Quarterly Journal Defensive Future Studies
IRI Military Command and Staff University
www.dfsr.ir
Online ISSN: 2645-7172
Print ISSN: 2588-428x

Three-Dimensional UAV Sound Source Localization Using Deep Audio Feature Learning

Fatemeh Alimadadi[1] ✉

1. Master's student in Artificial Intelligence & Robotics, Department of Computer Engineering, Malek ashtar University, Tehran, Iran. (Corresponding author). E-mail: fatemalimadadi@mut.ac.ir

Article Info

Article type:
Research Article

Article history:
Received:
2025-9-2
Received in revised form:
2026-1-8
Accepted:
2026-1-24
Published online:
2026-2-20

Keywords:
UAV Sound Source Localization, Deep Learning, Multi Channel Audio, Defense Systems.

ABSTRACT

Objective: Acoustic localization of unmanned aerial vehicles (UAVs) plays a critical role in military and surveillance applications, as it enables the detection and tracking of hostile drones under real-world conditions. This study proposes a deep learning-based framework that leverages joint time-frequency features for three-dimensional UAV localization.

Methodology: A publicly available benchmark dataset consisting of multichannel UAV flight recordings was employed for training and evaluation. Spectral representations were extracted using Mel-spectrograms and subsequently analyzed through an integrated time-frequency processing scheme based on the Mamba architecture, allowing accurate estimation of spatial parameters including range, altitude, azimuth and elevation angles.

Findings: Experimental results demonstrate that the proposed model achieves precise estimation of UAV spatial parameters and maintains robust performance under noisy conditions and across varying microphone distances.

Conclusion: The proposed approach, leveraging deep learning and multichannel audio data, can serve as an effective tool for defense and surveillance systems in the acoustic detection and tracking of UAVs.

Cite this article: Alimadadi,F. (2026). Three-Dimensional UAV Sound Source Localization Using Deep Audio Feature Learning. (e735729). *Defensive Future Studies*, 11 (40),79-114.

DOI: https://doi.org/10.22034/dfsr.2026.2070645.1943

CC BY NC ND

Publisher: IRI Military Command and Staff University

Extended Abstract

INTRODUCTION

Drones have rapidly emerged as critical tools in military and security operations, playing key roles in reconnaissance, surveillance, and offensive missions over sensitive airspaces and urban infrastructure (1). The threat posed by unauthorized or hostile drone flights has become a major national security concern. Accurate three-dimensional localization of drones is essential, and acoustic source localization—a technique that estimates the position of one or more sound sources relative to a reference, typically a microphone—offers significant advantages (2). Inspired by human auditory perception, which can simultaneously estimate direction and distance of sounds (3), acoustic systems operate effectively in low-visibility or covered conditions and with low-cost sensors. To address gaps in prior methods, this study proposes a deep Mamba-based network (4) that extracts both temporal and spectral features from Mel-spectrograms to precisely estimate drone distance, elevation, and angles in complex environments.

METHODOLOGY

We propose a deep architecture for 3D localization of drone acoustic sources using multichannel audio signals. The model first extracts log-Mel spectrograms from each of the eight microphone channels, capturing perceptually relevant spectral energy distributions (5). Spectrograms are divided into overlapping 2D temporal and frequency patches, which are processed using 2D convolutional layers and Mamba blocks. Each Mamba block integrates convolution and a state-space model to capture both local and long-range dependencies along time and frequency axes. Temporal and spectral features are fused via cross-attention to generate a unified representation, which is then used to estimate drone distance, elevation, azimuth, and elevation angle. The model is trained on the publicly available multi-channel drone dataset (3) using AdamW optimizer with CosineAnnealingLR. Performance is evaluated with mean absolute error and root mean squared error, providing robust 3D localization suitable for real-time applications on GPU and edge devices.

RESULT

Table 1 summarizes the evaluation results for four key parameters: a) distance (meters), b) elevation (meters), c) azimuth (degrees), and d) elevation angle (degrees). In the baseline model from the original dataset (3), the best performance was achieved using log-Mel spectrogram features with a sampling rate of 44.1 kHz, FFT length of 1024, hop size of 512, and 128 Mel filters. In contrast, our proposed model uses higher-resolution log-Mel spectrograms with a 96 kHz sampling rate, FFT length of 4096, hop size of 1024, and 256 Mel filters.

Table (1) Comparison of performance evaluation of the base model with the model proposed in this paper with MAE and RMSE

model	MAE				RMSE			
	a	b	c	d	a	b	c	d
Baseline Model	0/39	0/41	0/91	8/61	0/58	0/55	1/36	26/88
Proposed Model	0/05	0/06	0/14	8/51	0/19	0/47	0/62	26/7

The increased sampling rate and finer Mel filter resolution improved feature representation, enhancing temporal and spectral discrimination. For distance (a), the baseline model achieved MAE: 0/39, RMSE: 0/58, while our model achieved MAE: 0/051, RMSE: 0/192, representing reductions of ~87% in MAE and ~67% in RMSE. Elevation (b) showed MAE/RMSE reductions of 83% and 13%, respectively. Azimuth (c) improved from MAE 0/91, RMSE 1/36 to MAE 0/143, RMSE 0/628, indicating more precise horizontal angle estimation. Elevation angle (d) showed minimal improvement (MAE 8/514, RMSE 26/708), highlighting inherent challenges in vertical angle estimation due to the limited vertical baseline of mostly planar microphone arrays and environmental reflections.

As shown in Figure 1, the changes in the loss function over the training epochs are illustrated. This plot demonstrates the overall decrease in loss despite minor fluctuations between epochs, clearly reflecting the learning progress of the proposed model.

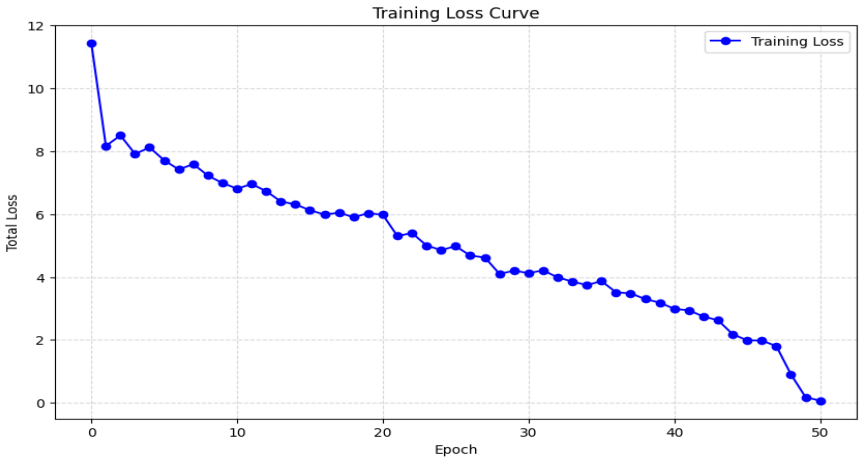


Figure 1: Loss curve during the training process of the model.

These results indicate that the proposed model achieves highly accurate distance and horizontal angle estimation, while vertical angle estimation remains limited by geometric constraints. The combination of temporal and spectral feature extraction, Mamba blocks, and cross-attention allows robust 3D localization even in noisy environments. From a practical perspective, the improved precision in distance and azimuth enables real-time drone tracking and early warning in military or surveillance contexts. Acoustic-based localization can thus serve as a complementary or alternative system in electronic warfare scenarios where radar may be compromised, enhancing situational awareness and defense capabilities.

DISCUSSION and CONCLUSIONS

This study demonstrates that the proposed Mamba-based architecture achieves superior performance in 3D drone localization compared to baseline methods. Significant reductions in distance and azimuth estimation errors indicate that the model can accurately and in real-time track drones, even under noisy conditions. Unlike previous studies focusing primarily on 2D localization (azimuth and elevation), this work provides a comprehensive 3D solution encompassing distance, height, azimuth, and elevation, addressing a key gap in the literature.

The main innovation is the simultaneous extraction and integration of temporal and spectral features within the Mamba architecture. This approach emulates human auditory perception while maintaining lower computational complexity and greater generalizability than conventional transformer-based models. Results show improved accuracy in distance and angle estimations,

demonstrating the model's potential for deployment in operational defense applications.

Nonetheless, some limitations remain. Most training data were collected under controlled conditions, which may limit performance in real-world urban or battlefield environments with complex noise, reflections, or varying weather. Experiments were also conducted on a limited set of drones; evaluation across drones with diverse acoustic signatures is needed to ensure robust generalization.

Future work should include: (1) testing on more diverse datasets representing urban and military scenarios; (2) applying data augmentation to improve robustness against dynamic noise; (3) integrating acoustic sensors with low-power radar, optical, or thermal cameras for multi-modal tracking; and (4) optimizing the model for lightweight, portable deployment.

The findings have practical significance for defense systems. Accurate 3D acoustic localization can complement radar and optical sensors, particularly in electronic warfare or low-visibility conditions, enabling rapid alerts, autonomous drone tracking, and enhanced intelligent air defense.

REFERENCES

1. Jekateryńczuk, G. & Piotrowski, Z. (2023). A survey of sound source localization and detection methods and their applications. *Sensors*, *24*(1), 68. DOI: https://doi.org/10.3390/s24010068

2. Khan, A. Waqar, A. Kim, B. & Park, D. (2025). A review on recent advances in sound source localization techniques, challenges, and applications. *Sensors and Actuators Reports*, 100313. DOI: https://doi.org/10.1016/j.snr.2025.100313

3. Jekateryńczuk, G. Szadkowski, R. & Piotrowski, Z. (2025). UaVirBASE: A Public-Access Unmanned Aerial Vehicle Sound Source Localization Dataset. *Applied Sciences*, *15*(10), 5378. DOI: https://doi.org/10.3390/app15105378

4. Gu, A. & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*. DOI: https://doi.org/10.48550/arXiv.2312.00752

5. Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366. DOI: https://doi.org/10.1109/TASSP.1980.1163420

فصلنامه علمی آینده‌پژوهی دفاعی
شاپای چاپی: ۲۲۸۸-۲۲۸X شاپای الکترونیکی: ۲۶۴۵-۷۱۷۲ http://www.dfsr.ir/
دانشگاه فرماندهی و ستاد آجا

مکان‌یابی سه‌بعدی پهپاد با یادگیری عمیق ویژگی‌های صوتی

فاطمه علی مددی[۱] ✉

۱. کارشناسی ارشد هوش مصنوعی و رباتیک، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی مالک اشتر، تهران، ایران. (نویسنده مسئول)، رایانامه: fatemalimadadi@mut.ac.ir

اطلاعات مقاله

نوع مقاله:
مقالهٔ پژوهشی

تاریخچه مقاله:

تاریخ دریافت:
۱۴۰۴/۰۶/۱۱

تاریخ بازنگری:
۱۴۰۴/۱۰/۱۸

تاریخ پذیرش:
۱۴۰۴/۱۱/۰۴

تاریخ انتشار:
۱۴۰۴/۱۲/۰۱

کلیدواژه‌ها:
مکان‌یابی سه‌بعدی پهپاد، یادگیری عمیق، صوت چند کاناله، سامانه‌های دفاعی.

چکیده

هدف: مکان‌یابی صوتی پهپادها در کاربردهای نظامی و نظارتی اهمیت بالایی دارد، زیرا امکان شناسایی و ردیابی پهپادهای متخاصم را در محیط‌های واقعی فراهم می‌سازد. در این پژوهش، یک رویکرد یادگیری عمیق مبتنی بر ویژگی‌های زمانی-طیفی برای تخمین موقعیت سه‌بعدی پهپاد ارائه‌شده است.

روش: برای آموزش و ارزیابی مدل، از یک پایگاه‌داده‌ی استاندارد موجود شامل داده‌های صوتی چندکاناله‌ی پرواز پهپادها استفاده گردید. ویژگی‌های طیفی با استفاده از طیف‌نگار مل استخراج و سپس با بهره‌گیری از ترکیب‌پردازش ویژگی‌های زمانی و فرکانسی مبتنی بر معماری Mamba مورد تحلیل قرار گرفتند. تا پارامترهای مکانی شامل فاصله، ارتفاع، زاویه‌ی افقی و زاویه‌ی عمودی تخمین زده شوند.

یافته‌ها: نتایج آزمایش‌ها نشان داد که مدل پیشنهادی قادر به تخمین دقیق پارامترهای مکانی پهپاد بوده و عملکرد مطلوبی در شرایط نویزی و فاصله‌های مختلف از میکروفون ارائه می‌دهد.

نتیجه‌گیری: رویکرد پیشنهادی با تکیه بر یادگیری عمیق و داده‌های صوتی چندکاناله می‌تواند به‌عنوان ابزاری کارآمد برای سامانه‌های دفاعی و نظارتی در تشخیص و رهگیری صوتی پهپادها مورد استفاده قرار گیرد.

استناد: علی مددی، فاطمه. (۱۴۰۵). مکان‌یابی سه‌بعدی پهپاد با یادگیری عمیق ویژگی‌های صوتی. مجله، آینده‌پژوهی دفاعی، ۱۱ (۴۰)، ۷۷۰-۱۱۴.

DOI: https://doi.org/10.22034/dfsr.2026.2070645.1943

ناشر: دانشگاه فرماندهی و ستاد ارتش جمهوری اسلامی ایران

CC BY NC ND

مقدمه

پهپادها به‌عنوان یکی از مهم‌ترین ابزارهای نظامی و امنیتی در سال‌های اخیر به‌سرعت گسترش‌یافته‌اند و نقش تعیین‌کننده‌ای در عملیات شناسایی، نظارتی و تهاجمی برای حریم‌های هوایی نظامی و زیرساخت‌های حیاتی شهری ایفا می‌کنند (Jekateryńczuk & Piotrowski, ۲۰۲۳). درعین‌حال، تهدیدات ناشی از پرواز پهپادهای ناشناس یا متخاصم بر فراز مناطق حساس نظامی و شهری، به چالش جدی برای امنیت ملی بسیاری از کشورها تبدیل شده است. همچنین، گسترش روزافزون پهپادهای جاسوسی و تهاجمی بدون سرنشین توسط کشورهای فرا منطقه‌ای که دامنه تهدیدات را علیه منافع ملی ج.ا. ایران وسیع‌تر نموده‌اند، ایجاب می‌کند که همه نیروهای دفاعی کشور برای مقابله با آن‌ها به بازبینی و بهینه‌سازی فرآیندهای عملیاتی خود در حوزه پهپادها نیز بپردازد (حبیبی، ۱۳۹۶).

یکی از مؤلفه‌های کلیدی در مقابله با این تهدیدات، توانایی دقیق و سریع در تعیین موقعیت سه‌بعدی پهپادها است. مکان‌یابی منبع صدا تکنیکی است که شامل تعیین مکان یک یا چند منبع صدا نسبت به یک مرجع انتخاب‌شده، معمولاً مکان میکروفون، با تجزیه‌وتحلیل سیگنال‌های صوتی دریافتی از منبع است (Khan, Waqar, Kim & Park, ۲۰۲۵). انسان‌ها به‌طور طبیعی از توانایی چشمگیری در مکان‌یابی منابع صوتی[1] برخوردارند و می‌توانند هم‌زمان جهت و فاصله‌ی منبع صدا را با دقت قابل‌توجهی تشخیص دهند. این قابلیت ذاتی نقش مهمی در درک صحنه‌های پویا و پیچیده دارد، چرا که به مغز امکان می‌دهد اطلاعات شنیداری را با داده‌های دیداری و دیگر حواس ادغام کرده و تصویری جامع از محیط پیرامون بسازد (Jekateryńczuk, Szadkowski & Piotrowski, ۲۰۲۵). الهام از این سازوکار زیستی، زمینه‌ساز شکل‌گیری و رشد حوزه‌ی پژوهشی مکان‌یابی منبع صدا شده است؛ حوزه‌ای کلیدی در پردازش سیگنال‌های صوتی که به دنبال طراحی الگوریتم‌ها و سامانه‌هایی است که بتوانند توانایی‌های ادراکی انسان در مکان‌یابی صدا را شبیه‌سازی و حتی در برخی شرایط ارتقا دهند.

1- Sound Source Localization

استفاده از سامانه‌های صوتی برای مکان‌یابی پهپادها مزایای متعددی دارد. این سامانه‌ها مستقل از شرایط آب‌وهوایی و محدودیت‌های دیداری عمل می‌کنند و می‌توانند حتی در شب یا در شرایط پوششی مانند دود، مه و گردوغبار نیز کارایی بالایی داشته باشند. هرچند سامانه‌های مبتنی بر روش‌های اپتیکی و بینایی ماشین در شرایط نوری مناسب و میدان دید باز از دقت بالایی در شناسایی و ردیابی پهپادها برخوردارند، اما عملکرد آن‌ها به‌شدت به نور، خط دید مستقیم و شرایط محیطی وابسته است و در سناریوهای عملیاتی پیچیده با افت کارایی مواجه می‌شود. از سوی دیگر، سامانه‌های راداری اگرچه توانایی کشف اهداف در بردهای بلند و در شرایط نامساعد جوی را دارند، اما در مواجهه با پهپادهای کوچک، کم‌ارتفاع و دارای سطح مقطع راداری پایین با چالش‌های جدی روبه‌رو بوده و علاوه بر هزینه‌ی بالا، نیازمند زیرساخت پیچیده و مصرف انرژی قابل‌توجه هستند (چالاکی، ۱۴۰۳). در مقابل، سامانه‌های صوتی با اتکا به انتشار امواج صوتی پهپاد و بدون نیاز به دید مستقیم یا تجهیزات پرهزینه، محدودیت‌های رایج روش‌های اپتیکی و راداری را تا حد زیادی برطرف کرده و به دلیل استفاده از حسگرهای صوتی کم‌هزینه و نیاز به زیرساخت محدود، امکان پیاده‌سازی گسترده و سریع در محیط‌های شهری و نظامی را فراهم می‌کنند.

در روش‌های کلاسیک مکان‌یابی منبع صدا، اغلب به تخمین جهت رسیدن منابع[1]، به‌ویژه زاویه و ارتفاع، ساده می‌شود، درحالی‌که تخمین فاصله تا آرایه میکروفون را حذف می‌کند (Luo, Lu, Huang, Ran & He, ۲۰۲۳)؛ و یا در روش‌های شکل‌دهی پرتو[2]، انرژی سیگنال را در جهات مختلف محاسبه می‌کنند و تنها جهت منبع را تخمین می‌زنند (Song et al, ۲۰۲۵) و یا روش‌های مبتنی بر همبستگی متقابل تعمیم‌یافته با تبدیل فاز[3] نیز تنها زاویه منبع را تخمین می‌زنند و قادر به تخمین فاصله نیست (Ma et al, ۲۰۲۵).

این روش‌ها عموماً به مکان‌یابی دوبعدی یعنی تخمین زاویه افقی[4] و زاویه ارتفاع[5] منبع می‌پردازند، بااین‌حال، یک محدودیت اساسی در این روش‌ها وجود دارد که تخمین فاصله منبع تا آرایه میکروفون برای مکان‌یابی سه‌بعدی است. دلیل این امر، حساسیت بالای

1- Direction Of Arrival (DOA)
2- Beamforming
3 -Generalized Cross-Correlation with Phase Transform (GCC-PHAT)
4- Azimuth
5- Elevation

تخمین فاصله به نویز محیطی، بازتاب‌ها و تغییرات شدت صوت در فضا است. اهمیت این موضوع به‌ویژه در زمینه‌ی مکان‌یابی صوتی پهپادها دوچندان است، زیرا دانستن صرفاً جهت منبع برای رهگیری یا خنثی‌سازی پهپاد کافی نیست و تخمین دقیق فاصله نیز برای برآورد موقعیت سه‌بعدی و پیش‌بینی مسیر حرکت آن ضروری است؛ بنابراین، گذار از روش‌های کلاسیک مبتنی بر تخمین زاویه به سمت روش‌های نوین سه‌بعدی، یکی از شکاف‌های اصلی پژوهشی در حوزه‌ی مکان‌یابی منبع صدای پهپاد به شمار می‌رود.

با وجود پیشرفت‌های اخیر، هنوز چالش‌های قابل توجهی در زمینه توسعه الگوریتم‌های دقیق، سریع و مقاوم به نویز برای مکان‌یابی سه‌بعدی پهپادها وجود دارد، به‌ویژه در محیط‌هایی با نویز پس‌زمینه قابل توجه یا منابع صوتی غیر ثابت و محیط‌های شهری و نظامی با پیچیدگی‌های آکوستیکی بالا. در نتیجه، توسعه سیستم‌های دقیق و مقاوم به نویز به مجموعه داده‌های بزرگ و با کیفیت بالا و همچنین منابع محاسباتی قابل توجه نیاز دارد تا دقت و قابلیت اطمینان سیستم حفظ شود.

در پاسخ به شکاف‌های عدم وجود الگوریتم‌های مقاوم و دقیق برای مکان‌یابی سه‌بعدی پهپادها در شرایط نویزی، عدم بهره‌گیری همزمان از ویژگی‌های زمانی و طیفی و محدودیت کارایی روش‌های پیشین در شرایط واقعی، در این مقاله یک مدل شبکه عمیق مبتنی بر Mamba (Gu & Dao, ۲۰۲۳) طراحی و ارائه شده است که به‌طور ویژه برای تخمین دقیق فاصله تا منبع صدا، ارتفاع پهپاد، زاویه افقی و زاویه ارتفاع از روی داده‌های صوتی توسعه یافته است. در این مدل، سیگنال‌های صوتی ابتدا به Mel-Spectrogram به‌عنوان ورودی مدل تبدیل شده و سپس دو مسیر پردازش موازی برای استخراج ویژگی تعریف می‌شود. یک مسیر با تمرکز بر ویژگی‌های زمانی، اختلاف زمانی رسیدن[1] سیگنال‌ها به میکروفن‌ها را استخراج می‌کند و از این طریق اطلاعات دقیقی درباره فاصله و جهت منبع صوتی فراهم می‌سازد. مسیر دیگر بر ویژگی‌های طیفی متمرکز است و با تحلیل کاهش انرژی طیفی[2]، الگوهای مرتبط با تداخل محیطی و ویژگی‌های آکوستیکی منبع صدا را مدل‌سازی می‌کند. در گام بعد، این دو دسته ویژگی با یکدیگر ادغام می‌شوند تا نمایش یکپارچه‌ای از اطلاعات زمانی و طیفی به‌دست آید. این معماری، علاوه بر شبیه‌سازی توانایی‌های ادراکی انسان در

1 -Temporal Difference of Arrival (TDoA)
2 -Spectral Attenuation

شنود و مکان‌یابی، می‌تواند در محیط‌های پر نویز و پیچیده نیز عملکرد قابل‌اعتمادی داشته باشد.

برای آموزش و ارزیابی مدل شبکه عمیق طراحی‌شده، از مجموعه داده مقاله (Jekateryńczuk, Szadkowski & Piotrowski, ۲۰۲۵) بهره می‌بریم که شامل ضبط‌های چند میکروفونی همزمان است که تحت شرایط کنترل‌شده ضبط شده‌اند و شامل تغییرات در فواصل، ارتفاعات، زوایا و جهت‌های پهپاد نسبت به یک آرایه میکروفون ثابت هستند. جهت‌های پهپاد شامل پیکربندی‌های رو به جلو، عقب، چپ و راست است. این مقاله شکافی را که توسط پایگاه‌های داده موجود که اغلب فاقد چنین تغییرات خاصی هستند، پر می‌کند.

مرور پیشینه و مبانی نظری

پیشینه پژوهش

اکثر پیاده‌سازی‌های سخت‌افزاری فعلی روش‌های مکان‌یابی منبع صوتی با آرایه‌های میکروفونی انجام می‌شوند (Chung, Chou & Lin, ۲۰۲۲). در این سامانه‌ها، چندین میکروفون به‌صورت هندسی در یک آرایه قرار می‌گیرند تا با بهره‌گیری از اختلاف‌های زمانی یا فازی بین سیگنال‌های دریافتی، جهت رسیدن منبع صوتی تخمین زده شود. روش‌های کلاسیک متداول در زمینه مکان‌یابی منبع صوتی، مانند اختلاف زمانی رسیدن و زاویه رسیدن سیگنال، معمولاً با استفاده از آرایه میکروفون‌ها انجام می‌شوند و بر اساس اصول عملکردشان به سه دسته اصلی تقسیم می‌شوند: روش‌های مبتنی بر تأخیر زمانی، روش‌های مبتنی بر شکل‌دهی پرتو و روش‌های با تفکیک‌پذیری بالا در حوزه طیف[1]. روش‌های کلاسیک بر مدل‌های فیزیکی طراحی‌شده دستی و فرضیات ایده‌آل متکی هستند که اغلب در سناریوهای پویا و واقعی ناکارآمد می‌شوند. چالش‌های خاص شامل ناهمگونی مکانی-زمانی سرعت صوت در محیط، بازتاب زیاد همراه با نسبت سیگنال به نویز[2] پایین در فضاهای داخلی و ویژگی‌های نویز غیرایستا در پهپادها است. این عوامل می‌توانند دقت روش‌های کلاسیک مکان‌یابی صوتی را به‌طور قابل توجهی کاهش دهند یا حتی منجر به شکست کامل شوند (Jekateryńczuk & Piotrowski, ۲۰۲۳).

1- High-Resolution Spectral-based
2- Signal-to-Noise Ratio (SNR)

روش‌های مبتنی بر تأخیر زمانی (Wang & Zhang, ۲۰۲۴)، بر اساس اندازه‌گیری اختلاف زمان رسیدن سیگنال صوتی به میکروفن‌های قرار گرفته در موقعیت‌های مختلف عمل می‌کند. با استفاده از این اختلاف زمان و سرعت صوت در محیط، می‌توان فاصله منبع صدا را از هر میکروفن تخمین زد. به‌کارگیری روش اختلاف زمان رسیدن مستلزم داشتن اطلاعات دقیق درباره موقعیت میکروفن‌ها و ویژگی‌های آکوستیکی آن‌ها شامل حساسیت و جهت‌گیری است. با استفاده از این اطلاعات و الگوریتم‌های محاسباتی مناسب، امکان تعیین موقعیت منبع صوتی فراهم می‌شود. در این زمینه، تابع همبستگی متقابل تعمیم‌یافته بیشترین کاربرد را دارد که با محاسبه همبستگی سیگنال‌ها و اعمال وزن‌دهی، دقت تخمین اختلاف زمان رسیدن در محیط‌های نویزی افزایش می‌یابد.

در روش‌های مبتنی بر شکل‌دهی پرتو (Chen, Yao & Hudson, ۲۰۰۳)، سیگنال‌های صوتی که از یک جهت مشخص می‌آیند تقویت می‌شوند و درعین‌حال سیگنال‌های ناخواسته از سایر جهات کاهش می‌یابند. این تکنیک داده‌های دریافت‌شده توسط آرایه میکروفن را پردازش می‌کند تا یک پرتو صوتی متمرکز ایجاد کند که انرژی صوتی را به سمت منبع هدف هدایت می‌کند. با این کار، موقعیت منبع صدا در محیط با دقت بیشتری قابل شناسایی است. این فرآیند شامل تخمین جهت سیگنال‌های ورودی و تقویت آن‌ها از زوایای مشخص و همزمان سرکوب نویز و تداخل‌های محیطی است. شکل‌دهی پرتو به دلیل توانایی آن در تمرکز انرژی صوتی و کاهش اثر انعکاس‌ها و اختلالات محیطی، به‌عنوان یک روش مؤثر و قابل اعتماد در بسیاری از کاربردها شناخته می‌شود؛ اما در مواردی که آرایه‌های میکروفونی گسترده هستند، نیازهای محاسباتی می‌توانند نسبتاً زیاد باشند. یکی دیگر از چالش‌های این روش، دشواری شناسایی منابع در فرکانس‌های پایین و در محیط‌هایی با سطوح بازتابنده محدود یا کاملاً منعکس‌کننده است. در چنین شرایطی، تکنیک‌های متداول شکل‌دهی پرتو ممکن است نتوانند نقشه‌ای دقیق و واقعی از موقعیت منابع ارائه دهند. علاوه بر این، وجود موانع و ساختارهای پیچیده محیط، فرآیند تخمین موقعیت را دشوارتر می‌کند، زیرا این عوامل به‌طور کامل در مدل‌های استاندارد مکان‌یابی در نظر گرفته نمی‌شوند (Gombots, Nowak & Kaltenbacher, ۲۰۲۱).

روش‌های با تفکیک‌پذیری بالا در حوزه طیف (Schmidt, ۱۹۸۶)، با مدل‌سازی سیگنال‌های دریافتی و تحلیل ماتریس کوواریانس آرایه، اطلاعات دقیق زاویه‌ای استخراج می‌کنند. روش‌های مختلفی برای تعیین زاویه‌ها و جهت منبع صدا وجود دارد. این روش‌ها شامل تخمین تأخیر زمانی، الگوریتم MUSIC (Tang, ۲۰۱۴) با تجزیه مقادیر ویژه ماتریس کوواریانس و الگوریتم ESPRIT (Ning, Ma, Meng & Wu, ۲۰۲۰) با استفاده از ساختار آرایه است. علاوه بر این، فرکانس موج صدا در تحلیل طیفی می‌تواند برای تخمین جهت ورود استفاده شود. دقت این روش به تعداد میکروفون‌ها بستگی دارد، اما انسجام سیگنال‌ها نیز بسیار مهم است. از چالش‌های این روش، حساسیت به نویز و محاسبات پیچیده است. این رویکرد، قابلیت تفکیک‌پذیری بسیار بالا و توانایی شناسایی منابع نزدیک به هم را دارا است.

در رویکردی دیگر، پژوهشگران به‌کارگیری الگوریتم‌های یادگیری ماشین کلاسیک را برای مکان‌یابی منبع صوتی مورد بررسی قرار دادند. به‌عنوان مثال، الگوریتم ماشین‌های بردار پشتیبان[1] با ویژگی‌های استخراج‌شده از روش‌های کلاسیک به‌عنوان ورودی مانند ویژگی‌های اختلاف زمانی رسیدن (Chen & Ser, ۲۰۱۱) یا ویژگی‌های شکل‌دهی پرتو (Salvati, Drioli & Foresti, ۲۰۱۶) مورد مطالعه قرار گرفتند. یا در مثالی دیگر، الگوریتم k نزدیک‌ترین همسایه[2] بر ویژگی‌های اختلاف فاز و جابجایی زمانی بین میکروفون‌ها (Gadre, Patole & Metkar, ۲۰۲۳) به کار گرفته شده است. به‌طورکلی، روش‌های یادگیری ماشین کلاسیک در مکان‌یابی منبع صوتی باعث بهبود نسبی نسبت به روش‌های کاملاً کلاسیک شده‌اند، اما وابستگی به ویژگی‌های دستی، محدودیت در تعمیم‌پذیری و ضعف در مدل‌سازی روابط پیچیده موجب شد که پژوهشگران به سمت یادگیری عمیق حرکت کنند، جایی که شبکه‌ها توانایی استخراج خودکار ویژگی‌ها و یادگیری مستقیم نگاشت‌های پیچیده را دارند.

در سال‌های اخیر، در راستای رفع محدودیت‌های روش‌های کلاسیک، رویکردهای مبتنی بر یادگیری عمیق برای مکان‌یابی منبع صوتی توسعه یافته‌اند. این روش‌ها برخلاف الگوریتم‌های کلاسیک که عمدتاً بر مدل‌های فیزیکی و فرضیات ایده‌آل متکی هستند،

1- Support Vector Machines (SVM)
2- K-Nearest Neighbours (KNN)

قادرند به‌طور مستقیم الگوهای پیچیده در داده‌های چندکاناله را بیاموزند و در شرایط واقعی با نویز، بازتاب و تغییرات محیطی عملکرد بهتری ارائه دهند (Takeda & Komatani, ۲۰۱۷; Perotin et al. ۲۰۱۸).

در حوزه مکان‌یابی منبع صوتی، تاکنون تعداد بسیار کمی از مقالات از معماری Mamba یا مدل‌های مبتنی بر آن استفاده کرده‌اند. اغلب پژوهش‌ها همچنان بر پایه شبکه‌های مرسوم مانند شبکه‌های عصبی کانولوشنی[۱]، بازگشتی[۲]، ترنسفورمرها[۳] و ترکیب‌های آن‌ها متمرکز هستند (Xu et al, ۲۰۲۵). در مقابل، استفاده از Mamba به دلیل تازگی این معماری و تمرکز اولیه آن بر پردازش دنباله[۴]، هنوز در مکان‌یابی منبع صوتی رایج نشده و تنها در موارد محدودی به‌صورت آزمایشی یا در قالب الهام‌گیری از ساختار آن به‌کار رفته است.

در مقاله (Xiao & Das, ۲۰۲۴)، مدل مکان‌یابی منبع صوتی با نام TF-Mamba معرفی شده است که ترکیبی از لایه‌های زمان و فرکانس را در قالب ساختار BiMamba در خود دارد. ورودی مدل شامل قسمت‌های حقیقی و موهومی STFT[۵] سیگنال‌هاست و با گرفتن اطلاعات زمانی و طیفی، نمایش یکپارچه‌ای برای تخمین دقیق زاویه منبع صدا ارائه می‌دهد؛ اما این مدل طراحی شده تنها به تخمین زاویه محدود است و فاصله‌ی منبع را در نظر نمی‌گیرد.

در مقاله (Mu et al, ۲۰۲۴)، مدل طراحی شده مبتنی بر معماری Mamba علاوه بر تشخیص جهت، فاصله‌ی منبع صوتی را نیز تخمین می‌زند. مدل از معماری دو مرحله‌ای آموزش بهره می‌برد: ابتدا تشخیص رویداد صوتی و زاویه‌ی ورود، سپس مرحله دوم شامل تخمین فاصله منبع. از Mamba برای حفظ اثربخشی محاسباتی و اطلاعات متنی/طیفی استفاده شده است. محدودیت مدل مطرح شده این است که پیچیدگی آموزش دو مرحله‌ای ممکن است برای کاربرد در محیط پرسرعت یا پر نویز مشکل‌ساز باشد.

1 -Convolutional Neural Network
2- Recurrent Neural Network
3- Transformer
4- Sequence modeling
5 -Short-Time Fourier Transform

در سناریوهای عملی، مکان‌یابی منبع صوتی اغلب به تخمین جهت رسیدن منابع، به‌ویژه زاویه‌های افقی و ارتفاع، ساده‌سازی می‌شود، درحالی‌که تخمین فاصله تا آرایه میکروفون را حذف می‌کند. رویکرد ساده شامل استفاده از حداقل دو یا چند میکروفون برای ضبط امواج صوتی، در کنار الگوریتم‌هایی است که سیگنال‌ها را تجزیه‌وتحلیل کرده و جهت منبع صدا را تعیین می‌کنند (Khan, Waqar, Kim & Park, ۲۰۲۵).

مدل پیشنهادی مقاله (Qayyum et al, ۲۰۲۰) یک شبکه عصبی کانولوشنی یک‌بعدی است که زاویه و ارتفاع منبع صوتی را مستقیماً از سیگنال صوتی خام استخراج می‌کند و برای تخمین فاصله یا موقعیت سه‌بعدی کامل طراحی نشده است. مزیت اصلی این مدل آن است که نیازی به ویژگی‌های صوتی دست‌ساز یا روش‌های پیچیده کاهش نویز خودکار ندارد و می‌تواند به‌صورت انتها-به-انتهای زاویه منبع صوتی را تخمین بزند.

مقاله (Hu, Song, He & Yu, ۲۰۲۳)، شبکه‌ی عمیق مبتنی بر ساختار باقیمانده[1] و مکانیزم توجه طراحی نموده که ترکیبی از طیف‌نگار لگاریتمی مل[2] و همبستگی متقابل تعمیم‌یافته با تبدیل فاز را به‌عنوان ویژگی‌های ورودی در نظر می‌گیرد و اطلاعات زمان-فرکانس را برای تخمین موقعیت صوتی استخراج می‌کند. این مدل فقط برای تخمین زاویه‌ی افقی و زاویه‌ی ارتفاع طراحی‌شده و تخمین فاصله در خروجی مدل لحاظ نشده است.

یکی از چالش‌های کلیدی در مکان‌یابی منبع صوتی مبتنی بر یادگیری عمیق، اطمینان از در دسترس بودن کمی و کیفی داده‌های آموزشی است، به‌گونه‌ای که مدل بتواند الگوهای لازم را به‌طور مؤثر بیاموزد. جمع‌آوری داده‌های چندکاناله واقعی با برچسب دقیق موقعیت منبع صوتی دشوار و پرهزینه است و نویزهای غیرایستا، بازتاب‌های متعدد و شرایط آکوستیکی متغیر می‌توانند کیفیت و تنوع داده‌ها را کاهش دهند که این امر منجر به افت تعمیم‌پذیری مدل در محیط‌ها و موقعیت‌های جدید می‌شود. به‌عنوان مثال، مجموعه داده (Ruiz-Espitia, Martinez-Carranza & Rascon, ۲۰۱۸) برای مکان‌یابی پهپاد به پهپاد است و از ۸ میکروفون نصب‌شده روی پهپاد استفاده می‌کند. این مجموعه شامل ضبط‌های صوتی دو پهپاد در حال پرواز تا فاصله ۳ متر است. این مجموعه به‌طور خاص چالش‌های نویز ناشی از پره‌های پهپاد را در نظر گرفته است، اما محدودیت‌هایی از جمله محدودیت

1- Residual
2- log-Mel spectrogram

فاصله منبع صوتی و تنوع محیط‌ها و منابع صوتی را دارا است. یکی دیگر از منابع داده در دسترس (Strauss, Mordel, Miguet & Deleforge, ۲۰۱۸) است که هشت میکروفون در قالب یک مکعب زیر پهپاد چهارملخه، در حال ضبط سیگنال‌های گفتار یا نویز ضبط‌شده در حین پرواز با نسبت سیگنال به نویز منفی و پروازهای جداگانه برای ثبت نویز پهپاد با تمرکز روی آزمایشگاه و شرایط کنترل‌شده است. پهپاد در دو فضای محصور با طول، عرض و ارتفاع به ترتیب ۱۰، ۱۰ و ۲/۵ متر و ۱۲، ۱۲ و ۳/۵ متر به پرواز درآمده است که محدودیت در تنوع محیط‌های عملی واقعی را دارد. مجموعه داده (Wang, Sanchez-Matilla & Cavallaro, ۲۰۱۹) نیز تمرکز بر سناریوهای آزمایشگاهی و فضاهای محدود و کمتر منعکس‌کننده محیط‌های واقعی را دارد زیرا حداکثر دو گوینده در نه مکان از پیش تعریف شده بین دو تا شش متر از پهپاد قرار می‌گیرند.

اخیراً مجموعه داده (Jekateryńczuk, Szadkowski & Piotrowski, ۲۰۲۵) با هدف نظارت پهپاد از زمین ارائه شده است که شامل صوت‌های با کیفیت بالا از یک پهپاد در محیط‌های واقعی است که از فواصل، ارتفاعات، زوایا و جهت‌گیری جانبی مختلف پهپاد گرفته شده است و شامل اطلاعات دقیقی مانند مختصات و شرایط محیطی است.

با وجود پیشرفت‌های اخیر در مکان‌یابی منبع صوتی، بسیاری از مطالعات به نویز و بازتاب‌های واقعی محیطی، محدودیت‌های مجموعه‌داده‌ها و وابستگی به ویژگی‌های دستی توجه نکرده‌اند؛ این محدودیت‌ها زمینه را برای توسعه رویکردهای دارای توانایی استخراج خودکار ویژگی‌ها و مدل‌سازی روابط پیچیده چندکاناله را فراهم کرده است که می‌تواند در سامانه‌های نظامی و کاربردهای بلادرنگ برای شناسایی و ردیابی منابع صوتی حیاتی باشد.

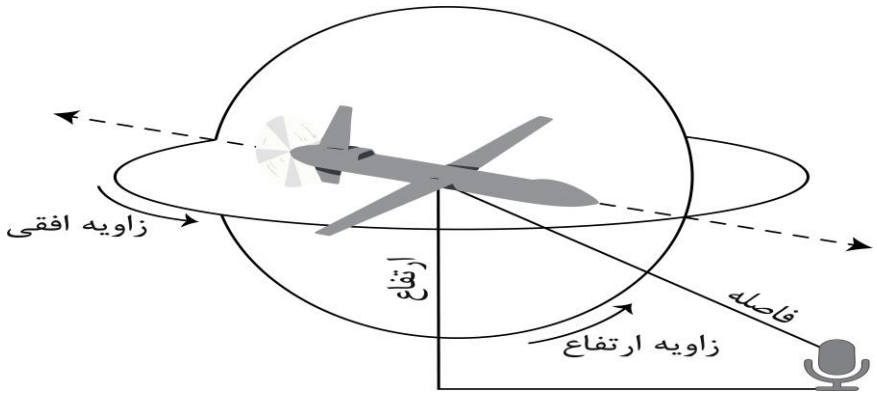
مبانی نظری

مکان‌یابی منبع صوتی به فرآیندی اطلاق می‌شود که در آن موقعیت یک یا چند منبع صوتی نسبت به یک مرجع مشخص، معمولاً آرایه‌ای از میکروفون‌ها، تخمین زده می‌شود. این موقعیت می‌تواند در ابعاد مختلفی تعریف شود که مکان‌یابی سه بعدی منبع صدا لازم است موارد زیر را دارا باشد:

- فاصله: فاصله شعاعی میکروفون از مرکز آرایه، بر حسب متر که موقعیت سه‌بعدی منبع صوتی را نسبت به آرایه مشخص می‌کند.

- ارتفاع: موقعیت عمودی میکروفون نسبت به سطح زمین، برحسب متر که تعیین می‌کند منبع صوتی در چه ارتفاعی قرار دارد.
- زاویه افقی و زاویه ارتفاع: این پارامترها زاویه‌های عمودی و افقی میکروفون نسبت به مرکز آرایه را مشخص می‌کنند و تضمین می‌کنند که میکروفون‌ها با دقت مناسب در راستای مرکز قرار گرفته و هم‌راستا شوند.

تعیین دقیق شاخص‌های موقعیت پهپاد، به‌ویژه در کاربردهای نظامی و امنیتی، اهمیت حیاتی دارد؛ چرا که صرفاً دانستن جهت صدا برای رهگیری یا ارائه واکنش به موقع کافی نیست. مکان‌یابی دقیق پهپادها امکان شناسایی به‌موقع تهدیدات هوایی، رهگیری پهپادهای غیرمجاز و برنامه‌ریزی مؤثر اقدامات دفاعی را فراهم می‌کند و به نیروهای نظامی اجازه می‌دهد پاسخ‌های سریع و هدفمند ارائه دهند و خطر نفوذ یا حملات غیرمنتظره را به حداقل برسانند.



شکل (۱) شاخص‌های تخمین موقعیت سه‌بعدی منبع صوتی نسبت به آرایه میکروفون

اکثر سامانه‌ها و پژوهش‌های موجود همان‌طور که در بخش پیشینه پژوهش شرح داده شد، محدود به مکان‌یابی دو بعدی هستند؛ یعنی تنها زاویه افقی و گاهی زاویه ارتفاع را تخمین می‌زنند و فاصله منبع از آرایه نادیده گرفته می‌شود. این محدودیت‌ها باعث بروز چالش‌های نظیر ناتوانی در برآورد مسیر واقعی حرکت منابع پرنده یا متحرک. کاهش دقت رهگیری و واکنش در محیط‌های پیچیده و پر نویز و محدودیت کاربرد در شرایط عملیاتی واقعی به‌ویژه در محیط‌های شهری یا نظامی با بازتاب‌های متعدد و نویز پس‌زمینه

می‌شوند. لذا، توسعه مکان‌یابی سه‌بعدی که شامل زاویه افقی، زاویه ارتفاع و فاصله و ارتفاع پهپاد از سطح زمین باشد، یکی از شکاف‌های اصلی پژوهشی در ادبیات مکان‌یابی صوتی پهپاد است و پیشرفت در این زمینه می‌تواند به‌طور قابل توجهی قابلیت سامانه‌های نظارتی و دفاعی را ارتقا دهد.

توانایی انسان در تعیین موقعیت منابع صوتی و تخمین جهت و فاصله منابع صوتی با دقت بالا، حتی در محیط‌های شلوغ و پر نویز، نشان‌دهنده یک سیستم پردازشی قدرتمند است که به‌طور همزمان ویژگی‌های زمانی و فرکانسی صدا را تحلیل می‌کند. الهام‌گیری از این توانایی‌ها باعث شکل‌گیری رویکردهای نوین در طراحی الگوریتم‌های مکان‌یابی صوتی شده است. به‌ویژه، ترکیب ویژگی‌های زمانی (مانند اختلاف زمان رسیدن امواج صوتی به میکروفون‌ها) و ویژگی‌های طیفی (مانند انرژی در باندهای فرکانسی مختلف) می‌تواند همانند سیستم ادراکی انسان، امکان تخمین دقیق زاویه‌ها، ارتفاع و فاصله را فراهم کند. همان‌طور که در بخش قبل نیز شرح داده شد، شبکه‌های عصبی کانولوشنی توانایی بالایی در استخراج ویژگی از نگاشت‌های طیفی صوت دارند، به‌ویژه برای تخمین دوبعدی جهت رسیدن منابع صوتی موفق بوده‌اند؛ اما در مدل‌سازی روابط بلندمدت و وابستگی‌های زمانی (روند تغییرات سیگنال در طول بازه‌های طولانی) و مکانی (فاز و اختلاف زمان رسیدن صدا به هر میکروفون از آرایه) در داده‌های چندکاناله، دارای محدودیت می‌باشند. ترنسفورمرها نیز به دلیل آن‌که بر اساس مکانیزم توجه عمل می‌کنند و قادرند روابط بلندمدت بین ویژگی‌های زمانی و مکانی را مدل کنند، در مسائل پیچیده‌تر مانند مکان‌یابی سه‌بعدی، کارایی بالاتری نسبت به شبکه‌های عصبی کانولوشنی نشان داده‌اند و دارای مقیاس‌پذیری و توانایی یادگیری نمایش‌های غنی از داده‌های بزرگ می‌باشند؛ اما روش‌های مبتنی بر توجه دارای پیچیدگی محاسباتی درجه دوم نسبت به طول دنباله هستند. این ویژگی باعث می‌شود که پردازش سیگنال‌های صوتی طولانی و چندکاناله، به‌ویژه در مسائل مکان‌یابی سه‌بعدی، از نظر محاسباتی بسیار پرهزینه و سنگین باشد. در مکانیزم خود-توجه[1]، هر نمونه‌ی دنباله باید با همه‌ی نمونه‌های دیگر مقایسه شود. پس

1- Self-Attention

اگر طول دنباله N باشد، پیچیدگی زمانی و حافظه $O(N^2)$ است که برای صوت چندکاناله که دنباله‌های طولانی تولید می‌شود، این بسیار پرهزینه است.

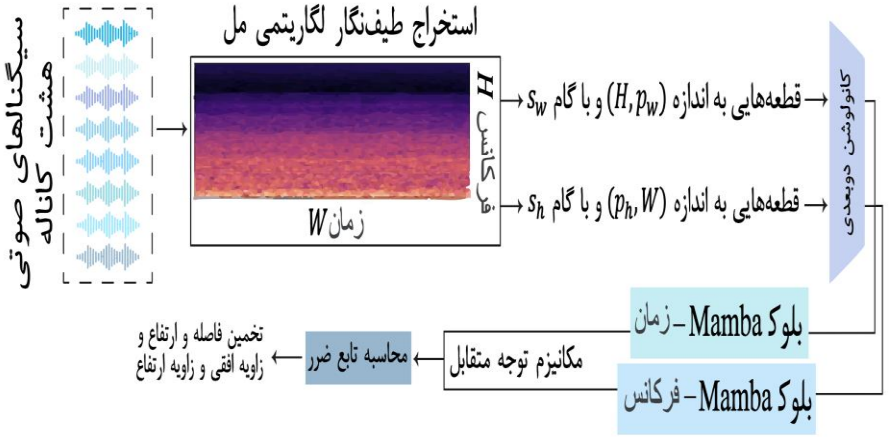
به‌منظور غلبه بر محدودیت محاسباتی ترنسفورمرها، مدل Mamba بر پایه مدل‌های توالی فضای حالت[1] ساخته شده است که این مدل‌ها برخلاف مکانیزم توجه، به‌صورت بازگشتی روی توالی حرکت می‌کنند و دارای پیچیدگی خطی به‌صورت $O(N)$ هستند؛ یعنی Mambaمی‌تواند جایی که ترنسفورمرها به‌دلیل محدودیت در حافظه یا زمان شکست می‌خورند، همچنان کارا بماند.

مدل پیشنهادی این مقاله توانایی تخمین دقیق فاصله، ارتفاع و زوایای افقی و ارتفاعی منبع صوتی را دارد و با ادغام ویژگی‌های زمانی و طیفی به کمک معماری Mamba که امکان پردازش بازگشتی و کارآمد توالی‌های طولانی را با پیچیدگی خطی فراهم می‌کند، کارایی قابل‌توجهی در محیط‌های پر نویز و شرایط واقعی ارائه می‌دهد. این ویژگی‌ها امکان توسعه سیستم‌های نظارتی و دفاعی بلادرنگ با دقت و اطمینان بالا را فراهم می‌سازند.

روش‌شناسی

در این بخش، معماری عمیق پیشنهادی برای تخمین سه‌بعدی موقعیت منبع صوتی پهپاد با استفاده از سیگنال‌های صوتی چندکاناله ارائه می‌شود. معماری کلی در شکل ۲ نشان داده شده است. مدل پیشنهادی شامل مراحل زیر است: استخراج طیف‌نگار لگاریتمی مل از سیگنال‌های صوتی هشت کاناله، تقسیم طیف‌نگار ورودی به قطعه‌های دوبعدی در حوزه‌ی زمان و فرکانس و استخراج ویژگی با استفاده از لایه‌های کانولوشن دوبعدی، پردازش توالی قطعه‌های زمانی و فرکانسی با بلوک‌های Mamba، ادغام ویژگی‌های استخراج شده از حوزه‌های زمان و فرکانس و در نهایت تخمین اطلاعات مکانی پهپاد از طریق لایه‌های خروجی است.

1- State Space Sequence Models



شکل (۲) معماری عمیق پیشنهادی برای مکان‌یابی سه‌بعدی منبع صوتی پهپاد

در ابتدا، برای هر یک از هشت کانال آرایه میکروفون، طیف‌نگار لگاریتمی مل استخراج می‌شود. طیف‌نگار لگاریتمی مل به‌عنوان یکی از پرکاربردترین ویژگی‌ها در حوزه‌ی پردازش سیگنال صوت شناخته می‌شود. این ویژگی بیانگر توزیع انرژی سیگنال بر روی محور فرکانسی مل است که از طریق نگاشت داده‌های طیفی خطی به مقیاس مل با استفاده از بانک فیلتر مل[1] حاصل می‌گردد (Davis & Mermelstein, ۱۹۸۰). بدین‌ترتیب، فرایند استخراج این ویژگی با برجسته‌سازی مؤلفه‌های ادراکی مرتبط با سیستم شنوایی انسان، امکان تحلیل دقیق‌تر و کارآمدتر سیگنال‌های صوتی را فراهم می‌سازد.

فرض کنید موج صوتی هشت کاناله حاصل از آرایه میکروفون به‌صورت $x(t) = [x_1(t), x_2(t), \dots, x_8(t)]$ باشد که در آن $x_c(t)$، سیگنال زمانی کانال c-ام است و t شاخص نمونه‌های زمانی گسسته است. برای استخراج طیف‌نگار لگاریتمی مل از سیگنال‌های صوتی، روند زیر را طی می‌کنیم. ابتدا برای هر کانال، سیگنال به قاب‌های کوتاه تبدیل شده و سپس روی هر قاب، تبدیل فوریه سریع[2] به‌صورت معادله (۱) محاسبه می‌شود که $w(n)$ پنجره به طول N، H اندازه گام، m شاخص قاب زمانی و k بازه‌های فرکانسی در تبدیل فوریه سریع است:

1- Mel Filter Bank
2- Fast Fourier Transform (FFT)

$$X_c(m,k) = \sum_{n=0}^{N-1} x_c(n+mH)w(n)e^{-j2\pi kn/N}, k = 0, \dots, N-1$$

(۱)

سپس طبق معادله (۲)، با استفاده از بانک فیلتر مل به M قسمت در مقیاس مل نگاشت می‌شود که $H_i(k)$ فیلتر مل i-ام است و هر فیلتر بر محدوده فرکانسی مربوط به باند مل خود تاکید می‌کند:

$$S_c(m,i) = \sum_{k=0}^{N-1} H_i(k).|X_c(m,k)|^2, i = 1, \dots, M$$

(۲)

ازآنجایی‌که درک انسان از فرکانس غیرخطی است، تبدیل از مقیاس فرکانس خطی (هرتز) به مقیاس مل به‌صورت معادله (۳) است که در آن f_{mel} فرکانس در مقیاس مل و f_z فرکانس بر حسب هرتز است:

$$f_{mel} = 2595 \log_{10}(1 + \frac{f_{Hz}}{700})$$

(۳)

در نهایت، برای تقلید از درک بلندی صدا توسط انسان (تقریباً لگاریتمی)، طیف‌نگار لگاریتمی مل به‌صورت معادله (۴) به دست می‌آید که ϵ عدد کوچکی برای جلوگیری از لگاریتم صفر است.

$$LogMel_c(m,i) = \log(S_c(m,i) + \epsilon)$$

(۴)

در شکل ۳، طیف‌نگار لگاریتمی مل یک نمونه از صوت‌های هشت کاناله میکروفون آرایه‌ای است که هنگام پرواز پهپاد ضبط شده است. آرایه میکروفون‌ها یک آرایه مربعی دو سطحی (در دو ارتفاع) است که می‌تواند برای تخمین جهت و ارتفاع منبع صوتی مناسب باشد. کانال‌های یک تا چهار روی محورهای اصلی ۰, ۹۰, ۱۸۰, ۲۷۰ درجه و کانال‌های پنج تا هشت روی محورهای مورب ۴۵, ۱۳۵, ۲۲۵, ۳۱۵ درجه نصب شده‌اند. در هر طیف‌نگار لگاریتمی مل، محور افقی نشانگر زمان (۰ تا حدود ۱۴۰ قاب زمانی) و محور عمودی نشانگر فرکانس (۰ تا ۲۵۶ مل) و رنگ نیز نشانگر شدت انرژی (زرد = قوی، بنفش و مشکی = ضعیف) است. در همه کانال‌ها یک نویز زمینه پهن‌باند دیده می‌شود که انرژی بیشتری در فرکانس‌های پایین دارد که بیانگر نویز محیط اطراف و صدای ملخ‌های پهپاد

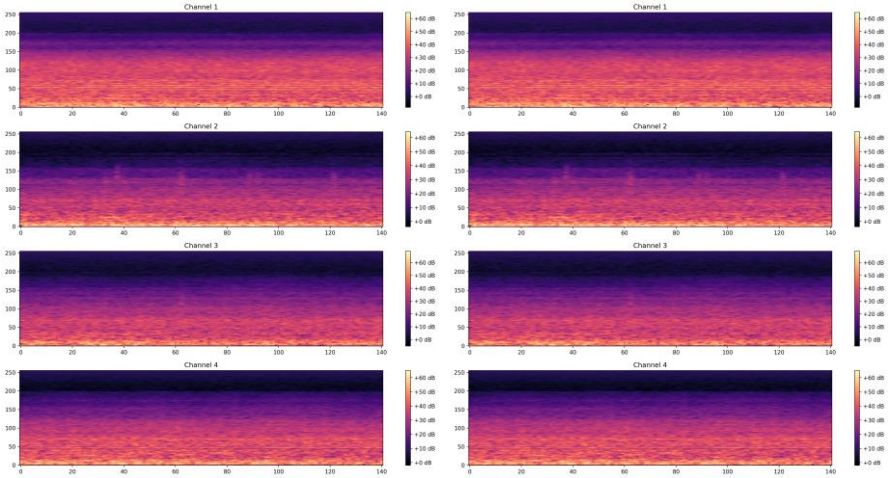
است. این تفاوت‌ها به علت جهت رسیدن صدا به آرایه است؛ میکروفون‌هایی که رو به پهپاد هستند یا مانع کمتری دارند، سیگنال واضح‌تری از ملخ‌ها ضبط می‌کنند، درحالی‌که سایر کانال‌ها بیشتر نویز پراکنده را می‌گیرند. این اختلاف‌ها دقیقاً سرنخ‌هایی هستند که در مکان‌یابی منبع صوتی استفاده می‌شوند:

اختلاف سطح بین کانال‌ها: بعضی میکروفون‌ها شدت بیشتری از هارمونیک‌ها می‌گیرند.

اختلاف زمان رسیدن: در اصل هارمونیک‌ها کمی زودتر یا دیرتر در کانال‌ها ظاهر می‌شوند.

وضوح طیفی متفاوت: بعضی کانال‌ها امضای پهپاد را واضح‌تر دارند.

ترکیب داده‌های هشت کاناله باعث می‌شود مدل بتواند موقعیت سه‌بعدی پهپاد را تخمین بزند. در مثال سیگنال صوتی شکل ۳، صدای پهپاد در کانال‌های ۱، ۴، ۷ و ۸ واضح است، درحالی‌که در کانال‌های ۵ و ۶ با فاصله‌ای دور شنیده می‌شود و در کانال‌های ۲ و ۳ به سختی قابل تشخیص است.



شکل (۳) طیف‌نگار لگاریتمی مل مربوط به سیگنال صوتی هشت کاناله از پهپاد در حال پرواز

در این مقاله، پارامترهای استخراج طیف‌نگار لگاریتمی مل شامل نرخ نمونه‌برداری، طول تبدیل فوریه سریع (N)، طول گام (H) و تعداد فیلترهای مل (M) به ترتیب ۹۶ kHz،

۴۰۹۶، ۱۰۲۴ و ۲۵۶ انتخاب شدند. این پیکربندی باعث ایجاد توازن بین دقت فرکانسی و زمانی برای شناسایی ویژگی‌های طیفی مرتبط با صدای پهپاد شد.

حال، طیف‌نگارهای لگاریتمی مل هشت کاناله با ابعاد $LogMel \in \mathbb{R}^{C\times H\times W}$ که H تعداد باند مل (محور فرکانس) برابر با ۲۵۶، W تعداد قاب زمانی (محور زمان) برابر با ۹۴ و C تعداد کانال برابر با ۸، به قطعه‌های دوبعدی دارای هم‌پوشانی در حوزه‌های زمان و فرکانس تقسیم می‌شوند. در حوزه‌ی زمان، طیف‌نگار ابتدا به قطعه‌هایی به اندازه (H, p_w) و با گام s_w تقسیم می‌شود و سپس کانولوشن دوبعدی روی هر قطعه برای استخراج ویژگی‌ها اعمال می‌گردد؛ در حوزه‌ی فرکانس نیز، طیف‌نگار ابتدا به قطعه‌هایی به اندازه (p_h, W) و با گام s_h تقسیم می‌گردد و کانولوشن دوبعدی روی هر قطعه اعمال می‌شود. تعداد قطعه‌های فرکانسی و زمانی طبق معادله (۵) به دست می‌آیند. در این مقاله، به‌صورت تجربی مقادیر p_w ، s_w ، p_h ، و s_h به ترتیب ۳۰ ، ۴ ، ۳۲ و ۱۴ تعیین شده‌اند.

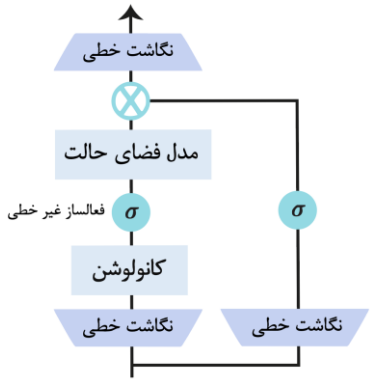
$$N_t = \frac{W-P_w}{s_w} + 1,\ \ N_f = \frac{H-P_h}{s_h} + 1 \quad (۵)$$

پس از به دست آمدن این توالی قطعه‌های زمانی و فرکانسی، پردازش توسط بلوک‌های Mamba که ساختار آن در شکل ۴ نشان داده شده، انجام می‌شود. هر بلوک Mamba ترکیبی از لایه کانولوشن و یک مدل فضای حالت است که امکان استخراج هم‌زمان وابستگی‌های کوتاه‌برد و بلندبرد را فراهم می‌کند. هنگامی‌که این بلوک در راستای محور زمان اعمال می‌شود، پویایی‌های زمانی سیگنال را مدل‌سازی می‌کند؛ درحالی‌که در راستای محور فرکانس، همبستگی‌های طیفی را می‌آموزد.

بلوک Mamba در شکل ۴، مدل‌سازی زمانی و طیفی را در یک ساختار واحد ادغام می‌کند. هر بلوک ابتدا یک نگاشت خطی را برای گسترش فضای ویژگی اعمال می‌کند که به دو جریان تقسیم می‌شود: یکی تحت کانولوشن قرار می‌گیرد و به دنبال آن یک مدل فضای حالت برای ثبت وابستگی‌های دوربرد عمل می‌کند، درحالی‌که جریان دیگر به‌عنوان یک سیگنال دروازه‌ای عمل می‌کند. فعال‌سازی‌های غیرخطی و دروازه ضربی این مسیرها را ترکیب می‌کنند و نتیجه از طریق یک نگاشت خطی فشرده می‌شود. این طراحی، بلوک را قادر می‌سازد تا به‌طور مؤثر همبستگی‌های محلی (از طریق کانولوشن)

و وابستگی‌های متوالی سراسری (از طریق مدل فضای حالت) را مدل‌سازی کند که آن را برای محلی‌سازی منبع صدای پهپاد در محیط‌های پر سروصدا مناسب می‌کند. در واقع، یک مدل فضای حالت، چگونگی تکامل یک حالت پنهان در طول زمان و نحوه تولید خروجی‌ها را طبق فرمول (۶) شرح می‌دهد که $\tilde{X}$ ویژگی استخراج‌شده از لایه کانولوشن است.

$$Y = SSM(\tilde{X});\ where \begin{cases} h_{t+1} = Ah_t + B\tilde{x}_t \\ y_t = Ch_t + D\tilde{x}_t \end{cases} \quad (۶)$$



شکل (۴) ساختار بلوک Mamba

پس از عبور ویژگی‌های حوزه زمان و فرکانس از بلوک‌های Mamba، ویژگی‌های زمانی به‌عنوان جست‌وجو[1] و ویژگی‌های فرکانسی به‌عنوان کلید[2] و مقدار[3] وارد مکانیزم توجه متقابل[4] طبق معادله (۷) می‌شوند تا اطلاعات فرکانسی روی ویژگی‌های زمانی اعمال و ادغام گردد. در این معادله، W_Q، W_K و W_V ماتریس‌های قابل یادگیری برای پرس‌وجو، کلید و مقدار و همچنین d_h ابعاد هر سر[5] توجه است.

1- Query
2- Key
3- Value
4- Cross Attention
5 -head

$$CrossAttention(t,f) = Softmax(^{tW_Q(fW_K)^\top}/_{\sqrt{d_h}})(fW_V) \quad (۷)$$

در نهایت، برای تخمین فاصله و ارتفاع بر حسب متر و همچنین زاویه افقی و زاویه ارتفاع بر حسب درجه، ما تابع ضرر[1] را برای هر پارامتر به‌طور جداگانه با استفاده از فاصله اقلیدسی جفتی[2] طبق معادله (۸) محاسبه می‌کنیم و بدون هیچ‌گونه وزن‌دهی اضافی با یکدیگر جمع می‌کنیم؛ بنابراین، همه‌ی پارامترها شامل فاصله، ارتفاع و نمایش سینوسی و کسینوسی زوایا در تابع خطا وزن مساوی دارند و مدل هرکدام را با اهمیت یکسان بهینه می‌کند. هدف ما از این طراحی، جلوگیری از جانبداری تابع خطا به نفع یک متغیر خاص و ساده نگه‌داشتن فرآیند آموزش بوده است. زوایای افقی و ارتفاع با استفاده از مؤلفه‌های سینوس و کسینوس نمایش داده می‌شوند تا دایره‌ای بودن مقادیر زاویه‌ای در نظر گرفته شود.

$$\mathcal{L}_x = \| x_{pred} - x_{target} \|_2 \quad (۸)$$

برای آموزش و ارزیابی مدل، از پایگاه داده (Jekateryńczuk, Szadkowski & Piotrowski, ۲۰۲۵) که یک مجموعه‌داده در دسترس عمومی برای پژوهش‌های مرتبط با مکان‌یابی منبع صوتی پهپادها که شامل دو منبع صوتی اصلی نویز محیطی و صدای پهپاد است، استفاده می‌کنیم. این پایگاه داده شامل صدای پهپادها و نویز محیطی در شرایط مختلف فاصله، ارتفاع و جهت‌گیری بوده و امکان ارزیابی الگوریتم‌های پردازش سیگنال و یادگیری عمیق را در سناریوهای واقعی فراهم می‌کند. در جدول (۱)، اطلاعات این پایگاه داده نشان داده شده است. این پایگاه داده شامل نویز روتور پهپاد ثبت شده در ارتفاعات، فواصل و زوایای مختلف نسبت به آرایه میکروفون در جهات جلو، عقب، راست و چپ پهپاد همچنین و نویز پس‌زمینه از محیط شهری است.

جدول (۱) اطلاعات پایگاه داده مورد استفاده در این مقاله

مدت زمان (بر حسب ثانیه)	منبع صدا

1- Loss function
2 -pairwise Euclidean distance (L2)

۴۱۶	نویز محیط
۵۱۲۰	پرواز پهپاد

این ضبط‌ها درمجموع ۳۳۰۹۰ ثانیه طول دارند و با فرکانس ۹۶ کیلوهرتز و عمق ۳۲ بیت نمونه‌برداری شده‌اند که ضبط صدا با وضوح بالا را تضمین می‌کند. این مجموعه داده نقاط قوت و همچنین محدودیت‌هایی را ارائه می‌دهد که هنگام استفاده از آن برای تحقیقات مکان‌یابی منبع صدای پهپاد باید در نظر گرفته شوند. یکی از مزایای اصلی، حجم قابل توجه صدای پهپاد موجود است: با ۱۲۸ ضبط در مجموع تقریباً ۵۱۲۰ ثانیه، این مجموعه داده‌ها تنوع صوتی غنی را در شرایط پرواز، مسیرها و تنظیمات ضبط مختلف ارائه می‌دهد. چنین تنوعی برای آموزش مدل‌های قوی که قادر به تعمیم به محیط‌های دنیای واقعی هستند، به‌ویژه در وظایف محلی‌سازی که نیاز به یادگیری تغییرات مداوم در نشانه‌های زاویه ورود دارند، مفید است. حجم کل بالای داده‌ها تقریباً ۱۶ گیگابایت، بیشتر نشان می‌دهد که ضبط‌ها احتمالاً اطلاعات مکانی چند کاناله با وضوح بالا را حفظ می‌کنند که برای مکان‌یابی دقیق، شکل‌دهی پرتو و استخراج ویژگی‌های مکانی ضروری است. علاوه بر این، این مجموعه داده شامل ضبط‌های نویز محیطی واقعی است که اگرچه تعداد آن‌ها محدود است، اما پایه مفیدی برای مدل‌سازی پس‌زمینه‌های محیطی واقع‌گرایانه و کاهش پیش‌پردازش برای داده‌های پاک فقط پهپاد فراهم می‌کند. همچنین در دسترس بودن این پایگاه داده به‌صورت آزاد، تکرارپذیری و مقایسه‌پذیری نتایج تحقیقات را نیز افزایش می‌دهد، درحالی‌که بسیاری از پایگاه داده‌های مربوط به این حوزه، یا در دسترس نیستند یا از غنای کافی شرایط متفاوت برخوردار نمی‌باشند.

فرآیند آموزش بر بهینه‌سازی مدل یادگیری عمیق با استفاده از بهینه‌ساز AdamW متمرکز است. در این بهینه‌ساز، نرخ یادگیری برابر با ۰/۰۰۰۰۱ و مقادیر β نیز در محدوده ۰/۵ تا ۰/۹ انتخاب شده‌اند تا همگرایی پایدار و کارآمد تضمین گردد. آموزش مدل با ۵۰ دوره[1] و اندازه دسته[2] ۳۲ انجام می‌شود که تعادلی میان کارایی محاسباتی و کیفیت عملکرد برقرار می‌سازد. هر تکرار آموزش شامل سه مرحله اصلی است: انتشار رو به جلو[3]

1- Epoch
2- Batch Size
3 -Forward propagation

که مدل داده‌های ورودی را پردازش کرده و خروجی پیش‌بینی‌شده را تولید می‌کند؛ محاسبه تابع ضرر برای ارزیابی خطا از فاصله اقلیدسی جفتی استفاده می‌شود؛ انتشار رو به عقب[1] که وزن‌های مدل بر اساس گرادیان‌های محاسبه‌شده به‌روزرسانی می‌شوند.
به‌منظور بهبود پویایی فرآیند یادگیری، از زمان‌بند نرخ یادگیری CosineAnnealingLR استفاده شده است که نرخ یادگیری را به‌تدریج از مقدار اولیه به سمت یک مقدار حداقل کاهش می‌دهد و سپس به‌صورت دوره‌ای با الگوی کسینوسی نوسان می‌کند که باعث می‌شود مدل ضمن همگرایی پایدار، از افتادن در مینیمم‌های محلی جلوگیری کرده و قابلیت اکتشاف فضای پارامترها را افزایش دهد.
برای ارزیابی عملکرد مدل پیشنهادی، ما از دو معیار رایج استفاده کردیم: میانگین مطلق خطا[2] و ریشه میانگین مربعات خطا[3]. این معیارها برای اندازه‌گیری خطاهای پیش‌بینی در فاصله و ارتفاع (بر حسب متر) و همچنین زوایای ارتفاع و افقی (بر حسب درجه) انتخاب شدند و به کمی‌سازی دقت تخمین‌های مدل برای هر پارامتر کمک کردند. میانگین مطلق خطا، میانگین اختلاف مطلق بین مقادیر پیش‌بینی‌شده و واقعی را طبق فرمول (۹) محاسبه می‌کند. استفاده از این دو معیار به‌طور خاص انتخاب شد، زیرا در مقاله‌ی پایه نیز دقیقاً همین معیارها برای گزارش نتایج مورد استفاده قرار گرفته بودند؛ بنابراین، به‌کارگیری این دو شاخص به ما امکان می‌دهد مقایسه‌ای مستقیم، منصفانه و قابل اتکا با نتایج مقاله مرجع انجام دهیم. درزمینهٔ فاصله و ارتفاع، این درک ساده‌ای از میزان انحراف پیش‌بینی‌های مدل از مقادیر واقعی بر حسب متر ارائه می‌دهد. برای زوایای افقی و ارتفاع که بر حسب درجه بیان می‌شوند، میانگین مطلق خطا به ما می‌گوید که موقعیت‌های زاویه‌ای پیش‌بینی‌شده به‌طور متوسط چقدر با زوایای واقعی متفاوت هستند. در مقابل، ریشه میانگین مربعات خطا، جذر میانگین مربعات اختلاف بین مقادیر پیش‌بینی‌شده و واقعی را طبق فرمول (۱۰) محاسبه می‌کند. این معیار مستعد خطاهای بزرگتر است و تأکید بیشتری بر مواردی دارد که پیش‌بینی‌های مدل به‌طور قابل‌توجهی از مقادیر واقعی انحراف دارند.

1- Backpropagation
2- Mean Absolute Error (MAE)
3- Root Mean Squared Error (RMSE)

$$\mathcal{L}_{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (۹)$$

$$\mathcal{L}_{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (۱۰)$$

مدل پیشنهادی ما با توجه به این که بر اساس مدل‌های توالی فضای حالت و بلوک‌های Mamba ساخته شده است، یک مدل دارای پیچیدگی خطی نسبت به تعداد توکن‌ها به صورت O(N) است که هم حافظه و هم زمان تست کمی مصرف می‌کند و برای پردازش بلادرنگ در سیستم‌های دفاعی قابل استفاده است. همچنین لازم به ذکر است که مدل توالی فضای حالت و بلوک‌های Mamba استفاده شده در این مدل، برای استفاده در GPU و تراشه‌های edge با ترایتون (Triton) طراحی و آموزش داده شده است به این معنا که هم قابلیت موازی‌سازی روی GPU و هم پیاده‌سازی سبک روی تراشه‌های edge با کارایی بالا و حافظه بهینه را داشته باشند. با استفاده از Triton، عملیات ماتریسی داخل Mamba و محاسبات مدل توالی فضای حالت می‌تواند به‌صورت بلاک‌بندی شده و همزمان اجرا شود که باعث افزایش سرعت و کاهش مصرف حافظه می‌شود. به‌طور خاص، توابعی که شامل ضرب ماتریس‌ها، محاسبات توجه و اعمال فعال‌سازی هستند، می‌توانند روی هسته‌های GPU بهینه شوند. این ویژگی باعث می‌شود که هم حافظه و هم زمان تست مدل بسیار کم باشد و پردازش بلادرنگ در سیستم‌های دفاعی ممکن شود. مدل ما با عمق L=۱۲ بلوک و ابعاد فضای تعبیه D=۱۹۲ طراحی شده است که در هر بلوک، یک پرسپترون چندلایه B=۷۶۸ وجود دارد و برای هر نمونه، تعداد توکن‌ها N=۶۲ است. تعداد پارامترهای مدل نیز P=۷/۵ M است که با فرمول (۱۱)، حافظه مدل به دست می‌آید. (FP۳۲ یک قالب ۳۲ بیتی برای نمایش اعداد حقیقی در رایانه‌ها است که برای ایجاد تعادل بین دقت و سرعت در برنامه‌هایی مانند گرافیک، یادگیری ماشین و محاسبات علمی استفاده می‌شود.)

$$Memory_{model} \approx P \times 4\ (FP32) \approx 30MB \quad (۱۱)$$

این مقدار به دست آمده برای کاربردهای عملی که مدل تبدیل به FP۱۶ یا Int۸ می‌شود، به نصف کاهش می‌یابد. همچنین، پیچیدگی محاسباتی هر بلوک که شامل لایه‌ها و مکانیزم توجه متقابل است، در فرمول (۱۲) نشان داده شده است که ۱۸.۳M Flops/Block برای لایه‌ها و۷۳۷K Flops/Block برای مکانیزم توجه است. (مکانیزم توجه متقابل روی تعداد محدودی توکن در حال اجرا است و بار محاسباتی بالا در زمان تست ندارد و فقط حدود چهار درصد از کل Flops را تشکیل می‌دهد.)

$$Flops \approx (2 \times N \times D \times B)_{MLP} + (N^2 \times D)_{Attn} \quad (۱۲)$$

تجزیه‌وتحلیل یافته‌ها

در جدول (۲)، نتایج از a تا d برچسب‌گذاری شده‌اند که هر برچسب مربوط به یک پارامتر خاص است: a نشان دهنده فاصله (بر حسب متر) است؛ b نشان دهنده ارتفاع (بر حسب متر) است؛ c مربوط به زاویه افقی (بر حسب درجه) است؛ d نشان دهنده زاویه ارتفاع (آن هم بر حسب درجه) است. در مدل ارائه شده در مقاله پایه پایگاه داده (Jekateryńczuk, Szadkowski & Piotrowski, ۲۰۲۵)، بهترین عملکرد با استفاده از ویژگی‌های طیف‌نگار لگاریتمی مل و پارامترهای نرخ نمونه‌برداری ۴۴/۱ kHz، طول تبدیل فوریه سریع ۱۰۲۴، طول گام ۵۱۲ و تعداد فیلترهای مل ۱۲۸ به دست آمده است. درحالی‌که در مدل پیشنهادی ما، از ویژگی‌های طیف‌نگار لگاریتمی مل با رزولوشن بالاتر یعنی پارامترهای نرخ نمونه‌برداری ۹۶ kHz، طول تبدیل فوریه سریع ۴۰۹۶، طول گام ۱۰۲۴ و تعداد فیلترهای مل ۲۵۶ استفاده شده است.

نرخ نمونه‌برداری بالاتر و اندازه طول تبدیل فوریه سریع بزرگ‌تر همراه با تعداد فیلترهای مل بیشتر در مدل پیشنهادی باعث افزایش رزولوشن ویژگی‌ها و تمایز بهتر نشانه‌های زمانی و طیفی می‌شود. مدل پیشنهادی در پارامترهای مرتبط با فاصله و زاویه عملکرد بهتری دارد و کاهش MAE وRMSE نشان‌دهنده مکان‌یابی دقیق‌تر و مقاوم‌تر منبع صوتی پهپاد است. خطاهای گزارش‌شده را در جدول (۲) مشاهده می‌کنیم.

جدول (۲) مقایسه ارزیابی عملکرد مدل پایه با مدل پیشنهادی این مقاله با مقادیر MAE و RMSE

model	MAE				RMSE			
	a	b	c	d	a	b	c	d
مدل پایه	۰/۳۹	۰/۴۱	۰/۹۱	۸/۶۱	۰/۵۸	۰/۵۵	۱/۳۶	۲۶/۸۸
مدل پیشنهادی	**۰/۰۵**	**۰/۰۶**	**۰/۱۴**	**۸/۵۱**	**۰/۱۹**	**۰/۴۷**	**۰/۶۲**	**۲۶/۷**

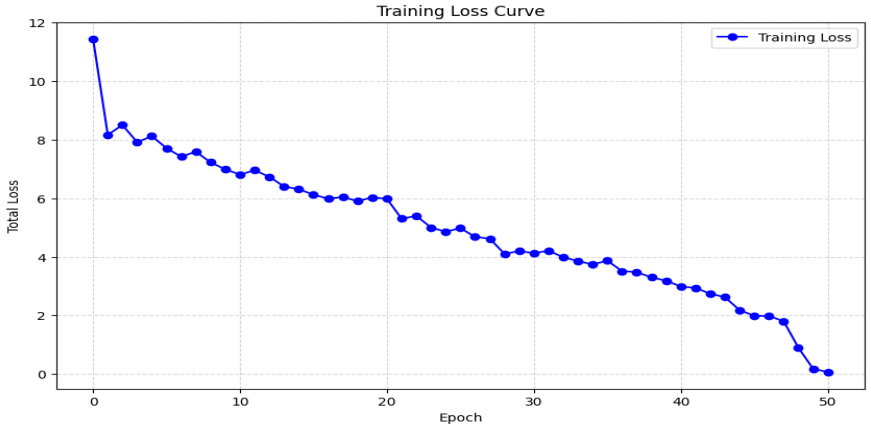
برای فاصله (پارامتر a) در مدل پایه MAE: ۰/۳۹ :وRMSE: ۰/۵۸ :و

در مدل پیشنهادی ما MAE: ۰/۰۵۱ :و RMSE: ۰/۱۹۲ :گزارش شده است که کاهش حدود ۸۷ درصد در MAE و ۶۷ درصد در RMSE که نشان‌دهنده تخمین بسیار دقیق فاصله است را مشاهده می‌کنیم. همچنین برای ارتفاع (پارامتر b) در مدل پایه MAE: ۰/۴۱ :و RMSE: ۰/۵۵ :و در مدل پیشنهادی ما MAE: ۰/۰۶۹ :وRMSE: ۰/۴۷۸ :است که کاهش قابل توجه ۸۳ درصد در MAE و کاهش متوسط ۱۳ درصد در RMSE را داریم. برای زاویه افقی (پارامتر c)، مدل پایه MAE: ۰/۹۱ :و RMSE: ۱/۳۶ :را در مقابل مدل پیشنهادی ما با MAE: ۰/۱۴۳ :و RMSE: ۰/۶۲۸ :دارد که نمایانگر کاهش ۸۴ درصد درMAE و ۵۴ درصد در RMSE و تخمین دقیق‌تر زاویه افقی است. همچنین برای زاویه ارتفاع (پارامتر d)، مدل پایه مقادیر MAE: ۸/۶۱ :و RMSE: ۲۶/۸۸ :را گزارش نموده که در مقابل، مدل ما مقادیر MAE: ۸/۵۱۴ :و RMSE: ۲۶/۷۰۸ :را دارا است. تغییرات ناچیز نشان می‌دهد تخمین زاویه ارتفاع هنوز چالش‌برانگیز است و ممکن است به دلیل ابهام‌های ذاتی در داده‌ها باشد.

از دلایلی که تخمین زاویه ارتفاع هم در مدل پایه و هم در مدل ما (با وجود تخمین بسیار دقیق در موارد دیگر) همچنان دقت بالایی ندارد، عمدتاً به دلیل محدودیت‌های هندسی، فیزیکی و طراحی آرایه میکروفون است. برخلاف زاویه افقی که از نشانه‌های افقی قوی بین میکروفونی بهره می‌برد، نشانه‌های ارتفاع ذاتاً ضعیف‌تر هستند زیرا اکثر آرایه‌های میکروفونی در یک صفحه عمدتاً افقی چیده شده‌اند و دیافراگم عمودی

محدودی را فراهم می‌کنند. این خط پایه عمودی کوچک تنها تفاوت‌های جزئی در تأخیر زمانی و دامنه ایجاد می‌کند، زمانی که منبع صدا در بالا یا پایین آرایه حرکت می‌کند و تخمین ارتفاع را به‌شدت به نویز و طنین حساس می‌کند. محیط‌های بیرونی از طریق بازتاب‌های زمینی و اثرات جوی که تفاوت‌های فاز عمودی را شدیدتر از تفاوت‌های افقی تحریف می‌کنند، پیچیدگی بیشتری ایجاد می‌کنند. در نتیجه، درحالی‌که تخمین زاویه افقی و فاصله می‌تواند به گرادیان‌های مکانی قوی و اطلاعات تأخیر زمانی پایدار متکی باشد، تخمین زاویه ارتفاع تحت تأثیر تفاوت‌های ضعیف بین کانال‌ها قرار می‌گیرد. درنتیجه، حتی زمانی که پیش‌بینی‌های زاویه افقی، فاصله یا ارتفاع قابل اعتماد باشند، دقت زاویه ارتفاع در مکان‌یابی صوتی سه‌بعدی پهپاد معمولاً محدود باقی می‌ماند.

همان‌طور که در شکل ۵ مشاهده می‌شود، روند تغییرات تابع خطا در طول دوره‌های آموزشی نشان داده شده است. این نمودار، کاهش کلی خطا را با وجود نوسانات جزئی بین دوره‌ها نمایش می‌دهد و روند یادگیری مدل پیشنهادی را به‌طور واضح نشان می‌دهد.



شکل (۵) منحنی تغییرات خطا در روند آموزش مدل

این سطح از دقت در مکان‌یابی صوتی پهپاد از منظر نظامی اهمیت بسزایی دارد. شناسایی موقعیت پهپادها در میدان نبرد یا مناطق نظارتی بدون نیاز به رادار و صرفاً بر پایه سیگنال‌های صوتی، می‌تواند به‌عنوان یک سامانه مکمل یا جایگزین در شرایط جنگ الکترونیک عمل کند؛ جایی که رادارها ممکن است دچار اختلال شوند. مدل پیشنهادی

با کاهش خطا در برآورد فاصله و زاویه، امکان رهگیری بلادرنگ پهپاد را فراهم کرده و می‌تواند نقش مهمی در تقویت سامانه‌های دفاعی و هشدار زودهنگام ایفا کند.

نتیجه‌گیری و پیشنهادها

یافته‌های این پژوهش نشان داد که مدل پیشنهادی مبتنی بر معماری Mamba توانسته است در مقایسه با روش‌های پایه، دقت بالاتری در مکان‌یابی سه‌بعدی پهپادها ارائه دهد. به‌ویژه کاهش قابل‌توجه خطای تخمین در پارامترهای فاصله و زاویه نشان می‌دهد که این مدل، توانایی بالایی در رهگیری دقیق و بلادرنگ پهپادها حتی در شرایط نویزی دارد. درحالی‌که پژوهش‌های گذشته بیشتر بر تخمین دوبعدی (زاویه افقی و ارتفاع) متمرکز بوده‌اند، این مطالعه با ارائه یک مدل جامع برای مکان‌یابی سه‌بعدی (فاصله، ارتفاع، زاویه افقی و زاویه ارتفاعی) گامی فراتر نهاده و شکاف مهمی در ادبیات را پر کرده است.

نوآوری اصلی این پژوهش در بهره‌گیری همزمان از ویژگی‌های زمانی و طیفی و ادغام آن‌ها در معماری Mamba است؛ رویکردی که علاوه بر شبیه‌سازی توانایی ادراکی انسان در شنود، پیچیدگی محاسباتی پایین‌تر و تعمیم‌پذیری بالاتری نسبت به ترنسفورمرها دارد. مقایسه نتایج با پژوهش‌های اخیر نشان می‌دهد که مدل حاضر علاوه بر افزایش دقت در تخمین فاصله، ازنظر کارایی محاسباتی نیز مزیت چشمگیری دارد و قابلیت پیاده‌سازی در سامانه‌های عملیاتی دفاعی را داراست.

باوجود این دستاوردها، پژوهش حاضر محدودیت‌هایی نیز دارد. نخست آنکه اکثر داده‌های آموزشی مورد استفاده در شرایط کنترل‌شده جمع‌آوری شده‌اند و ممکن است در محیط‌های شهری با نویزهای متغیر، بازتاب‌های پیچیده یا شرایط جوی متنوع، کارایی مدل کاهش یابد. همچنین، پایگاه داده‌های دیگر در شرایطی جمع‌آوری شده‌اند که فاصله پهپاد از آرایه میکروفونی کمتر از این پایگاه داده مورد استفاده است. این موضوع باعث می‌شود تخمین پارامترهایی مانند فاصله و زاویه ساده‌تر باشد و دقت مدل‌ها به‌طور مصنوعی بالاتر گزارش شود. در مقابل، پایگاه داده مورد استفاده در این پژوهش شامل نمونه‌هایی با فواصل و شرایط متنوع‌تر است که ارزیابی مدل پیشنهادی را واقع‌بینانه‌تر و چالش‌برانگیزتر می‌سازد. دوم آنکه آزمایش‌ها بر روی چند نوع پهپاد خاص انجام شده و

برای اطمینان از تعمیم‌پذیری، لازم است عملکرد مدل روی انواع مختلف پهپاد با اندازه‌ها و الگوهای صوتی متفاوت بررسی شود.

برای مطالعات آینده پیشنهاد می‌شود:

۱. توسعه و ارزیابی مدل بر روی مجموعه‌داده‌های متنوع‌تر شامل شرایط واقعی شهری و میدان‌های نظامی.
۲. بررسی روش‌های تقویت داده برای افزایش مقاومت مدل در برابر نویزهای غیرایستا.
۳. ترکیب سامانه صوتی با حسگرهای دیگر (چشم‌انداز چندوجهی شامل رادار کم‌قدرت یا سیستم‌های اپتیکی و دوربین‌های حرارتی) برای بهبود دقت شناسایی، کاهش نرخ خطا، افزایش پایداری در شرایط محیطی متغیر (مانند مه، تاریکی یا نویز بالا) و ارتقای قابلیت اعتماد سامانه در کاربردهای عملیاتی.
٤. بهینه‌سازی مدل برای اجرا در سامانه‌های سبک و قابل حمل به‌منظور استفاده عملیاتی در میدان نبرد.

کاربرد نتایج این تحقیق برای مدیران و سیاست‌گذاران دفاعی بسیار حائز اهمیت است. دقت بالاتر در مکان‌یابی صوتی پهپادها می‌تواند نقش مکملی برای سامانه‌های راداری و اپتیکی داشته باشد، به‌ویژه در شرایط جنگ الکترونیک یا محیط‌های با دید محدود. این نوآوری می‌تواند زمینه‌ساز توسعه سامانه‌های هشدار سریع، رهگیری خودکار و دفاع هوایی هوشمند باشد.

توصیه‌های کلیدی برای سیاست‌گذاران دفاعی

استفاده از سامانه‌های مکان‌یابی صوتی به‌عنوان مکمل رادار و اپتیک برای شناسایی پهپادهای متخاصم.

۲- سرمایه‌گذاری در توسعه و بومی‌سازی پایگاه‌های داده صوتی پهپاد در شرایط واقعی میدان نبرد

۳- حمایت از پژوهش‌های میان رشته‌ای در حوزه هوش مصنوعی و پردازش سیگنال برای ارتقای دقت و کارایی مدل‌ها

۴- طراحی سامانه‌های سبک و قابل حمل مبتنی بر مدل پیشنهادی این مقاله برای استفاده در مناطق مرزی و نقاط حساس

۵- بهره‌گیری از سامانه‌های صوتی مقاوم به نویز در چارچوب دفاع هوایی چندلایه برای افزایش سطح امنیت ملی.

تشکر و قدردانی

نویسندگان این مقاله از تمامی افرادی که در تهیه و انتشار این مقاله به‌ویژه سردبیر محترم نشریه و داوران این مقاله مؤثر بوده‌اند قدردانی می‌نمایند.

تضاد منافع:

بدین‌وسیله نویسنده تصریح می‌نماید که هیچ گونه تضاد منافعی در خصوص پژوهش حاضر وجود ندارد.

منابع

- چالاکی، محمدباقر؛ احمدزاده فرد، محمدحسن؛ رجب پور، مجید. (۱۴۰۳). به‌کارگیری پهپاد در مأموریت کشف نیروی پدافند هوایی ارتش ج.ا.ا، *فصلنامه مطالعات جنگ*، ۶(۲۲)، ۱-۲۴.
- حبیبی، نیک بخش. (۱۳۹۶). ارائه مدل اثربخش به‌کارگیری بهینه پهپاد در توانمندسازی عملیات آینده سازمان‌های دفاعی (مطالعه موردی عملیات پروازی نیروی هوایی)، فصلنامه آینده‌پژوهی دفاعی، ۲(۴), ۳۵-۶۲.

- Chen, H. & Ser, W. (2011). Sound source DOA estimation and localization in noisy reverberant environments using least-squares support vector machines. Journal of Signal Processing Systems, 63(3), 287-300. (DOI: https://doi.org/10.1007/s11265-009-0423-7)
- Chen, J. C. Yao, K. & Hudson, R. E. (2003). Acoustic source localization and beamforming: theory and practice. EURASIP journal on advances in signal processing, 2003(4), 926837. (DOI: https://doi.org/10.1155/S1110865703212038)
- Chalaki, M. ahmadzadeh fard, M. H. and rajabpour, J. (2024). The use of UAV in the Detection mission of the Air Defense Force of the Islamic Republic of Iran. War Studies, 6(22), 1-24. [in Persian] (DOI: https://doi.org/10.22034/qjws.2024.2026774.1203)
- Chung, M. A. Chou, H. C. & Lin, C. W. (2022). Sound localization based on acoustic source using multiple microphone array in an indoor environment. Electronics, 11(6), 890. (DOI: https://doi.org/10.3390/electronics11060890)
- Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously

spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4), 357-366. (DOI: https://doi.org/10.1109/TASSP.1980.1163420)

- Gadre, C. M. Patole, R. K. & Metkar, S. P. (2023, September). Comparative analysis of KNN and CNN for Localization of Single Sound Source. In 2023 International Conference on Network, Multimedia and Information Technology (NMITCON) (pp. 1-6). IEEE. (DOI: https://doi.org/10.1109/NMITCON58196.2023.10275895)
- Gombots, S. Nowak, J. & Kaltenbacher, M. (2021). Sound source localization–state of the art and new inverse scheme. e & i Elektrotechnik und Informationstechnik, 138(3), 229-243. (DOI: https://doi.org/10.1007/s00502-021-00881-6)
- Gu, A. & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752. (DOI: https://doi.org/10.48550/arXiv.2312.00752)
- Habibi, Nikbakhsh. (2017). Presenting an effective model for optimal utilization of unmanned aerial vehicles in empowering future operations of defense organizations (Case study: flight operations of the Air Force). Defense Futures Studies, 2(4), 35-62. [in Persian] (URL: https://www.dfsr.ir/article_30716.html)
- Hu, F. Song, X. He, R. & Yu, Y. (2023). Sound source localization based on residual network and channel attention module. Scientific Reports, 13(1), 5443. (DOI: https://doi.org/10.1038/s41598-023-32657-7)
- Jekateryńczuk, G. & Piotrowski, Z. (2023). A survey of sound source localization and detection methods and their applications. Sensors, 24(1), 68.(DOI: https://doi.org/10.3390/s24010068)
- Jekateryńczuk, G. Szadkowski, R. & Piotrowski, Z. (2025). UaVirBASE: A Public-Access Unmanned Aerial Vehicle Sound Source Localization Dataset. Applied Sciences, 15(10), 5378. (DOI: https://doi.org/10.3390/app15105378)
- Khan, A. Waqar, A. Kim, B. & Park, D. (2025). A review on recent advances in sound source localization techniques, challenges, and applications. Sensors and Actuators Reports, 100313. (DOI: https://doi.org/10.1016/j.snr.2025.100313)
- Luo, Z. Lu, B. Huang, J. Ran, C. & He, H. (2023). Sound source direction-of-arrival estimation method for microphone array based on ultra-weak fiber Bragg grating distributed acoustic sensor. Optics Express, 31(19), 31342-31353. (DOI: https://doi.org/10.1364/OE.498027)
- Ma, S. Wang, J. Abbas, S. Ding, X. & Tu, X. (2025, July). Self-supervised Sound Source Localization for UAVs Using GCC-PHAT in

Low SNR Environments. In International Conference on Intelligent Computing (pp. 498-510). Singapore: Springer Nature Singapore. (DOI: https://doi.org/10.1007/978-981-96-9894-3_41)

- Mu, D. Zhang, Z. Yue, H. Wang, Z. Tang, J. & Yin, J. (2024). Seld-mamba: Selective state-space model for sound event localization and detection with source distance estimation. arXiv preprint arXiv:2408.05057. (DOI: https://doi.org/10.48550/arXiv.2408.05057)
- Ning, Y. M. Ma, S. Meng, F. Y. & Wu, Q. (2020). DOA estimation based on ESPRIT algorithm method for frequency scanning LWA. IEEE Communications Letters, 24(7), 1441-1445. (DOI: https://doi.org/10.1109/LCOMM.2020.2988020)
- Perotin, L. Serizel, R. Vincent, E. & Guérin, A. (2018, September). CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector. In 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC) (pp. 241-245). (IEEE. DOI: https://doi.org/10.1109/IWAENC.2018.8521403)
- Qayyum, A. B. A. Hassan, K. N. Anika, A. Shadiq, M. F. Rahman, M. M. Islam, M. T. ... & Haque, M. A. (2020). DOANet: a deep dilated convolutional neural network approach for search and rescue with drone-embedded sound source localization. EURASIP Journal on Audio, Speech, and Music Processing, 2020(1), 16. (DOI: https://doi.org/10.1186/s13636-020-00184-2)
- Ruiz-Espitia, O. Martinez-Carranza, J. & Rascon, C. (2018, June). AIRA-UAS: an evaluation corpus for audio processing in unmanned aerial system. In 2018 International Conference on Unmanned Aircraft Systems (ICUAS) (pp. 836-845). IEEE. (DOI: https://doi.org/10.1109/ICUAS.2018.8453466)
- Salvati, D. Drioli, C. & Foresti, G. L. (2016). A weighted MVDR beamformer based on SVM learning for sound source localization. Pattern Recognition Letters, 84, 15-21. (DOI: https://doi.org/10.1016/j.patrec.2016.07.003)
- Schmidt, R. (1986). Multiple emitter location and signal parameter estimation. IEEE transactions on antennas and propagation, 34(3), 276-280. (DOI: https://doi.org/10.1109/TAP.1986.1143830)
- Song, X. Qin, Q. Wang, S. Yao, F. Qiu, H. Wang, M. & Jiang, H. (2025). Embedding and Beamforming Network for Sound Source Localization in Spherical Harmonic Domain. IEEE Sensors Journal. (DOI: https://doi.org/10.1109/JSEN.2025.3595385)
- Strauss, M. Mordel, P. Miguet, V. & Deleforge, A. (2018, October). DREGON: Dataset and methods for UAV-embedded sound source localization. In 2018 IEEE/RSJ International Conference on Intelligent

Robots and Systems (IROS) (pp. 1-8). IEEE. (DOI: https://doi.org/110.1109/IROS.2018.8593581)

- Takeda, R. & Komatani, K. (2017, March). Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2217-2221). IEEE. (DOI: https://doi.org/10.1109/ICASSP.2017.7952550)
- Tang, H. (2014). DOA estimation based on MUSIC algorithm.
- Wang, K. & Zhang, M. (2024). Sound source localization system based on TDOA algorithm. In Intelligent Computing Technology and Automation (pp. 350-356). IOS Press. (DOI: https://doi.org/10.3233/ATDE231207)
- Wang, L. Sanchez-Matilla, R. & Cavallaro, A. (2019, November). Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 5320-5325). IEEE. (DOI: https://doi.org/10.1109/IROS40897.2019.8968183)
- Xiao, Y. & Das, R. K. (2024). Tf-mamba: A time-frequency network for sound source localization. arXiv preprint arXiv:2409.05034. (DOI: https://doi.org/10.48550/arXiv.2409.05034)
- Xu, K. Zong, Z. Liu, D. Wang, R. & Yu, L. (2025). Deep Learning-Based Sound Source Localization: A Review. Applied Sciences, 15(13), 7419. (DOI: https://doi.org/10.3390/app15137419)