POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Sound Event Localization And Detection With Mel-Scaled Frequency-Sliding Generalized Cross-Correlation

Laurea Magistrale in Music And Acoustic Engineering

Author: Federico Angelo Luigi Ferreri

Advisor: Luca Comanducci

Co-advisors: Francesca Ronchini, Maximo Cobos

Academic year: 2023-2024

1. Introduction

The analysis and recognition of sound events are evolving research areas with applications in acoustic surveillance, virtual reality, and human-machine interaction. Sound Event Localization and Detection (SELD) is a fundamental challenge in acoustic engineering and signal processing, aiming to enhance accuracy in sound localization and distance estimation for comprehensive three-dimensional acoustic scene understanding. This study builds on Time Delay Estimation (TDE), crucial for localization through Time Difference of Arrival (TDOA) estimation using microphone arrays. Traditional methods like Generalized Cross-Correlation (GCC) and its variant GCC-PHAT have long been used for TDE by computing the cross-correlation function between two signals and identifying the peak corresponding to the estimated time delay. However, they assume equal frequency contributions, which is often invalid in real-world environments where reverberation, noise, and multipath effects distort the cross-correlation function, leading to inaccurate delay estimates [2]. Frequency-Sliding Generalized Cross-Correlation (FS-GCC) addresses these limitations with a sliding-window approach, enhancing robustness by decomposing signals into sub-bands and computing cross-correlations separately for each frequency range [1]. FS-GCC outperforms GCC-PHAT in noisy and reverberant conditions by adaptively emphasizing reliable spectral regions. Mel-Frequency Sliding Generalized Cross-Correlation (Mel-FSGCC) further refines this approach by incorporating mel-frequency bands, improving frequency distribution in a perceptually relevant manner. This enhances resolution at lower frequencies, where delay information is most valuable, while appropriately scaling higher frequencies, which are more susceptible to noise [3]. These techniques significantly advance time delay estimation and sound source localization, ensuring greater accuracy despite increased computational demand. This research evaluates the integration of Mel-FSGCC into the SELD framework, replacing GCC to improve source localization and distance estimation. The study systematically compares these methods across datasets with varying reverberation times (T60), noise levels, and microphone array configurations to assess robustness and generalization. Extracted features, encoding critical temporal and spatial information, serve as input to the SELD-

net baseline model, aligned with the DCASE 2024 Task 3 Challenge for standardized evaluation. Performance metrics ensure comparability with existing SELD methodologies. Mel-FSGCC integrates mel-scale spectral analysis to emphasize perceptually relevant frequency components, mitigating noise and reverberation effects while enhancing spatial localization. The increasing demand for accurate SELD systems in applications such as environmental monitoring, robotics, and automated assistance underscores the relevance of this research. By refining feature extraction and time delay estimation, this study provides valuable insights for both theoretical advancements and practical implementations in acoustic signal processing.

2. Problem Formulation

The formal mathematical formulation of this problem begins with the estimation of TDOA using GCC, expressed as:

$$R_{x_1x_2}[\tau] = \int_{-\pi}^{\pi} W(\omega) P_{x_1x_2}(\omega) e^{j\omega\tau} d\omega, \quad (1)$$

where $P_{x_1x_2}(\omega)$ represents the cross-power spectral density between signals, $W(\omega)$ is the weighting function, which in our study corresponds to the PHAT (Phase Transform), defined as $(W(\omega) = 1/|P_{x_1x_2}(\omega)|)$, normalizing the cross-power spectral density to its magnitude for improved robustness in reverberant environments. τ is the estimated delay and $\omega = 2\pi f$. FS-GCC improves this by segmenting the spectrum into multiple sub-bands:

$$R[\tau, i] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Psi(\omega + \omega_i) \Phi(\omega) e^{j\omega\tau} d\omega, \quad (2)$$

where ω_i defines the frequency shift for the i-th band, $\Phi(\omega)$ is a spectral window function centered at $\omega = 0$, and $\Psi(\omega)$ represents the phase-transformed cross-power spectrum. Mel-FSGCC further refines this by employing mel-scale frequency decomposition to refine spectral resolution. This transformation ensures higher frequency resolution at lower frequencies, enhancing spatial feature extraction and mitigating the impact of reverberation and noise. The mel-scale transformation is given by:

$$f_m(\omega) = 2595 \log_{10}\left(1 + \frac{\omega}{700}\right), \quad (3)$$

with the inverse transformation determining the center frequencies of mel-scaled sub-bands:

$$\omega_i = 700\left(10^{\frac{m_i}{2595}} - 1\right), \quad i = 0, \ldots, L-1, \quad (4)$$

where m_i represents the index of the sub-band on the mel scale, and L is the total number of sub-bands.

3. Proposed Method

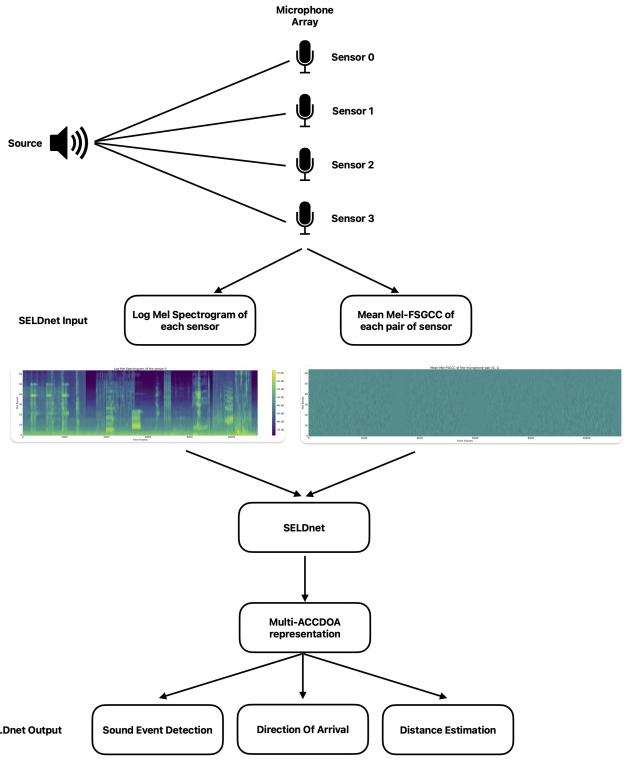


Figure 1: Diagram of the feature extraction process and SELDnet architecture. It illustrates the integration of Mel-FSGCC within the MIC + Multi-ACCDOA framework. The diagram encompasses audio acquisition via a microphone array, extraction of log-Mel spectrograms and mean Mel-FSGCC for microphone pairs, and the network's output, which comprises sound event detection and spatial localization of sound sources. Additionally, the log-Mel spectrogram and the mean Mel-FSGCC of an audio file are represented, specifically showing the log-Mel spectrogram of the first microphone (sensor 0) and the mean Mel-FSGCC of the first microphone pair (sensors 0,1).

3.1. Mel-FSGCC Within SELDnet

This study builds on the DCASE 2024 Task 3 challenge[1] by investigating the impact of modifying feature extraction within the SELDnet architecture. Specifically, it replaces the GCC with Mel-FSGCC to enhance localization accuracy and distance estimation by addressing limitations in traditional time-delay representations. The evaluation focuses on the MIC + GCC + Multi-ACCDOA baseline configuration for the Audio-only Track, where GCC is substituted with Mel-FSGCC to assess its effect on localization performance and robustness. Multi-ACCDOA enables the detection and localization of overlapping sound events from the same class, further refining the system's capability to handle complex acoustic scenarios. The objective is to determine whether the MIC + Mel-FSGCC + Multi-ACCDOA configuration improves source localization and distance estimation compared to the original MIC + GCC + Multi-ACCDOA setup. The study utilizes multichannel audio recordings captured with a microphone array, from which four equidistant microphones are selected to ensure a uniform spatial representation. Feature extraction involves computing log-Mel spectrograms for each microphone channel and obtaining Mel-FSGCC for each microphone pair. To enhance robustness against noise and reverberation, the mean Mel-FSGCC is computed across microphone pairs. A Mel filter bank with 64 bands is applied both to compute log-Mel spectrograms and to serve as a Mel-scaled frequency partition for Mel-FSGCC computation.

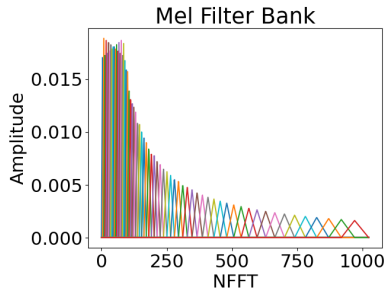


Figure 2: Mel Filter Bank employed in our study.

The final feature set consists of log-Mel spectrograms and mean Mel-FSGCC features, which provide a comprehensive representation of the acoustic scene. Using the Multi-ACCDOA representation, SELDnet produces two main outputs: a sound event detection component that identifies the presence of a specific sound event at a given moment, and a sound localization component that predicts the spatial position of the source, including its direction of arrival and estimated distance.

[1]DCASE 2024 Challenge link

4. Experimental Setup and Evaluation

In this section, we describe the experimental setup used to evaluate the performance of the proposed method. Subsection (4.1) presents the dataset utilized for training and testing. Subsection (4.2) outlines the baseline system configurations and the feature extraction process. We further discuss the evaluation protocols and metrics established for assessing SELD performance in (4.3).

4.1. Dataset

The dataset used in this study was generated using SpatialScaper, a library designed for simulating and augmenting soundscapes for SELD. To ensure a rigorous evaluation under varying acoustic conditions, separate datasets were generated, each tailored to specific T60, noise levels, and microphone array configurations. These datasets were created within a virtual rectangular room of 15 m × 20 m × 3.5 m, modeled using Pyroomacoustics, with five sound sources and a microphone array. For each configuration, distinct Room Impulse Responses (RIRs) were computed and subsequently used for spatialization and audio file generation.

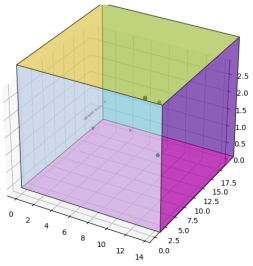


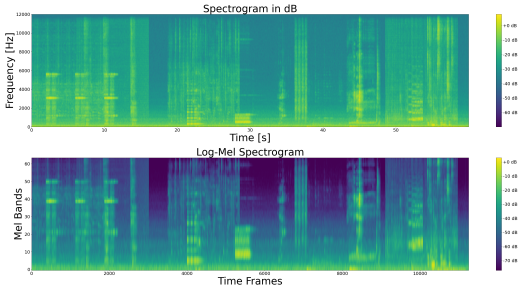
Figure 3: Visualization of the virtual room used for dataset generation. The 'x' markers indicate the positions of the microphones, while the other markers represent the five sound sources.

Each dataset consists of 12000 minutes of recordings, maintaining controlled conditions while

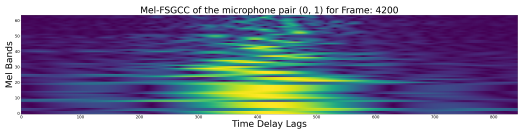
systematically altering T60, noise level, and microphone distances. This methodological approach enables a comprehensive assessment of the model's performance across diverse acoustic environments. The soundscapes include an average of 15 foreground sound events per recording, with a maximum of three overlapping sources per frame. Foreground sound samples originate from the FSD50K and FMA datasets, while background noise components are sourced from the TAU-SRIR database. Each sound event was generated with SNR varying from 5 dB to 30 dB randomly. All recordings are sampled at 24 kHz, with annotations provided at a 100 ms frame resolution. The dataset encompasses eight distinct sound event classes, including female and male speech, telephone ringing, laughter, domestic sounds, footsteps, music, and musical instruments. Each sound event is labeled with its azimuth (-180° to 180°), elevation (-90° to 90°), and absolute distance. The dataset is designed to align with DCASE 2024 Task 3, as the study builds upon the SELDnet baseline model established in the challenge.

4.2. Baseline System Configurations and Features Extraction

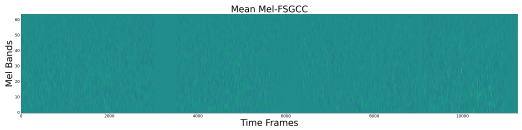
The experimental setup focuses on the Audio-only track, where SELDnet processes multichannel recordings from a subset of four microphones within the available microphone arrays. The log-Mel spectrograms are computed to capture the spectral characteristics of each channel, serving as essential input features for the deep learning model (Figure 4a). For the baseline MIC + GCC + Multi-ACCDOA configuration, GCC features are extracted for all six possible microphone pairs (Figure 4d), leveraging time delay cues for spatial localization. In contrast, the proposed MIC + Mel-FSGCC + Multi-ACCDOA configuration replaces GCC with Mel-FSGCC. The Mel-FSGCC is computed for each time frame and constrained to a limited range of lags based on the maximum time delay, which is determined by the maximum distance between microphone pairs within the selected sub-array of four microphones (Figure 4b). The mean of the Mel-FSGCC across all microphone pairs is extracted and used as an input feature (Figure 4c). In both configurations, the extracted features include the log-Mel spectrograms of the four microphone channels, combined with either GCC or Mel-FSGCC-derived spatial features. The resulting feature set is fed into SELDnet.



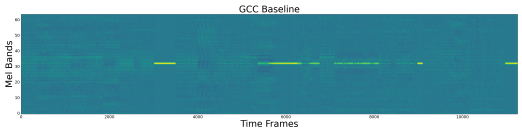
(a) Spectrogram and Log-Mel Spectrogram (in dB).



(b) Mel-FSGCC for the time frame 4200.



(c) Mean of the Mel-FSGCC.



(d) GCC.

Figure 4: Visualization of the extracted features for an audio sample from a dataset generated with a reverberation time (T60) of 0.6 s and a noise level of -43 dB. Figure (a) displays the Spectrogram and Log-Mel Spectrogram (in dB) of the audio signal. Figure (b) illustrates the Mel-FSGCC for time frame 4200, where overlapping sound events occur, extracted from the first microphone pair (sensors 0 and 1) with a spacing of 1 m. Figure (c) shows the mean of the Mel-FSGCC, while Figure (d) presents the corresponding GCC for the same microphone pair.

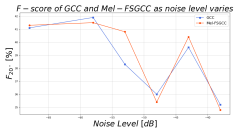
4.3. Evaluation Metrics

The DCASE 2024 Challenge[1] adopts a refined frame-wise evaluation framework for assessing sound event detection, localization accuracy, and distance estimation, ensuring precise time-step-based performance. In this study, we adopt the same evaluation metrics as DCASE 2024, specifically F-score (F20°), Direction of Arrival

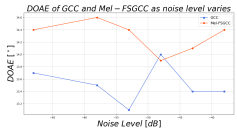
Error (DOAE), and Relative Distance Error (RDE), replacing previous segment-based assessments. The F-score measures detection accuracy, considering a prediction correct only if the class matches, the DOA error is below 20°, and the relative distance error is ≤ 1.0. DOAE and RDE, used as class-aware localization metrics, quantify angular and distance errors without fixed thresholds. The F-score applies separate angular and distance thresholds, ensuring distinct penalization for DOA and distance errors. The transition to frame-wise evaluation improves transparency and simplifies localization assessment. Unlike previous editions, DOAE no longer imposes a 180° penalty for undetected events, leading to a more balanced evaluation framework.

5. Results

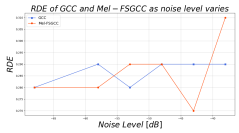
Three different experiments were conducted to analyze the Mel-FSGCC and GCC performances under different acoustic conditions by varying noise levels, reverberation time (T60), and microphone spacing.



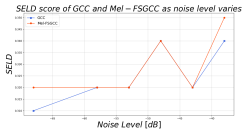
(a) F-score as a function of noise level.



(b) DOAE as a function of noise level.



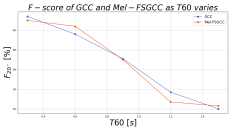
(c) RDE as a function of noise level.



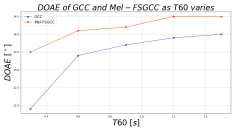
(d) Overall SELD performance as a function of noise level.

Figure 5: Comparison of GCC and Mel-FSGCC performance across different noise levels, ranging from -68 dB to -38 dB. The evaluation is conducted in a reverberant environment with T60 = 0.6 s and a microphone spacing of 1 m.

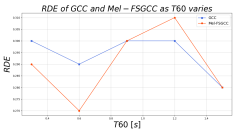
In the first experiment, the microphone array configuration and T60 were held constant while the noise level varied from -68 dB to -38 dB(Figure 5). The reverberation time was set to T60 = 0.6 s, with a microphone spacing of 1 m. Both methods exhibited comparable performance, with the F-score and RDE varying between them depending on the specific condition. In some cases, Mel-FSGCC outperformed GCC, while in others, GCC yielded better results. The DOAE remained slightly better for GCC across all noise levels, except for one specific condition. With an increase in the noise level, the performance of the Mel-FSGCC does not seem to improve. Overall, GCC demonstrated marginally greater stability and achieved a slightly better SELD score, but its improvement over Mel-FSGCC was not significant.



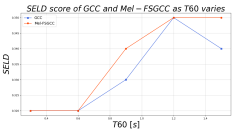
(a) F-score as a function of T60.



(b) DOAE as a function of T60.



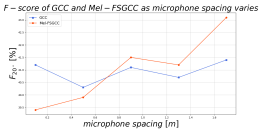
(c) RDE as a function of T60.



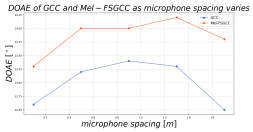
(d) Overall SELD performance as a function of T60.

Figure 6: Performance comparison of GCC and Mel-FSGCC across different reverberation times (T60), ranging from 0.3 s to 1.5 s. The evaluation is conducted in an environment with noise level = -43 dB and a microphone spacing of 1 m.

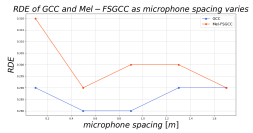
The second experiment maintained a fixed microphone distance and noise level while altering the T60 to examine the impact of reverberation (Figure 6). T60 ranged from 0.3 s to 1.5 s, with a noise level of -43 dB and a microphone spacing of 1 m. Similar to the experiment in which the noise level varied, both methods exhibited comparable performance, with fluctuations in the F-score and RDE depending on the specific condition. In some instances, Mel-FSGCC outperformed GCC, while in others, GCC performed better. The DOAE remained slightly more accurate for GCC across all conditions, with a maximum error variation of 1.5°. As the reverberation time increases, Mel-FSGCC performance does not show a significant enhancement. Overall, GCC achieved a marginally better SELD score, but the improvement over Mel-FSGCC was not substantial.



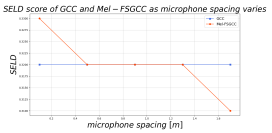
(a) F-score as a function of microphone spacing.



(b) DOAE as a function of microphone spacing.



(c) RDE as a function of microphone spacing.



(d) Overall SELD performance as a function of microphone spacing.

Figure 7: Performance comparison of GCC and Mel-FSGCC across different microphone spacings, ranging from 0.1 m to 1.7 m. The evaluation is conducted in an environment with noise level = -43 dB and T60 = 0.6 s.

In the third experiment, both T60 and noise level were kept constant while the distance between microphones was varied, allowing for an assessment of how spatial configuration influences localization accuracy (Figure 7). The microphone spacing varied from 0.1 m to 1.7 m, with a a noise level of -43 dB and a T60 = 0.6 s. Results show that increasing microphone spacing leads to a notable improvement in the F-score, with Mel-FSGCC generally outperforming GCC by a growing margin. This suggests that Mel-FSGCC benefits more from increased spacing, potentially enhancing detection accuracy. RDE also improved with spacing in Mel-FSGCC, eventually reaching a level comparable to GCC at the maximum tested distance, while DOAE remained slightly better for GCC across all spacing conditions. Additionally, at the maximum microphone spacing, Mel-FSGCC achieved a slightly better overall SELD score than GCC, indicating a possible advantage in larger-array configurations.

6. Conclusions

Mel-FSGCC does not offer a clear computational advantage, its effectiveness in sound event detection improves with increased microphone spacing, though its localization and distance estimation remain largely comparable to GCC. The mel-scale decomposition employed in Mel-FSGCC appears to enhance spectral feature extraction, particularly in wider array configurations, but its benefits are not consistent across all conditions, suggesting that its performance depends on specific spatial and acoustic factors. Theoretically, Mel-FSGCC could offer greater robustness in highly reverberant and noisy environments by emphasizing perceptually significant frequency bands, potentially leading to more stable time delay estimates and improved localization. However, further research is needed to explore its integration with deep learning-based SELD architectures and to validate its performance on larger, more diverse datasets. While Mel-FSGCC presents a promising alternative for feature extraction, its advantages over GCC remain context-dependent, requiring further investigation to determine its full potential in real-world applications.

References

[1] Maximo Cobos, Fabio Antonacci, Luca Comanducci, and Augusto Sarti. Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1270–1281, 2020.

[2] José Manuel Pérez-Lorenzo, Raquel Viciana-Abad, Pedro Jesús Reche-Lopez, Fernando Rivas, and José Escolano. Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments. *Applied Acoustics*, 73:698–712, 2012.

[3] S. S. Stevens and John E. Volkmann. The relation of pitch to frequency: A revised scale. *American Journal of Psychology*, 53:329, 1940.