequency band or temporal characteristics. This process requires prior knowledge of the properties of the target source [18]. The CLEAN-STFT method is an extension of CLEAN-T that applies to the field of short-time Fourier transforms (STFT), i.e. frequency and time. The method features adaptive time-frequency windowing to highlight all sources in the acoustic scene.

The method is based on temporal deconvolution with the following principle:

1. The time-domain signal at each grid point is obtained using DAS beamforming. After applying a STFT to all signals, a set of spectrograms is computed for all grid points, as illustrated in Fig. 4.
2. In the proposed method, a pixel corresponds to a point in the time-frequency domain. The number of pixels is equal to the number of frequency bins multiplied by the number of time increments in the spectrogram.
3. For each time-frequency pixel in a spectrogram associated with a grid point, its level is compared to the corresponding pixels across all spectrograms (Figure 5). Only the pixels with the maximum amplitude are retained, forming a maximized spectrogram in the direction of each grid point. If no pixel reaches a maximum at a given position, the corresponding spectrogram remains empty (Figure 6).

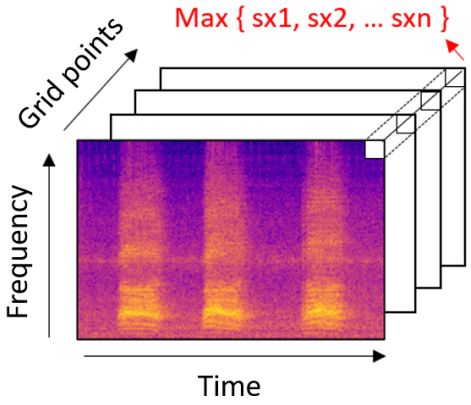

Figure 5. Selection of the maximum level of a time-frequency pixel to define the dominant direction.

4. The inverse Short-Time Fourier Transform (iSTFT) can then be applied to reconstruct a focused time-domain signal for each grid point from these maximized spectrograms. It is important to note that this time-domain signal differs from the initial signal obtained through DAS beamforming: here, it represents the signal whose time-frequency contributions are maximized in the direction of the grid point, rather than simply a weighted sum of the received signals.

2.3.2 *Temporal deconvolution*

Finally, the temporal deconvolution process described in Section 2.2 is applied to the signals obtained from the

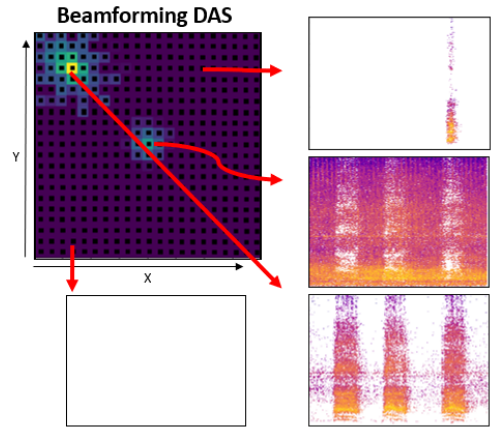

Figure 6. For each grid point, we obtain maximized spectrograms in their respective directions.

iSTFT of the maximized spectrograms. This process uses the pixel size of the spectrogram as the time-frequency window, allowing the emergence of sources with limited extent in this domain. Moreover, thanks to temporal deconvolution, it becomes possible to separate signals that were initially masked by others in the time-frequency domain after several iterations.

1. First, the dominant source is identified by locating the spectrogram containing the pixel with the maximum energy.

2. The contribution of the dominant source is removed through temporal deconvolution.

3. The process is repeated iteratively until all sources are isolated or a stopping criterion is reached (Figure 7).

This method allows the iterative extraction of the various sound sources present in a complex acoustic scene. The adaptive filtering provided by the time-frequency pixel resolution is especially effective for impulsive or frequency-restricted sources. It allows, for example, the emergence of sounds such as a dog barking, which would not be detected using Beamforming DAS or CLEAN-T with identical parameters (Figure 8).

3. SOUND IDENTIFICATION

After separation, the extracted sources are analyzed to identify their nature (speech, music, noise, etc.). This classification is achieved using a recognition model based on neural networks. The model used is based on the Vision Transformer (ViT), a deep learning architecture developed for image classification. This model is adapted and trained for spectrogram recognition to identify different sound sources [19].

The following steps are applied to predict the nature of a sound source:

- Time-frequency spectrograms representing each sound source, from CLEAN-STFT calculation are considered as images, input data of the deep learning algorithm.
- They are converted to the MEL scale, which mimics human perception of sound.
- They are then divided into patches (small square windows), while keeping their temporal and frequency position indices as a vector.
- The model learns to identify specific acoustic structures in these patches (high-pitched sounds, transients, etc.) on the basis of a large dataset of labeled spectrograms on which it has learned to recognize acoustic features.
- Signal analysis by the transformer assigns a probability score to each class in the model. The highest score indicates the most probable class.

Transformers offer many advantages:

- While convolutional neural networks (CNNs) analyze relationships between successive temporal sequences, transformers can establish correlations between temporally distant spectrogram patches.
- The transformer model uses spectrogram patches for identification, so it can handle signals of variable duration.
- The model used reached an accuracy of 93% on the ESC-50 dataset, which contains 2,000 environmental sounds divided into 50 classes [20]. This performance is due to its ability to model complex sound structures.

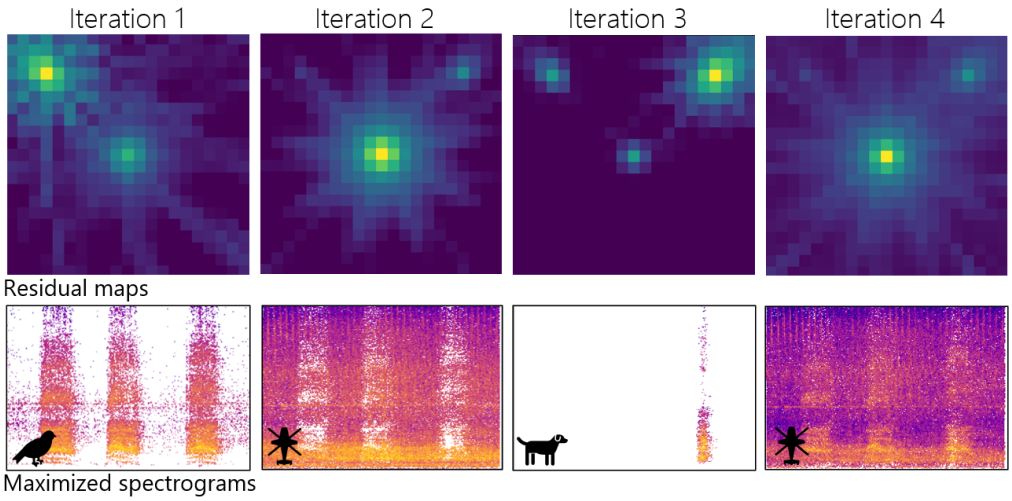

Figure 7. Temporal deconvolution using CLEAN-STFT for the three sources in the scene, with each source accompanied by an icon representing its nature. At iteration 4, the remaining spectral energy of the helicopter, which was previously masked by the more energetic crow caws, is recovered.

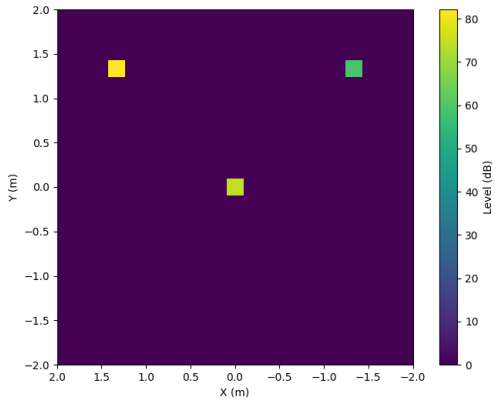

Figure 8. Acoustic map obtained with CLEAN-STFT.

4. RESULTS

To evaluate the proposed method, simulations were conducted using Python to simulate source propagation and generate various acoustic scenes. The sources used come from the ESC-50 dataset, which contains environmental sounds of diverse categories. The scene studied here includes three spatially close sources:

- A helicopter, a continuous broadband source.
- A crow, an impulsive, broadband source.
- A dog barking, whose energy is restricted in time-frequency space.

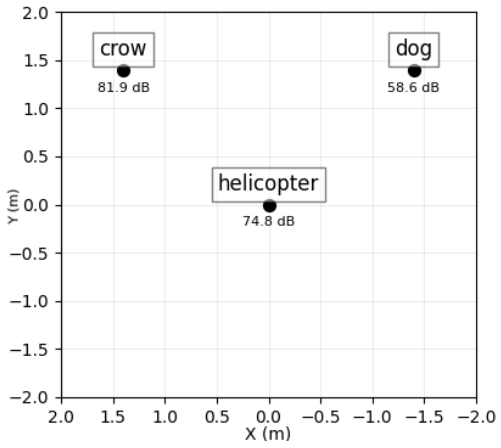

Figure 9. Ground truth of sound scene simulation.

This case study is particularly interesting, as the three sources share part of the time-frequency space, making separation complex. In particular, the dog's bark is masked by both the crow and the helicopter. This situation illustrates the efficiency of CLEAN-STFT in separating overlapping sources in the time-frequency domain, as well as the contribution of adaptive filtering,

which automatically extracts the bark despite its initial masking. The case and its ground truth are illustrated in Fig. 9.

The scene prediction is shown in Fig. 10. The performance obtained is satisfying, with the error on level estimation below 1%, while source localization and identification are correct. These results, and many others not presented here, confirm the validity of the method in a simulation framework. However, a study on a large set of simulated acoustic scenes is still required, as well as experimental validation to confirm the robustness of the model in real-life conditions.

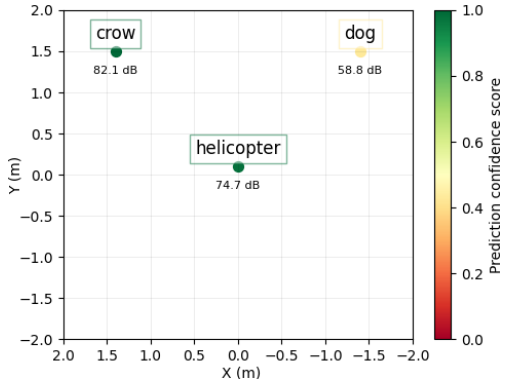

Figure 10. Result of the predicted acoustic scene.

5. CONCLUSION

The CLEAN-STFT method improves the description of sound scenes in acoustic monitoring by using adaptive filtering and exploiting the temporal and frequency dimensions of signals. It enables the isolation of sound sources restricted in time-frequency space, such as impulsive or tonal sounds, which are often masked by ambient noise, and so improves source separation. However, this method is still limited to stationary environments, and needs to be improved to handle dynamic scenes, such as tracking the trajectories of moving sources. Future work should focus on applying this method to real, more complex and mobile acoustic scenes, in order to test its performance under a variety of conditions.

6. REFERENCES

[1] L. Pinel Lamotte, S. Bouley, A. Purier, E. Tiffeneau, and F. Lepercque, "Acoustic Imaging and Machine Learning for Sources Localization and Identification : Application to In Situ Vehicle Pass-By Noise," in *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*, (Turin, Italy), 2024.

[2] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, 2007.

[3] J.-R. Gloaguen, *Estimation du niveau sonore de sources d'intérêt au sein de mixtures sonores urbaines: application au trafic routier*. PhD thesis, Université d'Orléans, 2018.

[4] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -RF: Efficient Networks for Universal Audio Source Separation," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, (Espoo, Finland), IEEE, 2020.

[5] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, 2019.

[6] J. S. Kim, H. J. Park, W. Shin, and S. W. Han, "Adyolo: You look only once in training multiple sound event localization and detection," 2023.

[7] V. Stabellini, *Impact of Noise on Different Neural Network Architectures for Environmental Sound Classification*. PhD thesis, Politecnico di Torino, 2023.

[8] F. Le Courtois, *Caractérisation des sources acoustiques sur le matériel ferroviaire par méthode d'antennerie*. PhD thesis, Université du Maine, 2012.

[9] V. Baron, S. Bouley, M. Muschinowski, J. Mars, and B. Nicolas, "Localisation et identification acoustique de drones par mesures d'antennerie et apprentissage supervisé," *GRETSI*, 2019.

[10] R. Leiba, F. Ollivier, R. Marchiano, N. Misdariis, J. Marchal, and P. Challande, "Acoustical Classification of the Urban Road Traffic with Large Arrays of Microphones," *Acta Acustica united with Acustica*, vol. 105, 2019.

[11] B. Van Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, 1988.

[12] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, 1986.

[13] P. Stoica, Zhisong W., and J. Li, "Robust Capon beamforming," *IEEE Signal Processing Letters*, vol. 10, 2003.

[14] K. Ehrenfried and L. Koop, "Comparison of iterative deconvolution algorithms for the mapping of acoustic sources," *AIAA Journal*, vol. 45, 2007.

[15] P. Sijtsma, "Clean Based on Spatial Source Coherence," *International journal of aeroacoustics*, vol. 6(4), 2007.

[16] O. Jaeckel, "Strengths and weaknesses of calculating beamforming in the time domain," in *1th Berlin Beamforming Conference BEBEC 2006*, 2006.

[17] R. Cousson, Q. Leclère, M.-A. Pallas, and M. Bérengier, "A time domain clean approach for the identification of acoustic moving sources," *Journal of Sound and Vibration*, vol. 443, 2019.

[18] R. Leiba, Q. Leclere, and E. Julliard, "Application of the cleant methodology to flyover noise measurements," in *9th Berlin Beamforming Conference BEBEC 2022*, (Berlin, Germany), 2022.

[19] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[20] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 336–340, IEEE, 2024.