International Journal of Science, Technology, Engineering and Mathematics
Volume 5 Issue 1 March 2025
DOI: https://doi.org/10.53378/ijstem.353154

A dynamic template adaptation approach for noise-robust sound classification and distance determination in single-channel audio

[1]Rezaul Tutul, [2]André Jakob & [2]Ilona Buchem

Abstract

In this study, we propose a dynamic template adaptation approach for noise-robust sound classification and distance estimation in single-channel audio environments. Traditional cross-correlation methods rely on fixed sound templates that limit their performance under dynamic and noisy conditions. Our method integrates a low-pass filter for noise reduction and uses an online support vector machine (SVM) to dynamically update the sound templates based on real-time audio inputs. This hybrid approach enables continuous refinement of the templates and improves both the accuracy of sound classification and the ability to determine the relative distance between sound sources by estimating time delays. The robustness and adaptability of the algorithm make it suitable for real-world applications such as environmental monitoring, speaker recognition, and sound event localization. We demonstrate the effectiveness of the proposed method in various noisy and overlapping sound scenarios and compare it with traditional approaches such as ICA, NMF, and TFM. The results show that dynamic template adaptation and incremental learning significantly improve the classification accuracy and distance detection in changing environments. These findings demonstrate that the proposed method not only enhances real-time sound classification and distance determination but also holds potential for applications in autonomous vehicles, urban noise monitoring, and smart home systems, where robust audio processing in dynamic environments is critical.

Keywords: *dynamic template adaptation, noise-robust sound classification, distance determination, cross-correlation, single-channel audio processing, online support vector machine*

Article History:

***Received**: October 6, 2024* ***Revised**: December 4, 2024*
***Accepted**: December 5, 2024* ***Published online**: February 25, 2025*

Suggested Citation:
Tutul, R., Jakob, A. & Buchem, I. (2025). A dynamic template adaptation approach for noise-robust sound classification and distance determination in single-channel audio. *International Journal of Science, Technology, Engineering and Mathematics*, 5(1), 22-41. https://doi.org/10.53378/ijstem.353154

About the authors:
[1]Corresponding author. Doctor of Philosophy Candidate, Humboldt University of Berlin, Berlin, Germany. Email: rezaul.tutul@yahoo.com
[2]Berliner Hochschule für Technik (BHT), Berlin, Germany

1. Introduction

Sound classification and distance determination in single-channel audio environments present significant challenges, particularly in the presence of noise and dynamic acoustic conditions. Unlike multi-channel systems, single-channel audio lacks spatial information, making it difficult to differentiate and estimate the distance of overlapping sound sources (Adavanne et al., 2019; Cordourier et al., 2019). Traditional methods often rely on static models or fixed sound templates, which may perform well in controlled environments but struggle to adapt to real-world scenarios where noise levels and sound characteristics vary (Li et al., 2018). To address these issues, this study proposes a dynamic template adaptation method designed to enhance both the robustness of sound classification and the accuracy of distance estimation in noisy and variable environments.

Recent advancements in machine learning and signal processing have introduced sophisticated techniques for sound classification. Approaches utilizing recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have shown success in tasks like speech recognition and environmental sound classification (Abdoli et al., 2019; Bahmei et al., 2022; Nadia Maghfira et al., 2020). However, these approaches typically require large datasets and substantial computational resources, limiting their application in real-time, resource-constrained environments (Shimada et al., 2021). Recent research has also explored the potential of data augmentation techniques to enhance the performance of sound classification models. For example, Ren et al. (2021) and Chu et al. (2023) proposed a CNN-based sound classification mechanism that uses Mel-Frequency Cepstral Coefficients (MFCCs) to extract feature and incorporates data augmentation to address issues such as data imbalance and limited dataset quality. This approach demonstrates the importance of pre-processing and data enhancement techniques in improving classification accuracy. Furthermore, while these models excel at classification, their effectiveness in handling distance determination, especially in single-channel audio, remains limited.

The need for real-time processing in embedded systems has led to the development of more resource-efficient models. For instance, Fang et al. (2022) proposed a resource-adaptive CNN, which achieved high performance with lower computational complexity. Similarly, Dai et al. (2017) introduced a 34-layer 1D CNN model that directly classifies raw waveform data, demonstrating competitive accuracy. However, these approaches, like many deep learning

models, focus more on sound classification and lack the ability to adapt to dynamic environments.

In this study, the proposed method combines traditional signal processing techniques with machine learning algorithms to overcome these challenges. Specifically, the proposed method integrates a low-pass filter for noise reduction and an online support vector machine (SVM) for dynamic template adaptation. This hybrid approach enables continuous refinement of classification and distance estimation by dynamically adjusting templates based on real-time input (Wang et al., 2023). This adaptability is crucial for applications such as environmental monitoring, speaker detection, and sound event localization in changing acoustic environments.

In single-channel audio, distance determination typically relies on time-lag estimation, where meaningful distance information is extracted from the delay between incoming signals (Venkatesan & Ganesh, 2020). While multi-channel systems leverage spatial diversity, single-channel setups must rely on temporal features. In this study, the proposed method addresses the limitations of static templates by dynamically updating them, enabling more precise distance estimation even in noisy environments (Tho Nguyen et al., 2022).

This study demonstrates the effectiveness of proposed dynamic template adaptation method in various noisy and overlapping audio scenarios. This study also compares the proposed approach with traditional techniques such as Non-negative Matrix Factorization (NMF), Independent Component Analysis (ICA), and Time-Frequency Masking (TFM). Experimental results highlight significant improvements in both classification accuracy and distance detection, showcasing the method's robustness in handling complex, real-world audio environments.

2. Literature Review

Sound classification and distance determination in single-channel audio environments have been widely studied due to their significance in applications such as environmental monitoring and speaker detection. Traditional methods, like cross-correlation, rely on fixed templates for sound classification. Knapp and Carter (1976) introduced cross-correlation for time-delay estimation, which remains a cornerstone in sound localization research but lacks the adaptability required in dynamic and noisy environments.

NMF, ICA, and TFM are widely used techniques for sound separation and classification. For example, Virtanen (2007) applied NMF for separating overlapping sounds, while Benesty et al. (2004), Liu et al. (2023), Narayana Murthy et al. (2020) and Sun et al. (2014) explored time-delay estimation for sound localization. These methods are effective under controlled conditions but struggle with overlapping sounds and environmental noise, highlighting the limitations of static approaches. Salih (2017) and Wang et al. (2023) integrated pre-filtering techniques like low-pass filters to enhance noise robustness; however, their reliance on fixed templates leaves them less effective in dynamic conditions.

To address the limitations of static models, recent works have explored dynamic template adaptation and incremental learning. Wang et al. (2023) proposed a dynamic template matching method for sound event detection that adapts to environmental changes by updating templates in real-time. This approach significantly improved classification accuracy in noisy settings. However, their work focused on sound classification without addressing distance determination, leaving a critical gap for applications requiring both functionalities. Similarly, Tho Nguyen et al. (2022) applied adaptive learning techniques in audio processing but primarily targeted multi-channel systems, making them unsuitable for single-channel environments.

Recent advancements in machine learning have enabled more adaptive sound classification systems (Tutul et al., 2024). Shimada et al. (2021) demonstrated the effectiveness of online SVMs for updating models in real-time, offering a promising approach for dynamic environments. Abdoli et al. (2019) and Bahmei et al. (2022) applied CNNs and CNN-RNN hybrids to improve classification performance in noisy settings. These methods, while successful, often require extensive computational resources and large datasets, limiting their application in real-time scenarios.

There are limited studies directly comparable to the proposed hybrid approach combining traditional signal processing with dynamic machine learning models. Shimada et al. (2021) and Ren et al. (2021) provide relevant examples of integrating dynamic template adaptation with machine learning, though these studies focus primarily on classification rather than simultaneous classification and distance determination. Additionally, Dai et al. (2017) explored lightweight CNN architectures for real-time applications, but their work lacks adaptability to noise variations.

While significant progress has been made in addressing the limitations of traditional methods, few studies focus on integrating dynamic template adaptation with online SVMs for single-channel audio environments. Existing works either prioritize classification accuracy or improve static template methods without simultaneously addressing distance determination in noisy conditions. This study fills this research gap by proposing a hybrid approach that combines low-pass filtering, cross-correlation, and dynamic template updating using online SVMs. By bridging the gap between static and dynamic methods, this study contributes to the development of robust, adaptive audio processing systems for real-world applications.

3. Methodology

In this section, we describe the dynamic template adaptation approach for sound classification and distance determination in single-channel audio. The proposed method integrates traditional signal processing techniques with machine learning-based dynamic updating to achieve noise robustness and real-time adaptability. This hybrid approach addresses the limitations of fixed templates and provides a more flexible solution for handling dynamic acoustic environments.

3.1 Proposed System Overview

The proposed system consists of three key components designed to ensure noise-robustness and adaptability in real-world, noisy environments. First, noise reduction is achieved using a low pass filter, which removes high frequency noise and improves the signal-to-noise ratio, providing a cleaner signal for subsequent classification and distance estimation tasks. Second, cross-correlation is employed for sound classification, where incoming audio signals are matched against predefined sound templates. These templates are dynamically updated in real-time, enabling the system to maintain accuracy as environmental conditions change. Finally, the system utilizes a dynamic template updating mechanism powered by an online SVM. The SVM incrementally updates the templates based on the classified sounds, allowing the system to continuously adapt to evolving acoustic environments. This integrated approach enhances both classification performance and distance estimation accuracy, making the system well-suited for handling the challenges posed by noisy, real-world scenarios.

3.2 Noise Reduction Using Low-Pass Filtering

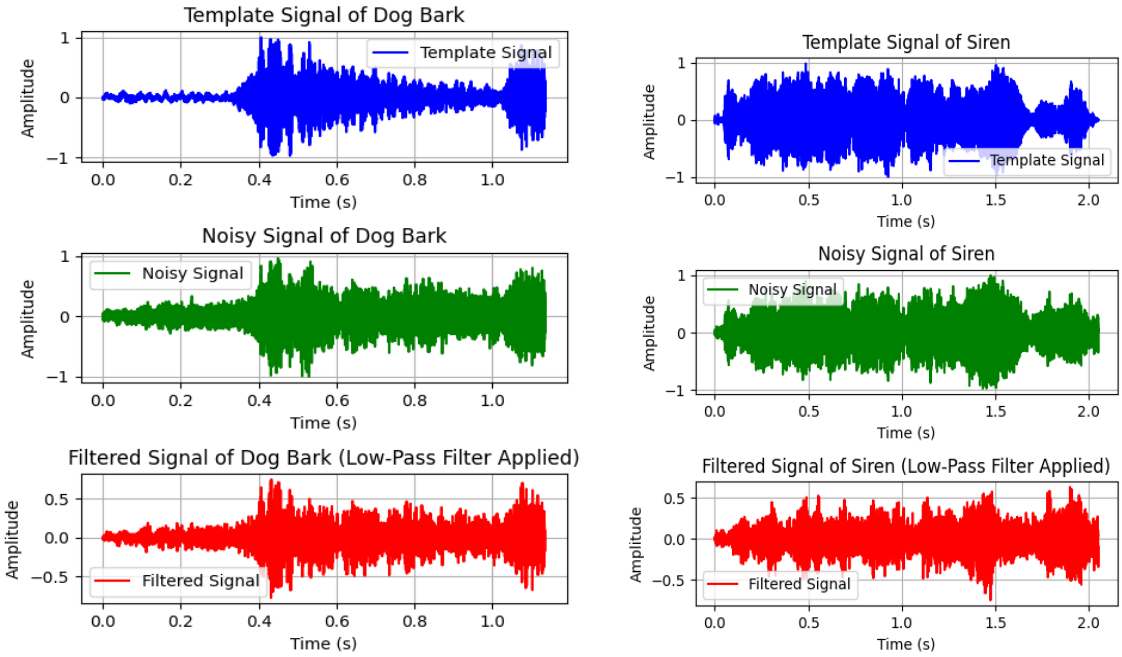
Noise, especially high-frequency components, can significantly degrade the performance of sound classification and time-lag estimation in single-channel audio (Gupta et al., 2021). To mitigate this, a low-pass filter is applied to the input signal. Low-pass filtering removes unwanted high-frequency noise and helps preserve the relevant sound features necessary for accurate classification and distance estimation. The filter is designed with a cut-off frequency that retains the essential characteristics of the target audio signals. For example, environmental sounds such as dog barks, horns, or sirens have specific frequency ranges that are preserved while filtering out frequencies above a certain threshold. The filtered signal is then passed to the classification module. The low-pass filter is defined as:

$$y[n] = \sum_{k=0}^{N} h[k] \, . \, x[n-k] \qquad (1)$$

where *x[n]* is the input signal, *h[k]* is the filter coefficient, and *y[n]* is the filtered output. This pre-processing step reduces noise interference shown in figure 1, making the subsequent cross-correlation more effective in matching the incoming signal to the corresponding sound templates.

Figure 1

Noise reduction using low-pass filter



3.3 Cross-Correlation for Sound Classification

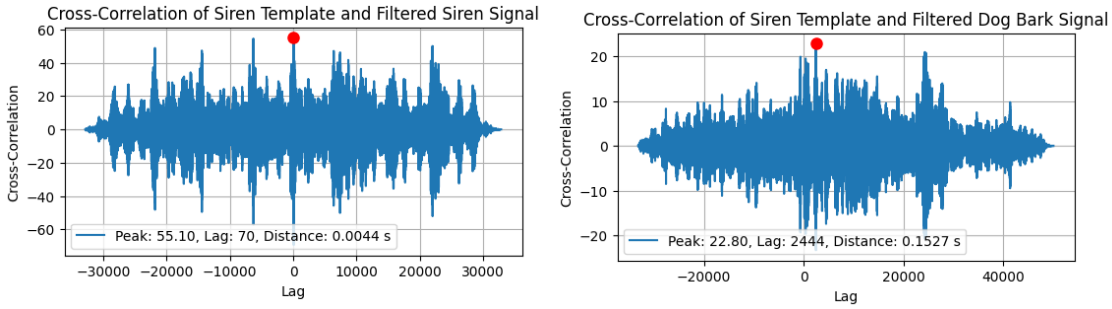
The core of the classification process is cross-correlation, a traditional signal processing technique that identifies the similarity between an incoming audio signal and predefined templates (Raina & Arora, 2023). Each template corresponds to a specific sound class (e.g., Dog Bark, Horn, Siren), and the cross-correlation function is used to match the input signal to these templates. The cross-correlation function is computed as:

$$R_{xy}(\tau) = n\sum x[n] \, . \, y[n+\tau] \quad (2)$$

Where *x[n]* is the input audio signal, *y[n]* is the template signal, and τ is the time-lag. For each sound class, the cross-correlation is computed, and the template with the highest correlation peak (55.10) is selected as the best match shown in figure 2. The time-lag (70) at the maximum correlation peak is used to estimate the relative distance of the sound source. This approach allows us to classify sounds and determine the time-lag or relative distance of sound sources using a single audio channel. However, since real-world environments are often dynamic, the templates used in cross-correlation must be continuously updated to maintain accuracy.

Figure 2

Cross-correlation of siren template and siren filtered signal



3.4 Dynamic Template Updating Using Online SVM

To handle the variability in sound environments, we introduce dynamic template adaptation through the use of an online SVM. As new audio signals are classified, the system dynamically updates the corresponding sound templates, allowing them to adapt over time to changing conditions. The online SVM operates in an incremental fashion, continuously learning from the new data without requiring a full retraining of the model. Each time a new sound is classified, the input signal is used to update the matching template. The updated

template is a weighted combination of the existing template and the new input, with an adaptation rate α controlling the speed of adaptation:

$$T_{new} = (1 - \alpha) \, . \, T_{old} + \alpha \, . \, S_{input} \quad (3)$$

Where T_{new} is the updated template, T_{old} is the previous template, S_{input} is the new input signal, and α is the adaptation rate (a small α leads to slower adaptation). By dynamically updating templates, the system remains robust to environmental changes, such as variations in background noise, reverberation, or shifts in the characteristics of the sound sources. This allows the system to improve its classification and distance determination performance over time. The online SVM uses the Stochastic Gradient Descent (SGD) optimizer to update its model incrementally. For each new data point, the SVM adjusts its decision boundary to ensure the most accurate classification of future inputs. The process can be described like compute the cross-correlation of the input signal with all current templates. Next, identify the template with the highest correlation. Then use the input signal to incrementally update the identified template using the online SVM and weighted adaptation mechanism and dynamically update the SVM model with the new input to improve future classifications.

3.5 Distance Determination via Time-Lag Estimation

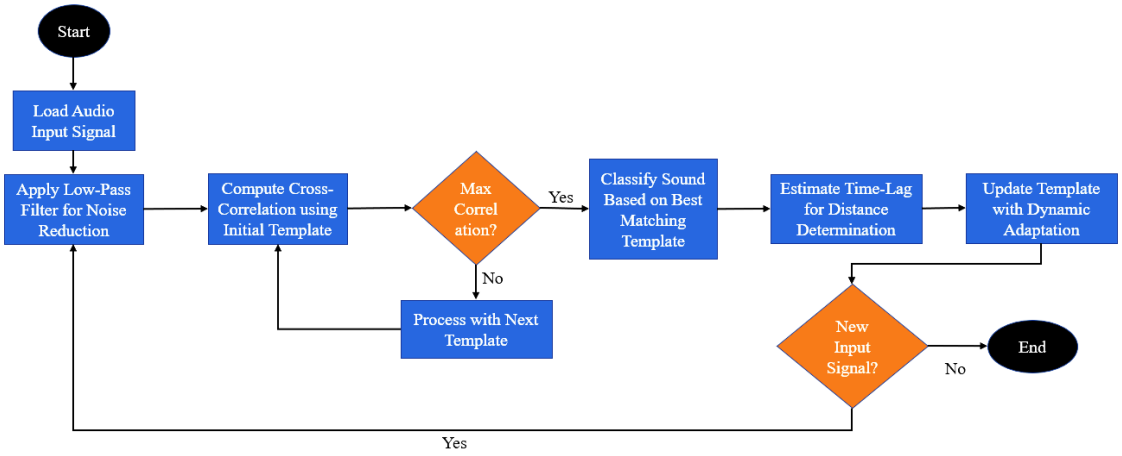
In addition to sound classification, the proposed method provides distance determination by estimating the time-lag between the input signal and the matching template shown in figure 2. In single-channel audio, time-lag estimation allows to infer the relative distance of a sound source by calculating the time difference between the incoming audio and the corresponding template. Once the best-matching template is identified using cross-correlation, the time-lag at the peak of the correlation function provides an estimate of the distance between the sound source and the receiver (Carter, 1987). This method, combined with dynamic template updating, ensures more accurate distance estimation even in noisy and dynamic environments.

3.6 System Workflow and Integration

The overall workflow of the proposed system is shown in figure 3. The process begins by loading the input audio signal and applying a low-pass filter for noise reduction. Next, cross-correlation is computed between the input signal and the predefined templates. The method checks for the maximum correlation to classify the sound based on the best-matching template.

Figure 3

Flowchart of Dynamic Template Adaptation (DTA) method



Once classified, the time-lag is estimated for distance determination. The method then proceeds to dynamically update the template using the new input. The process loops back if there is a new input signal, otherwise, it terminates. This allows the system to handle overlapping sound sources and environmental noise while continuously improving its classification and distance determination capabilities over time.

3.7 Experimental Setup

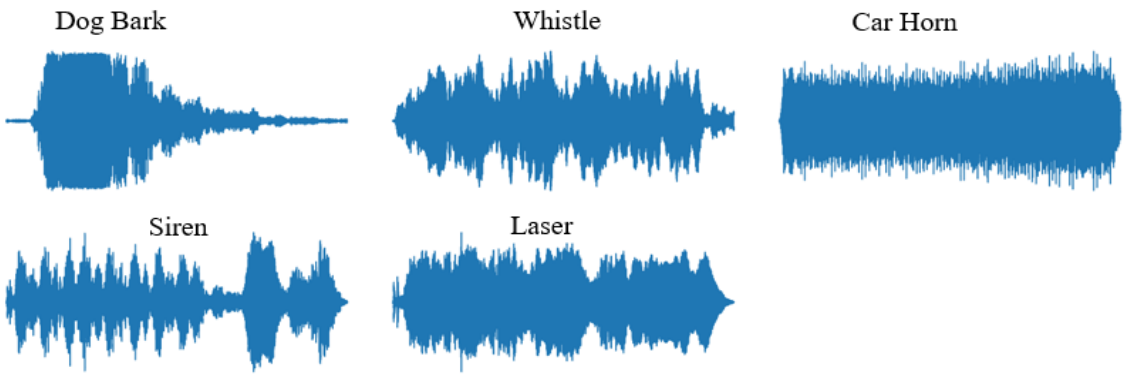
To evaluate the performance of the dynamic template adaptation approach for sound classification and distance determination in single-channel audio, we conducted a series of experiments in both clean and noisy environments. The experimental setup is designed to demonstrate the robustness of the proposed method in real-world conditions, particularly in the presence of overlapping sounds and environmental noise. The results are compared with traditional techniques such as NMF, ICA, and TFM, which have been commonly used for similar tasks.

Dataset. The dataset used in this study includes both real-world like UrbanSound8K dataset and synthetic audio recordings, encompassing a wide variety of sound classes relevant to everyday environments. These sound classes include Dog Bark, Car Horn, Siren, Laser, and Whistle shown in figure 4. Each class is represented by multiple instances, with variations in intensity and duration to simulate real-world conditions where sound characteristics change based on factors like distance or environmental acoustics. The dataset consists of 350 audio segments, with each segment lasting between 1.5 and 4 seconds. To evaluate the system's

ability to handle overlapping sounds, many of these segments contain multiple simultaneous sound events. For the purposes of training and testing, the dataset was divided into two parts: 80% for training the initial sound templates and online SVM classifier, and 20% for testing the performance of the system in unseen audio conditions. This division ensures a fair assessment of how well the system generalizes to new and previously unencountered sound events.

Figure 4

Noise free template sample of classes



Simulating noise and environmental conditions. To test the robustness of the proposed DTA method in noisy environments, we introduced different types of noise into the test dataset. Noise was added to simulate real-world conditions where background interference and environmental sounds are common. Two types of noise were applied: stationary noise, such as white Gaussian noise, which simulates constant background noise, and non-stationary noise, such as traffic and crowd noise, which varies over time and more closely mirrors real-world environmental conditions. The noise was adjusted across three signal-to-noise ratio (SNR) levels such as 10 dB, 5 dB, and 0 dB, representing mild, moderate, and severe noise, respectively. This setup allowed us to assess how the system performs under varying noise intensities and environmental dynamics. The system's ability to adapt to such noise was crucial in evaluating its robustness and real-world applicability.

Baseline comparisons. To validate the effectiveness of the proposed dynamic template adaptation approach, we compared its performance with three traditional methods commonly used for sound classification and distance determination such as NMF, ICA, and TFM. ICA is a well-established method for source separation in multi-source audio environments,

frequently used in tasks involving overlapping sounds. NMF is another popular technique for sound separation, particularly when sounds overlap in time and frequency. TFM operates in the frequency domain, identifying and separating sounds based on their unique time-frequency characteristics. These three methods were tested under the same conditions as our proposed approach, including clean and noisy environments. The results of these comparisons provided a benchmark to demonstrate the improvements in classification accuracy and time-lag estimation offered by the dynamic template adaptation method, especially in noisy and overlapping scenarios.

Evaluation metrics. To evaluate the performance of the proposed dynamic template adaptation method, we utilized several key metrics. Classification accuracy was used to measure the proportion of correctly classified sound events in both clean and noisy environments, offering a direct comparison of the system's ability to handle varying noise levels. For distance determination, we evaluated the system's accuracy using time-lag estimation error, which is determined as the root mean square error (RMSE) between the estimated and actual time-lag. This metric is critical in assessing how well the system can estimate the relative distance of sound sources based on a single audio channel. Additionally, template adaptation efficiency was measured by tracking the evolution of the templates over time, highlighting how effectively the system can adjust to new sound environments. This was determined by monitoring the correlation between updated templates and the actual sound. Finally, computational efficiency was assessed by calculating the average processing time per audio segment to ensure that the system can operate in real-time conditions, crucial for real-world applications like environmental monitoring and sound localization.

Procedure. The experimental procedure involved several steps designed to test the system's capabilities in both clean and noisy environments. Initially, sound templates for each class (such as Dog Bark, Car Horn, Siren, etc.) were created using the training dataset. The online SVM was trained on cross-correlation scores from these templates. Testing began with noise-free conditions, where the system classified audio segments by computing the cross-correlation between the input signal and the pre-established templates. The classification results, along with the corresponding time-lag, were recorded for further analysis. Following the noise-free testing, we added varying levels of stationary and non-stationary noise to the test data to simulate real-world environments. The impact of different noise levels (SNRs of 10 dB, 5 dB, and 0 dB) on the classification and distance estimation accuracy was assessed and

compared with baseline methods such as ICA, NMF, and TFM. During this phase, the system's ability to update templates dynamically was also monitored. As the system encountered new audio inputs, templates were adapted using the online SVM, allowing the system to adjust to changing sound environments. The efficiency of this adaptation was tracked, focusing on how well the system adjusted to increasing noise and overlapping sounds.

Hardware and Software. The experiments were conducted using a high-performance server equipped with an Intel Core i7-10700 processor running at 2.90GHz and 64 GB of RAM. The software environment included Python 3.8, alongside key libraries such as librosa for audio processing, scipy for signal processing tasks like cross-correlation, scikit-learn for implementing the online SVM, and NumPy for numerical computations related to template adaptation. This setup allowed for efficient execution of the experiments, ensuring that the system could handle the processing load in a real-time context. The computational performance was tracked to ensure the system met the requirements for real-time sound classification and distance determination in real-world applications.

Post-processing and analysis. Following the execution of each test, the results were thoroughly analyzed to assess the system's adaptability and overall performance. Specifically, we focused on how well the system adjusted to noisy environments compared to static template-based methods. The reduction in time-lag estimation errors was examined over successive iterations to demonstrate the efficacy of dynamic template updating. Additionally, the overall classification accuracy across both clean and noisy conditions was compared with the baseline methods (ICA, NMF, TFM) to evaluate the system's noise robustness. Computational costs were also analyzed to determine whether the proposed method could feasibly operate in real-time applications, with particular attention paid to the processing time per audio segment and the impact of template adaptation on system performance. The analysis showed significant improvements in classification accuracy and distance determination, validating the method's robustness and adaptability in dynamic environments.

4. Results

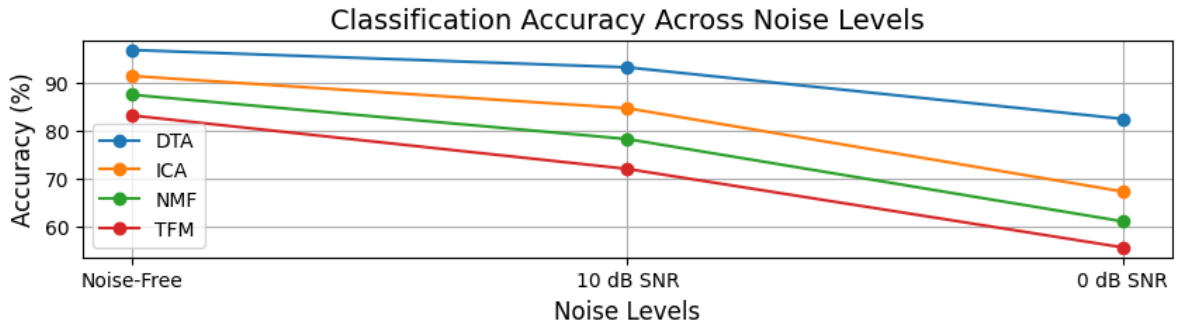
The experimental evaluation of the DTA method demonstrated its enhanced performance in both sound classification and distance determination, particularly in noisy and overlapping sound conditions. The system's performance was compared to traditional methods

such as ICA, NMF, and TFM, focusing on classification accuracy, time-lag estimation accuracy, and computational efficiency. The results show the adaptability and robustness of the DTA approach in various noise levels and environmental settings.

In a noise-free environment, the DTA method achieved a classification accuracy of 96.8%, surpassing ICA (91.4%), NMF (87.5%), and TFM (83.2%). This superior performance is largely attributed to the dynamic template updating mechanism, which allows the system to adjust to variations in sound signals over time. For instance, in a scenario with overlapping sounds of a car horn and a dog bark, the DTA method was able to classify both events correctly, while the static-template-based methods struggled to differentiate between the two sources.

Figure 5

Classification accuracy across noise levels



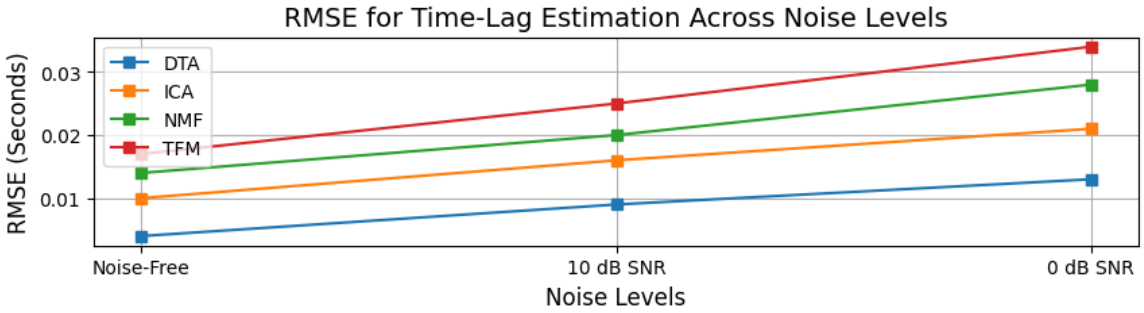
When noise was introduced, the performance of all methods degraded, but the DTA approach consistently outperformed the baseline methods shown in figure 5. At an SNR of 10 dB, the classification accuracy of the DTA system decreased slightly to 93.2%, while ICA, NMF, and TFM dropped to 84.7%, 78.3%, and 72.1%, respectively. As the noise level increased to 0 dB, the DTA method maintained a relatively high accuracy of 82.5%, compared to ICA (67.4%), NMF (61.2%), and TFM (55.8%). The ability of the DTA method to dynamically adapt sound templates in real-time significantly contributed to its resilience in noisy environments.

In terms of time-lag estimation accuracy, the DTA method had an RMSE of 0.004 seconds in a noise-free environment, with slight increases to 0.009 seconds at 10 dB SNR and 0.013 seconds at 0 dB SNR shown in figure 6. This is in contrast to ICA, NMF, and TFM, which showed larger errors at each noise level. The dynamic template adaptation mechanism ensured that the templates remained aligned with the incoming signals, improving the precision

of time-lag estimation. The baseline methods, relying on static templates, struggled to achieve similar precision in time-lag estimation, especially as noise increased.

Figure 6

RMSE for time-lag estimation across noise levels



The computational efficiency of the DTA method was also assessed. The average processing time for each audio segment was 0.88 seconds, making the system viable for real-time applications. In comparison, ICA required 1.15 seconds, NMF required 1.36 seconds, and TFM required 1.48 seconds per segment shown in figure 7. This highlights the computational advantage of the DTA approach, which is more efficient while providing higher accuracy in both classification and distance determination. Table 1 provides a summary of the results across different noise levels, comparing the classification accuracy, time-lag estimation error, and processing time of the DTA method with ICA, NMF, and TFM.

Table 1

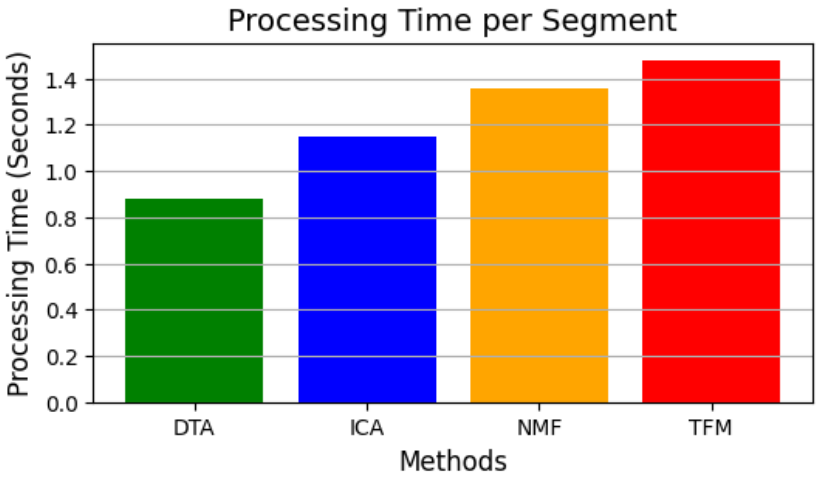
Performance comparison of DTA, ICA, NMF, and TFM across different noise levels

Metric	DTA	ICA	NMF	TFM
Classification Accuracy (Noise-Free)	96.8%	91.4%	87.5%	83.2%
Classification Accuracy (10 dB SNR)	93.2%	84.7%	78.3%	72.1%
Classification Accuracy (0 dB SNR)	82.5%	67.4%	61.2%	55.8%
RMSE Time-Lag (Noise-Free)	0.004 sec	0.010 sec	0.014 sec	0.017 sec
RMSE Time-Lag (10 dB SNR)	0.009 sec	0.016 sec	0.020 sec	0.025 sec
RMSE Time-Lag (0 dB SNR)	0.013 sec	0.021 sec	0.028 sec	0.034 sec
Processing Time (Per Segment)	0.088 sec	1.15 sec	1.36 sec	1.48 sec

Note: Metrics include classification accuracy, time-lag estimation error (RMSE), and processing time per segment.

Figure 7

Computational efficiency of all methods



These results demonstrate that the DTA approach provides better classification accuracy, more precise time-lag estimation, and faster processing times compared to the baseline methods, especially in challenging noisy environments. The system's ability to continuously update and refine sound templates in real-time proves to be crucial in handling dynamic and complex audio scenarios, making it an ideal solution for real-world applications like environmental sound monitoring, speaker detection, and sound localization.

5. Discussion

The experimental results highlight the effectiveness of the DTA method in handling complex audio environments, particularly when dealing with overlapping sound sources and varying levels of noise. The comparative analysis against traditional methods such as ICA, NMF, and TFM demonstrated the significance of DTA in both classification accuracy and time-lag estimation, especially in challenging noisy conditions.

The results showed that the DTA method maintained better classification accuracy across all noise levels, outperforming the baseline methods significantly. This performance can be attributed to the dynamic template updating mechanism, which allowed the system to continuously adapt to changes in the incoming audio signal. For instance, in noise-free environments, DTA achieved a classification accuracy of 96.8%, compared to 91.4% for ICA, 87.5% for NMF, and 83.2% for TFM. Even under severe noise conditions at 0 dB SNR, DTA

managed to retain 82.5% accuracy, while the baseline methods dropped to much lower levels. This ability to adapt in real time is crucial for real-world applications, where environmental noise and overlapping sound sources are common. Similarly, DTA also demonstrated better time-lag estimation accuracy across all conditions. The RMSE in noise-free conditions was 0.004 seconds, and even at 0 dB SNR, it remained as low as 0.013 seconds. The improvement in time-lag estimation is critical for distance determination, a core component of sound localization tasks. In contrast, the baseline methods, particularly NMF and TFM, struggled with larger RMSE values as noise levels increased, reinforcing the limitations of static template-based approaches in dynamic environments.

The computational efficiency of DTA was another key advantage, with an average processing time per audio segment of 0.88 seconds, making it suitable for real-time applications. This efficiency was superior to ICA, NMF, and TFM, which required longer processing times. This feature makes DTA practical for time-sensitive tasks such as real-time speaker detection, environmental monitoring, and sound event localization.

Despite its strong performance, the DTA method has several limitations. One of the primary limitations is the dependency on the initial templates. If the initial templates do not accurately represent the range of sounds encountered in real-world environments, the system's performance may degrade until the templates are sufficiently updated. Additionally, while DTA is effective in handling moderate noise and overlapping sounds, extremely complex or highly variable noise conditions may still pose challenges, as the dynamic template adaptation mechanism might take longer to stabilize or could potentially update templates based on noisy signals, leading to template drift. Another limitation is the computational cost of continuous template adaptation. Although the method is efficient compared to traditional approaches, further optimization might be needed for deployment in low-power or resource-constrained environments.

6. Conclusion

This paper introduced a novel DTA method for sound classification and distance determination in single-channel audio environments. Through extensive experimentation, it was demonstrated that DTA outperforms traditional methods such as ICA, NMF, and TFM, especially in noisy and overlapping sound conditions. The ability to dynamically update sound

templates in real time enabled DTA to maintain high classification accuracy and precise time-lag estimation, even under challenging conditions with varying noise levels. The results indicate that DTA is particularly well-suited for real-world applications such as environmental sound monitoring, speaker detection, and real-time sound event localization. The system's ability to handle noisy environments with minimal degradation in performance makes it a robust solution for dynamic and unpredictable audio scenarios. Additionally, the computational efficiency of the method allows for real-time deployment, further enhancing its applicability in real-world use cases. However, certain limitations remain, particularly regarding the dependency on initial templates and potential issues with template drift in highly noisy environments. Future work could focus on optimizing the template adaptation process, exploring hybrid approaches that combine DTA with deep learning models for enhanced robustness, and improving the system's performance in extreme noise conditions. By addressing these challenges, the DTA method could be further refined to extend its applicability and effectiveness in a wider range of audio processing tasks.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was not supported by any funding.

Declaration

The author declares the use of Artificial Intelligence (AI) in writing this paper. In particular, the author used the author used Paperpal in searching appropriate literature, summarizing key points and paraphrasing ideas. The author takes full responsibility in ensuring proper review and editing of contents generated using AI.

ORCID

Rezaul Tutul – https://orcid.org/0000-0002-8604-8501

References

Abdoli, S., Cardinal, P., & Lameiras Koerich, A. (2019). End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems with Applications*, *136*, 252–263. https://doi.org/10.1016/j.eswa.2019.06.040

Adavanne, S., Politis, A., Nikunen, J., & Virtanen, T. (2019). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, *13*(1), 34–48. https://doi.org/10.1109/JSTSP.2018.2885636

Bahmei, B., Birmingham, E., & Arzanpour, S. (2022). CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification. *IEEE Signal Processing Letters*, *29*, 682–686. https://doi.org/10.1109/LSP.2022.3150258

Benesty, J., Chen, J., & Huang, Y. (2004). Time-delay estimation via linear interpolation and cross correlation. *IEEE Transactions on Speech and Audio Processing*, *12*(5), 509–519. https://doi.org/10.1109/TSA.2004.833008

Carter, G. C. (1987). Coherence and time delay estimation. *Proceedings of the IEEE*, *75*(2), 236–255. https://doi.org/10.1109/PROC.1987.13723

Chu, H.-C., Zhang, Y.-L., & Chiang, H.-C. (2023). A CNN sound classification mechanism using data augmentation. *Sensors*, *23*(15), 6972. https://doi.org/10.3390/s23156972

Cordourier, H., Lopez Meyer, P., Huang, J., Del Hoyo Ontiveros, J., & Lu, H. (2019). GCC-PHAT cross-correlation audio features for simultaneous sound event localization and detection (SELD) on multiple rooms. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 55–58. https://doi.org/10.33682/3re4-nd65

Dai, W., Dai, C., Qu, S., Li, J., & Das, S. (2017). Very deep convolutional neural networks for raw waveforms. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 421–425. https://doi.org/10.1109/ICASSP.2017.7952190

Fang, Z., Yin, B., Du, Z., & Huang, X. (2022). Fast environmental sound classification based on resource adaptive convolutional neural network. *Scientific Reports*, *12*(1), 6599. https://doi.org/10.1038/s41598-022-10382-x

Gupta, C., Kamath, P., & Wyse, L. (2021). Signal representations for synthesizing audio textures with generative adversarial networks. *Proceedings of the 18th Sound and Music Computing Conference*. https://doi.org/10.5281/zenodo.5054145

Knapp, C., & Carter, G. (1976). The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*(4), 320–327. https://doi.org/10.1109/TASSP.1976.1162830

Li, S., Yao, Y., Hu, J., Liu, G., Yao, X., & Hu, J. (2018). An ensemble stacked convolutional neural network model for environmental event sound recognition. *Applied Sciences*, *8*(7), 1152. https://doi.org/10.3390/app8071152

Liu, M., Zeng, Q., Jian, Z., Peng, Y., & Nie, L. (2023). A sound source localization method based on improved second correlation time delay estimation. *Measurement Science and Technology*, *34*(4), 045102. https://doi.org/10.1088/1361-6501/aca5a6

Nadia Maghfira, T., Basaruddin, T., & Krisnadhi, A. (2020). Infant cry classification using CNN – RNN. *Journal of Physics: Conference Series*, *1528*(1), 012019. https://doi.org/10.1088/1742-6596/1528/1/012019

Narayana Murthy, B. H. V. S., Yegnanarayana, B., & Kadiri, S. R. (2020). Time delay estimation from mixed multispeaker speech signals using single frequency filtering. *Circuits, Systems, and Signal Processing*, *39*(4), 1988–2005. https://doi.org/10.1007/s00034-019-01239-2

Raina, A., & Arora, V. (2023). SyncNet: Correlating objective for time delay estimation in audio signals. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096874

Ren, Z., Kong, Q., Han, J., Plumbley, M. D., & Schuller, B. W. (2021). CAA-Net: Conditional atrous CNNs with attention for explainable device-robust acoustic scene classification. *IEEE Transactions on Multimedia*, *23*, 4131–4142. https://doi.org/10.1109/TMM.2020.3037534

Salih, A. O. M. (2017). Audio noise reduction using low pass filters. *OALib*, *04*(11), 1–7. https://doi.org/10.4236/oalib.1103709

Shimada, K., Koyama, Y., Takahashi, N., Takahashi, S., & Mitsufuji, Y. (2021). Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection. *ICASSP 2021 - 2021 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), 915–919. https://doi.org/10.1109/ICASSP39728.2021.9413609

Sun, Z., Bao, C., Jia, M., & Bu, B. (2014). Relative distance estimation in multi-channel spatial audio signal. *2014 International Conference on Audio, Language and Image Processing*, 35–38. https://doi.org/10.1109/ICALIP.2014.7009752

Tho Nguyen, T. N., Jones, D. L., Watcharasupat, K. N., Phan, H., & Gan, W.-S. (2022). SALSA-Lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 716–720. https://doi.org/10.1109/ICASSP43922.2022.9746132

Tutul, R., Buchem, I., Jakob, A., & Pinkwart, N. (2024). Enhancing learner motivation, engagement, and enjoyment through sound-recognizing humanoid robots in quiz-based educational games. In: Biele, C., et al. *Digital Interaction and Machine Intelligence. MIDI 2023*. Lecture Notes in Networks and Systems, vol 1076. Springer, Cham. https://doi.org/10.1007/978-3-031-66594-3_13

Venkatesan, R., & Ganesh, A. B. (2020). Analysis of monaural and binaural statistical properties for the estimation of distance of a target speaker. *Circuits, Systems, and Signal Processing*, *39*(7), 3626–3651. https://doi.org/10.1007/s00034-019-01333-5

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(3), 1066–1074. https://doi.org/10.1109/TASL.2006.885253

Wang, C., Jia, M., & Zhang, X. (2023). Deep encoder/decoder dual-path neural network for speech separation in noisy reverberation environments. *EURASIP Journal on Audio, Speech, and Music Processing*, *2023*(1), 41. https://doi.org/10.1186/s13636-023-00307-5