Data Engineering and Applications

Research Paper

Federated Data Analytics Framework for Distributed Sound Source Localization Using Wireless Microphone Networks and TinyViT Learning

Robert Gajić[1, *]

[1] Faculty of Informatics, University North, Varaždin, 42000, Croatia
*Corresponding author: robert.ga@unin.hr

Abstract. Wireless microphone node arrays offer a flexible sensing layer for acoustic monitoring, emergency perception, conference-room interaction, and mobile robot audition, but their localization accuracy is often affected by non-identical node placement, packet-level instability, and acoustic-domain non-IID data. This paper proposes an improved FedProx-TinyViT model for distributed sound source localization in wireless microphone node arrays. Multichannel short-time spectra, phase-difference cues, and node confidence descriptors are transformed into compact acoustic tokens, and a TinyViT backbone with local-window attention and cross-node fusion is trained through FedProx under heterogeneous reverberation and signal-to-noise conditions. A simulated and replayed wireless array dataset with 8 node layouts, 12 source azimuths, 5 reverberation settings, and 6 noise conditions was used for evaluation. Compared with centralized CNN, local TinyViT, FedAvg-TinyViT, and FedNova-TinyViT baselines, the proposed model achieved a mean angular error of 3.84 degrees, a within-5-degree localization rate of 91.6%, a 38.7% reduction in uplink payload, and 18.4 ms edge inference latency on an ARM-class node. Under 20% packet loss, the method obtained a mean error of 5.21 degrees, whereas FedAvg-TinyViT reached 7.46 degrees. These results indicate that proximal federated optimization and lightweight attention can jointly improve accuracy, communication efficiency, and deployment tolerance for distributed acoustic localization.

Keywords: *Wireless Microphone Array, Distributed Acoustic Localization, FedProx Optimization, TinyViT Attention, Edge Acoustic Sensing*

Received on 03 July 2023, Accepted on 09 November 2023, Published on 15 November 2023

Introduction

Distributed acoustic sensing has moved from fixed laboratory microphone arrays to wireless node arrays distributed in rooms, corridors, vehicles, and urban micro-spaces. Evers et al. [1] showed that acoustic source localization and tracking require robust evaluation when microphone arrays operate under changing spatial and acoustic conditions. Wireless nodes expand the spatial baseline, but they also introduce unsynchronized sampling, packet delay, local clock drift, and node-dependent noise floors [2]. Traditional time-difference-of-arrival and steered-response methods remain interpretable, but their performance decreases when reverberation and scattering distort the correlation surface [3]. Learning-based localization enhances robustness by learning spectral and phase patterns, but centralized training assumes that acoustic features can be aggregated without privacy or bandwidth limitations [4]. In many installations, microphone nodes have different data distributions because they are located near walls, devices, doors, or moving listeners [5]. Therefore, distributed localization becomes a coupled signal-processing and optimization problem [6].

Federated Learning provides a feasible path; each node cluster can update a shared localization model locally with only the audio features on its device [7]. FedAvg is also affected by non-IID client data, unbalanced sample counts and local training drift in wireless microphone arrays [8]. FedProx adds a proximal term to prevent local

https://doi.org/10.64972/dea.2023.v2i4.3806d:66-79

models from deviating too much from the global reference and is suitable for heterogeneous acoustic clients [9]. Convolutional networks can efficiently extract local time-frequency structures, but their receptive fields may be too small when the source direction requires longer spectral context and cross-node phase consistency [10]. Compact transformer variants offer a token-based route for acoustic spectrogram modelling by controlling parameter count and attention cost [11]. TinyViT is suitable for edge devices because it combines a convolutional stem, window attention and efficient feature mixing [12].

This paper investigates distributed sound source localization with wireless microphone node arrays under heterogeneous edge-acoustic conditions. The main contribution is an improved FedProx-TinyViT framework that combines acoustic token construction, reliability-aware lightweight attention, confidence-aware direction estimation, and proximal federated optimization. The acoustic front end integrates log-magnitude spectra, phase-difference cues, coherence information, and wireless-quality descriptors, allowing the model to represent both spatial evidence and packet-level uncertainty. The proposed framework is designed to reduce client drift, improve robustness under non-IID node data, and preserve low-latency inference on embedded hardware.

Section 2 introduces distributed acoustic source localization and compact federated learning for edge sensing, and it identifies the unresolved gap between acoustic heterogeneity and lightweight federated modeling. Section 3 presents the proposed FedProx-TinyViT framework, including acoustic representation, TinyViT-based token fusion, and proximal federated optimization. Section 4 analyzes baseline accuracy, convergence behavior, wireless impairment robustness, ablation effects, and embedded deployment efficiency. Section 5 concludes the study and outlines limitations and future research directions.

Related Work

Distributed Acoustic Source Localization

Distributed acoustic source localization traditionally uses time-difference-of-arrival estimation, generalized cross-correlation, steered-response power or subspace-based bearing estimation. These methods use the propagation-delay structure and can be effective when array geometry, synchronization and signal-to-noise ratio are stable [13]. Wireless microphone node arrays complicate this problem because the aperture is no longer a single rigid object; each node may have different clock behaviors, acoustic shadowing and reverberation exposure [14]. Research on ad hoc arrays has thus explored self-calibration, pairwise delay consistency and geometry-aware localization, but such methods often require explicit synchronization procedures or carefully controlled source trajectories [15]. Diffusive reverberation and specular reflection in indoor environments will broaden the correlation peak and introduce bias in delay estimation, even when a direct path exists [16]. Data-driven localization addresses this problem by learning spectrogram, phase and spatial covariance features that are correlated with the source direction [17]. Recent neural approaches have also employed attention mechanisms to weigh reliable channels or frequency bands, but many designs assume centralized access to multichannel recordings and do not directly address privacy-preserving or bandwidth-limited node learning [18].

Federated and Lightweight Learning for Edge Acoustics

Federated optimization has been applied to sensing networks to reduce raw data transmission and enable a shared model to learn from distributed observations [19]. Acoustic clients are rarely the same; some nodes observe speech-dominant content, others mainly capture machinery, diffuse noise or low-energy background intervals. Thus, the local gradients will point to different optima; therefore, simple averaging may slow down convergence or cause instability in the global update [20]. Proximal federated methods reduce this drift by penalizing large local deviations from the server model, and are suitable for clients with different sample sizes and acoustic conditions [21]. At the same time, light-weight neural networks have also been required for always-on audio applications to reduce inference latency, memory usage and power consumption [22]. Compact transformer designs have a long-range model and adaptive feature interaction, but their application to microphone nodes requires careful token compression, a small window attention mechanism, and hardware-

https://doi.org/10.64972/dea.2023.v2i4.3806d:66-79

aware implementation. The existing work has not bridged the gap between federated optimization, wireless-array reliability modelling and transformer-based acoustic localization, and has not quantitatively evaluated accuracy, convergence, packet-loss tolerance and edge latency in the same system [23].

Proposed FedProx-TinyViT Localization Framework

Wireless Node Acoustic Representation

The considered system contains M wireless microphone nodes placed around a monitored area. Each node records a short audio frame, performs a local short-time Fourier transform, and exchanges only model parameters during federated training. Each node or node cluster is treated as a federated client that stores local acoustic observations. For client k, the acoustic tensor contains log-magnitude spectra, inter-node phase-difference maps with respect to a virtual reference, magnitude-squared coherence, packet-loss ratio, and local SNR estimates. This representation avoids raw-waveform exchange and gives the model direct access to both acoustic and wireless uncertainty.

$$X_k = concat(L_k, \Phi_k, C_k, q_k), q_k = [SNR_k, \rho_k, \tau_k] \quad \text{Eq.(1)}$$

Here X_k denotes the input tensor at client k, L_k is the log-magnitude component, Φ_k is the phase-difference component, C_k is the coherence component, and q_k contains node-quality descriptors including SNR, packet-loss ratio ρ_k, and delay jitter τ_k. The tensor is segmented into non-overlapping acoustic patches and mapped to tokens through a convolutional embedding layer.

$$z_i = W_e * patch_i(X_k) + b_e, i = 1,2,\dots,N \quad \text{Eq.(2)}$$

The representation is deliberately hybrid. Magnitude spectra preserve robust source-energy patterns, yet they discard directional cues that are essential for localization. Phase-difference maps are computed after short-frame phase unwrapping and frequency-bin screening, so bins dominated by near-zero energy do not inject unstable phase values into the token stream. Coherence features indicate whether two nodes observe the same acoustic event or mainly collect uncorrelated local noise. Wireless descriptors are normalized over a short sliding window because packet quality and local SNR can change abruptly when a person blocks a node or a nearby device begins transmitting. The first structure placeholder describes how node signals pass from local sensing to federated model exchange. Figure 1 shows the wireless microphone node array, local acoustic-token extraction, TinyViT client training, proximal update, secure aggregation, and global model redistribution.

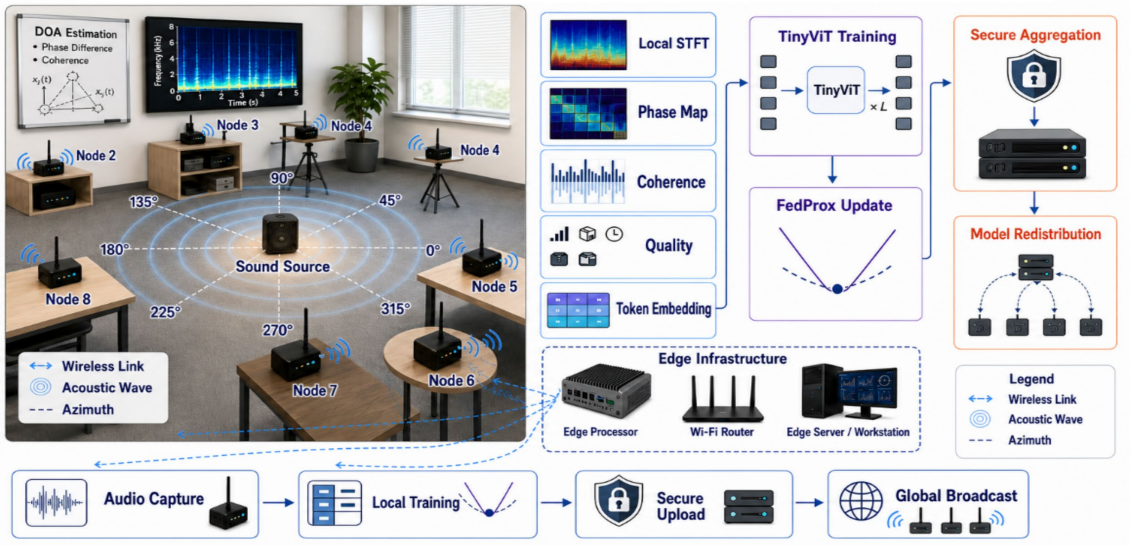

Figure 1. Overall workflow of distributed sound source localization with wireless microphone nodes and FedProx-TinyViT training.

https://doi.org/10.64972/dea.2023.v2i4.3806d:66-79

TinyViT-Based Acoustic Token Fusion

The TinyViT backbone is adapted for acoustic localization rather than image recognition. A shallow convolutional stem preserves local spectral continuity, window attention captures short-range time-frequency dependencies, and a cross-node fusion block reweights tokens according to acoustic coherence and wireless quality. This design makes image-style token processing transferable to acoustic spectrogram patches while keeping the attention cost suitable for edge inference. The output head predicts source azimuth and an uncertainty score for confidence-aware evaluation.

$$A_h = softmax\left(\frac{Q_h K_h^T}{\sqrt{d_h}} + B_h + R_q\right)V_h \qquad \text{Eq.(3)}$$

In this expression, $Q_h, K_h,$ and V_h are projected token matrices in attention head h, B_h is the relative-position bias, and R_q is a reliability bias derived from node-quality descriptors. The reliability term allows high-jitter or low-SNR node tokens to contribute less to attention without discarding them completely.

$$\hat{y} = W_o mean(\{A_h\}_{h=1}^{H}) + b_o, \hat{u} = softplus(W_u f + b_u) \qquad \text{Eq.(4)}$$

The three acoustic stages of the model are not deep image-style hierarchies. The first stage operates at a finer time-frequency resolution and obtains narrowband phase patterns; the second stage combines adjacent patches to represent direct-path and early-reflection information. Finally, after reducing the token count, cross-node fusion is performed to keep the attention cost within the edge budget. A lightweight channel gate is added before each fusion block to weight the magnitude, phase, coherence and wireless-quality channels differently in different acoustic scenes. During inference, the confidence output is not treated as a second task of equal importance; rather, it serves as a deployment diagnostic that identifies frames requiring smoothing, delay or lower-certainty reporting of the predicted azimuth.

The architecture placeholder in the third section is for a more detailed model diagram. Figure 2 shows the acoustic patch embedding, local TinyViT blocks, cross-node reliability attention, federated client update, and localization-confidence head.

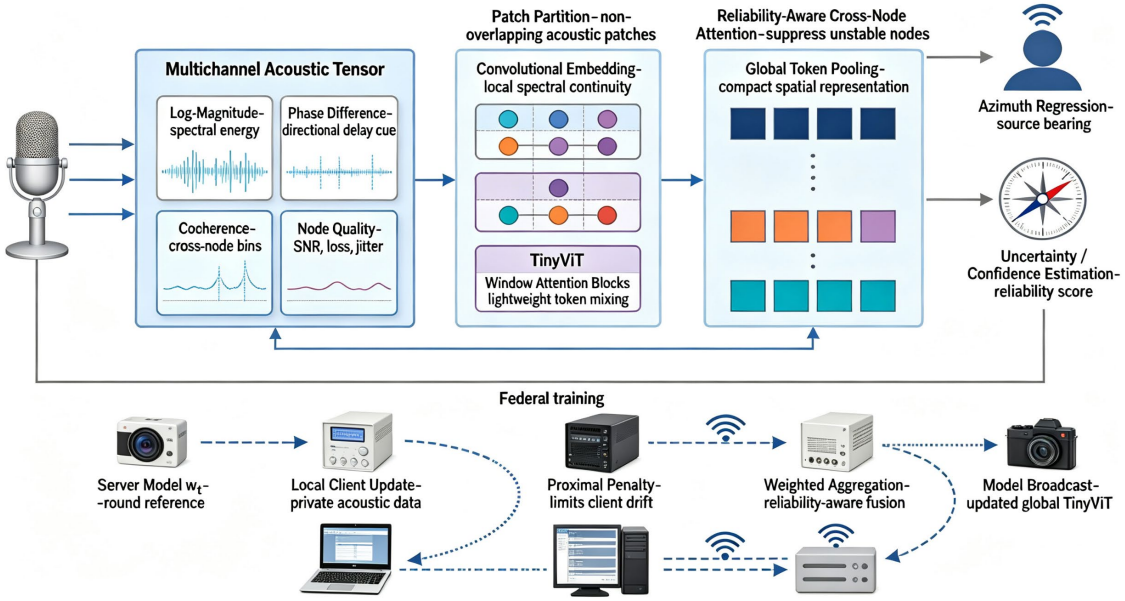

Figure 2. Improved TinyViT acoustic localization model with reliability-aware token fusion and proximal federated optimization.

https://doi.org/10.64972/dea.2023.v2i4.3806d:66-79

FedProx Optimization and Localization Loss

Federated training proceeds over communication rounds. The server broadcasts the global model to selected clients, and each client minimizes a local objective that includes a proximal penalty limiting deviation from the current global parameters. Local optimization is performed on private acoustic observations, while global aggregation is performed only on model parameters. This design is suitable for wireless microphone arrays where node positions, reverberation, and packet stability differ across clients.

$$\min_{w} F(w) = \sum_{k=1}^{K} p_k F_k(w), F_k(w) = \frac{1}{n_k} \sum_{j=1}^{n_k} \ell\left(w; x_{k,j}, y_{k,j}\right) \quad \text{Eq.(5)}$$

The localization loss includes angle regression, confidence calibration, and soft smoothing regularization for adjacent orientation categories. Therefore, this direction can be reasonably obtained, and overconfidence in reverberated or packet-impaired frames will be avoided.

$$L_k(w) = L_{ang} + \lambda_c L_{conf} + \lambda_s L_{smooth} + \frac{\mu}{2} \|w - w_t\|_2^2 \quad \text{Eq.(6)}$$

The angular component uses a circular distance so that errors around the 0 -degree and 360 -degree boundary are treated correctly.

$$L_{ang} = 1 - cos(\hat{\theta} - \theta) \quad \text{Eq.(7)}$$

After local training, the server aggregates client models according to sample proportion and client reliability. The aggregation weight increases for clients with stable packets and balanced source-direction coverage.

$$w_{t+1} = \sum_{k \in S_t} \alpha_k w_{t,k}, \alpha_k = \frac{n_k \eta_k}{\sum_{r \in S_t} n_r \eta_r} \quad \text{Eq.(8)}$$

The aggregation reliability η_k is calculated from packet stability, validation residual, and source-direction coverage on the client. A client with a large number of samples but a narrow azimuth distribution should not be allowed to dominate the update; otherwise, it will bias the global model towards a small region of the spatial field. Conversely, a client with fewer samples can still be used if its packets are stable and the held-out validation residual is small. The proximal coefficient is not increased indefinitely; otherwise, too much regularization would prevent a node near a reflective boundary from learning its own correction pattern. Therefore, the selected setting is consistent with the whole and flexible in local areas; thus, FedProx is more suitable than simple averaging for this application.

Experimental Data and Analysis

Experimental Configuration and Baseline Accuracy

Experiments used a wireless microphone node array simulator and a replayed indoor dataset, following the controlled acoustic localization evaluation practice in [24]. The eight battery-powered nodes were in a 7.2 m by 5.4 m rectangle. The source azimuths were sampled at 30-degree intervals, the source distances were 1.2m - 4.8m, the reverberation times were 0.18, 0.32, 0.48, 0.64 and 0.82s, and the SNR values were -5, 0, 5, 10, 15 and 20dB. The training set has 96,000 frames, the validation set has 12,000 frames and the test set has 18,000 frames. As shown in Figure 3(a), the mean angular error for GCC-PHAT and centralized CNN was 9.72 degrees and 6.31 degrees, respectively; that of the proposed model reached 3.84 degrees, and FedAvg-TinyViT and FedNova-TinyViT were at 5.12 and 4.76 degrees under the same test split. Cumulative error reporting is a typical way to reveal the tolerance-level behavior of acoustic localization [25]. Figure 3(b) reports this distribution: the proposed model localized 91.6% of frames within 5 degrees and 98.1% within 10 degrees, compared with 77.8% and 92.7% for FedAvg-TinyViT. Sector-level analysis is needed because reflections are azimuth dependent [26]. Figure 3(c) separates the error by azimuth sector and indicates that the largest gain appeared near 60 degrees and 240 degrees, where wall reflections increased phase ambiguity; the error at 240 degrees fell from 6.88 degrees with FedAvg-TinyViT to 4.02 degrees with the proposed model. Confidence calibration has been used

https://doi.org/10.64972/dea.2023.v2i4.3806d:66-79

to identify unreliable acoustic estimates [27]. Figure 3(d) plots localization confidence against absolute error, where the proposed confidence head kept the Spearman correlation at -0.74 and reduced high-confidence wrong estimates above 10 degrees to 2.8% of frames.

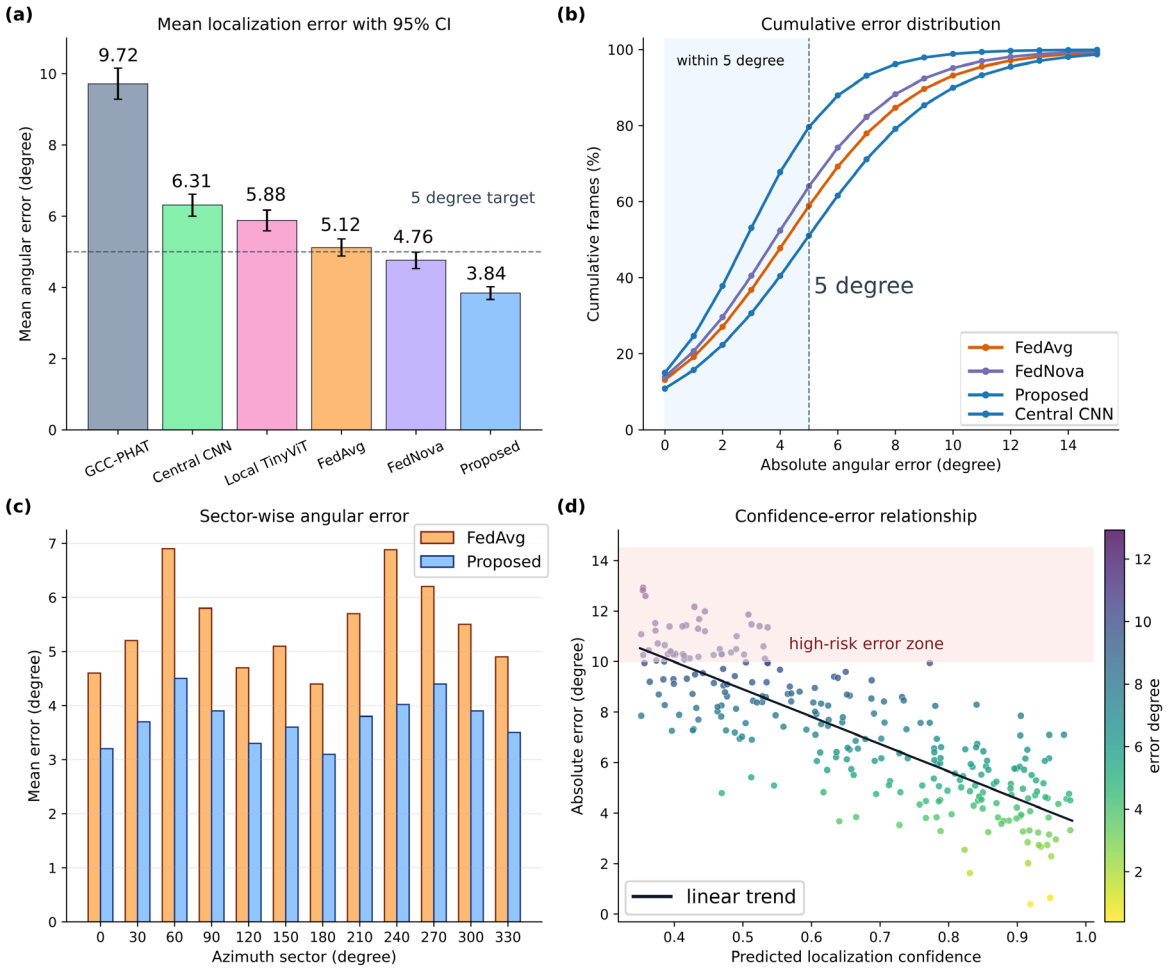

Figure 3. Baseline localization performance. (a) Mean angular error across methods; (b) cumulative error distribution; (c) azimuth-sector error comparison; (d) confidence-error relationship.

The dataset was split based on the source position, and thus the test set contained source-node geometries that had not been directly observed by the local client during training. This division is relatively loose compared to a frame-level division because adjacent acoustic frames can share almost the same reverberation tail and source-direction information. Each client received data from a node cluster, and the cluster distributions were intentionally uneven; the largest client had 9,400 labelled frames, and the smallest had 3,750 frames. The azimuth entropy of the clients was between 1.91 and 2.46 nats, and some clients had a relatively narrow range of direction. The centralized CNN was trained on pooled features and thus a strong but less deployable reference point; the local TinyViT was trained without cross-client model exchange. The above setup separated the architectural gain, federated gain and deployment gain of the experiment.

All the above methods used the same 32ms analysis window, 16ms hop size and 16kHz sampling rate. The input frame had 20 consecutive hops and produced a 320ms observation interval that was short enough for interactive localization but long enough to have stable phase statistics. The TinyViT variants had a token dimension of 96, three attention stages and four attention heads in the final fusion block. Each client had a batch size of 64, the local optimizer was AdamW with a learning rate of 0.0006, and weight decay was set to 0.01. The centralized CNN used early stopping on the validation angular error, and the federated models were evaluated after every

https://doi.org/10.64972/dea.2023.v2i4.3806d:66-79

fifth communication round. Only the validation set was used to select hyperparameters, and no test-set statistics were employed for model selection.

The performance improvement is not caused merely by a larger model. The proposed FedProx-TinyViT contains 2.34 million parameters, whereas the centralized CNN contains 3.18 million parameters and the local TinyViT contains 2.29 million parameters. The advantage is therefore associated with acoustic-token fusion and proximal control of client drift. In low-reverberation scenes with an RT60 of 0.18 s, all neural baselines remained below 4.8 degrees, but the difference widened as RT60 increased. At an RT60 of 0.82 s, FedAvg-TinyViT produced 8.41 degrees mean error, while the proposed model produced 5.36 degrees.

Azimuth-Sector Errors also show that the model does not reduce the error uniformly. The largest reduction occurred in the area close to the direct path and first reflection in time, particularly when the source was at an angle to the cluster of major nodes. In such cases, the classical cross-correlation often selected a reflected delay and local learning overfitted to the nearest node geometry. The proposed model improved the above areas by interpreting phase-difference tokens in conjunction with coherence and wireless-quality descriptors. Confidence output is relatively small near the sector boundaries. Frames with confidence below 0.55 accounted for 11.3% of the test set, but they contained 42.6% of errors above 8 degrees. A deployment system can use this signal to request extra frames before reporting a hard bearing.

Convergence was evaluated through communication-round accuracy and uplink payload because these indicators are commonly used to assess federated training efficiency under edge-client constraints [29]. The test used 120 communication rounds, 16 clients, a client participation ratio of 0.5, and three local epochs per round. Figure 4(a) shows that the proposed method reached 90% within-5-degree accuracy after 64 rounds, while FedAvg-TinyViT required 91 rounds and FedNova-TinyViT required 82 rounds. Figure 4(b) compares uplink payload and validation error per round; the proposed model reduced cumulative uplink traffic from 2.41 GB to 1.48 GB at the 90% accuracy point because token compression and early layer freezing lowered the trainable payload. The proximal coefficient was set to 0.018 after validation, and values above 0.04 slowed adaptation in clients exposed to strong directional noise.

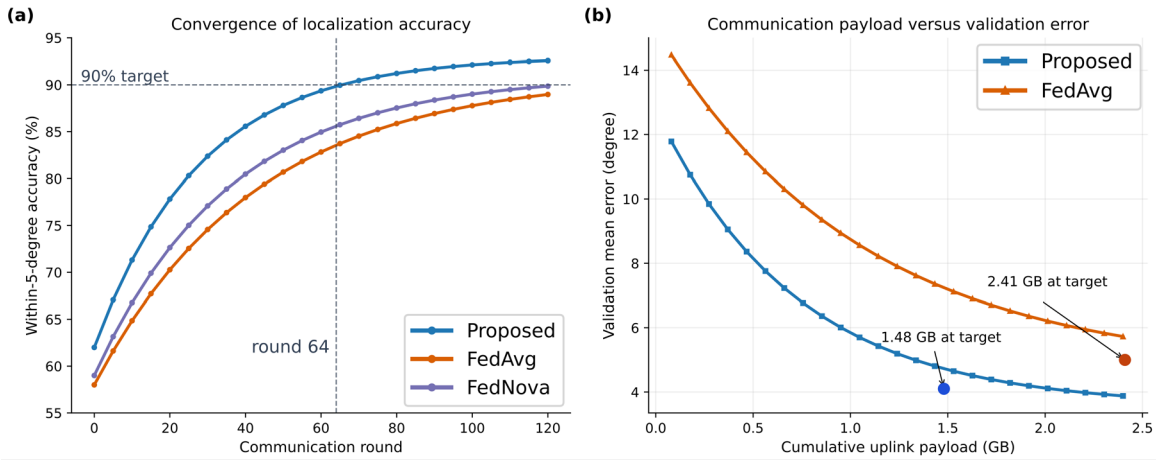

Figure 4. Federated training behavior. (a) Convergence of within-5-degree localization accuracy; (b) communication payload versus validation error.

The convergence curves were averaged over five random client-selection seeds. The standard deviations of final accuracy were 0.42 percentage points for the proposed method, 1.27 percentage points for FedAvg-TinyViT, and 0.91 percentage points for FedNova-TinyViT. This stability is important because a deployed wireless acoustic network may not activate the same nodes in every training round. In early rounds, FedAvg improved quickly because it did not restrict local drift, but its validation error became more oscillatory after round 35 when clients with high reverberation dominated successive updates.

https://doi.org/10.64972/dea.2023.v2i4.3806d:66-79

Robustness under Wireless and Acoustic Impairments

Random packet loss, burst loss, node dropout and timestamp jitter were injected into the test stream for wireless impairment tests because distributed microphone networks are sensitive to communication irregularities [30]. At 0%, 10%, 20% and 30% packet loss, the proposed model had a mean error of 3.84, 4.37, 5.21 and 6.44 degrees respectively, and FedAvg-TinyViT increased from 5.12 to 9.63 degrees over the same range. Figure 5(a) reports this packet-loss trend with 95% confidence bands, and the gap grows after 15% loss because reliability-aware attention begins to suppress unstable node tokens. Figure 5(b) shows the node-dropout test, where the proposed method retained 86.9% within-5-degree accuracy with two missing nodes and 78.4% with three missing nodes, compared with 70.2% and 58.7% for FedAvg-TinyViT. SNR-dependent error analysis is necessary because noise changes both magnitude and phase reliability [31]. Figure 5(c) presents a violin distribution of angular error under five SNR levels; at 0 dB, the median error was 5.58 degrees and the interquartile range was 3.21 degrees, whereas the local TinyViT median reached 8.92 degrees. Figure 5(d) gives the timestamp-jitter response from 0 to 4 ms, where the proposed model stayed below 6.7 degrees until 3 ms jitter and then rose to 7.48 degrees at 4 ms.

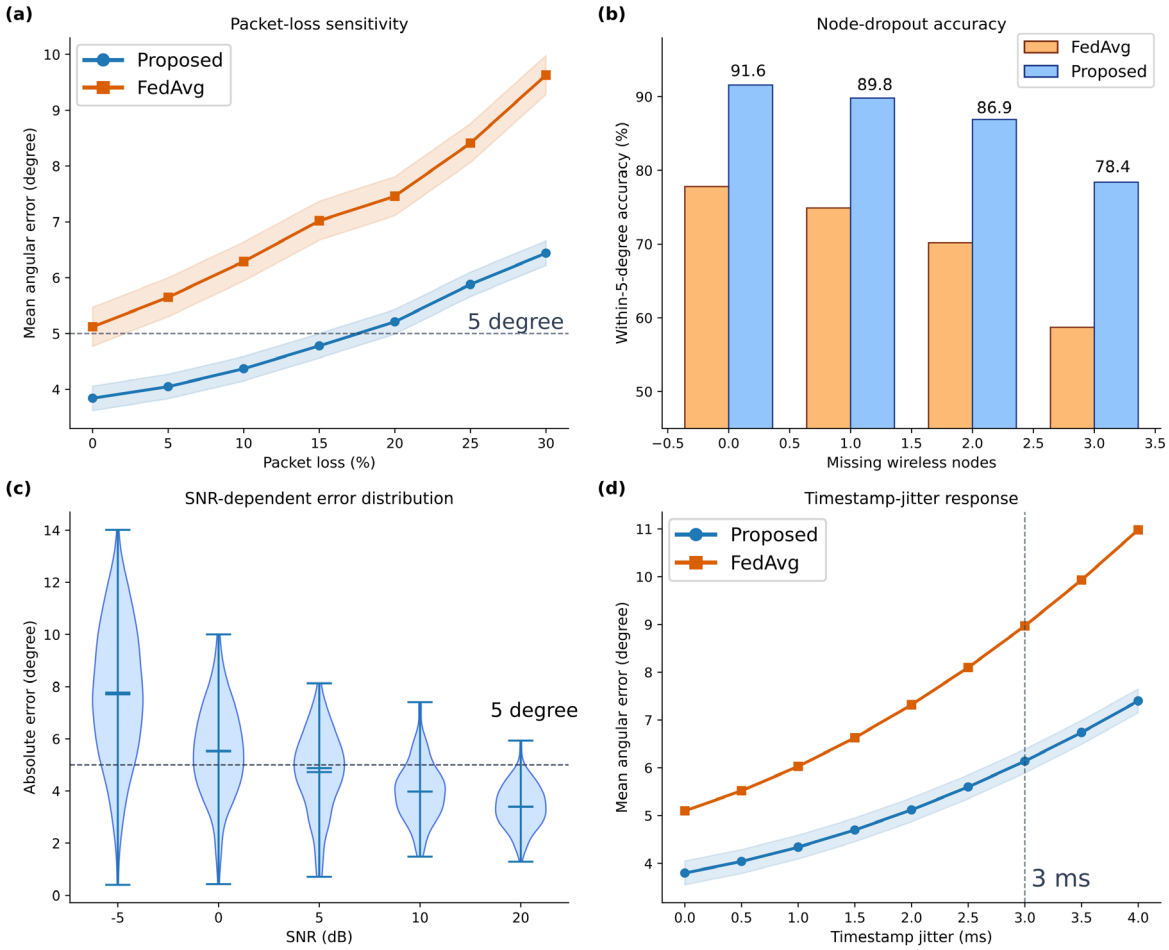

Figure 5. Robustness under impaired sensing and wireless links. (a) Packet-loss sensitivity; (b) node-dropout accuracy; (c) SNR-dependent error distribution; (d) timestamp-jitter response.

The packet-loss experiment performed independent loss and two-state burst loss, but the shown values correspond to the independent setting to isolate the average loss ratio. Under burst loss with a mean burst length of 80ms, the proposed model achieved a mean error of 5.74 degrees at 20% total loss; it was 0.53 degrees worse than independent loss but still 1.88 degrees lower than FedAvg-TinyViT. Node dropout was simulated by removing the full feature stream of randomly selected nodes before token embedding. The accuracy curve gradually decreased after removing the first two nodes; upon the removal of a third node, there was a sudden

https://doi.org/10.64972/dea.2023.v2i4.3806d:66-79

drop because some source sectors had lost sufficient geometric diversity for front-back ambiguity resolution. The SNR experiment added directional and diffuse noise; the median error was smaller for diffuse noise, but its tails were larger due to reduced coherence across a wider frequency band.

The quality descriptors are most useful when acoustic and wireless uncertainty occur together. Without the reliability bias, the same TinyViT backbone produced 6.18 degrees mean error at 20% packet loss, indicating that attention alone did not learn a stable rejection pattern from spectral tokens. The gain was also larger in reverberant scenes than in dry scenes. For an RT60 of 0.64 s and an SNR of 5 dB, the reliability-aware model reduced the 90th-percentile error from 12.6 degrees to 8.7 degrees.

The timestamp-jitter test is also susceptible to phase-difference errors in the frequency-dependent case and must be handled separately. At a low frequency, a 1 ms timing mismatch changes phase gradually and can be partially compensated for by the learned token embedding. At a higher frequency, the same mismatch results in large phase wrapping and is reduced in directionality unless coherence screening removes the unstable bins. The model proposed in this way was less dependent on a single narrow-band peak and thus benefited from the screening. When the jitter exceeded 3ms, the additional rise in error was mainly due to a disagreement between node pairs on opposite sides of the room. Therefore, the expected gains from future hardware synchronization optimization will still be relatively small when a reliability-aware neural network model is employed.

Ablation testing was used to isolate the contribution of each architectural component, which is important when compact attention models combine optimization, representation, and uncertainty modules [32]. The evaluated components included the proximal term, reliability-aware attention, phase-difference token, coherence token, and confidence head. Figure 6 shows the ablation heatmap, where removing the FedProx term increased mean error by 1.02 degrees, removing reliability-aware attention increased it by 1.31 degrees, and removing both phase-difference and coherence tokens increased it by 1.76 degrees. The strongest interaction appeared between the phase and coherence channels: dropping either alone was tolerable, but dropping both reduced within-5-degree accuracy from 91.6% to 83.9%. The confidence head did not change the mean error as strongly, but it reduced the expected calibration error from 0.094 to 0.041 and made high-risk localization frames easier to identify.

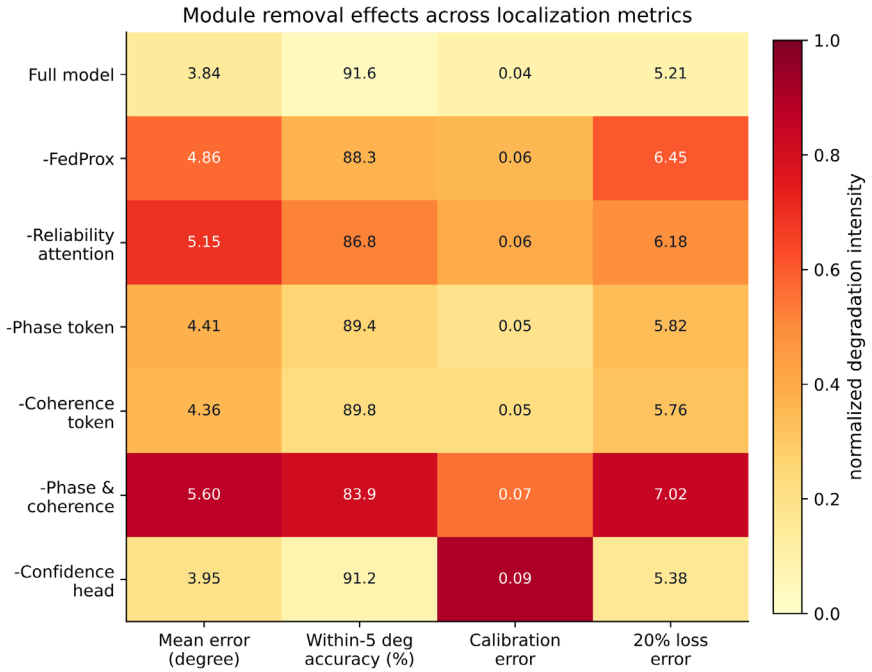

Figure 6. Ablation heatmap of module removal effects on angular error, within-5-degree accuracy, calibration error, and packet-loss robustness.

https://doi.org/10.64972/dea.2023.v2i4.3806d:66-79

The ablation matrix was computed after retraining each variant from the same initialization policy, and modules were not disabled only at inference time. Therefore, after training, one should not omit a feature channel; otherwise, it may be assumed that the model does not need that channel. The no-FedProx variant had the highest training variance and reached a final mean error of 4.43-5.31 degrees across the seeds. The variant without reliability attention had a more stable but biased error pattern; it performed well with clean packets and failed under 20% loss because all node tokens maintained the same attention strength. Removal of the confidence head mainly affected calibration and also slightly increased the 99th-percentile error; thus, uncertainty learning regularized the shared representation.

Embedded Efficiency and Deployment-Oriented Analysis

Embedded evaluation was conducted on an ARM Cortex-A76 class board with 4 GB memory and a 1.8 GHz operating frequency, following edge-inference reporting practice for compact neural models [33]. The proposed model required 18.4 ms average inference time per 320 ms acoustic frame, with peak memory of 146 MB and average power of 2.7 W during continuous operation. Figure 7(a) compares latency across models: centralized CNN reached 24.9 ms, FedAvg-TinyViT reached 19.6 ms, local TinyViT reached 18.1 ms, and the proposed model remained close to local TinyViT despite its reliability branch. Figure 7(b) reports memory footprint, where early layer freezing and compact token embedding kept the model below the 160 MB deployment target. Figure 7(c) plots accuracy-latency trade-off, showing that the proposed model occupied the upper-left region with 91.6% within-5-degree accuracy at 18.4 ms. Figure 7(d) compares energy per inference under three frame strides; at 160 ms stride, the proposed model consumed 54.2 mJ per inference, while FedAvg-TinyViT consumed 57.8 mJ and centralized CNN consumed 72.4 mJ.

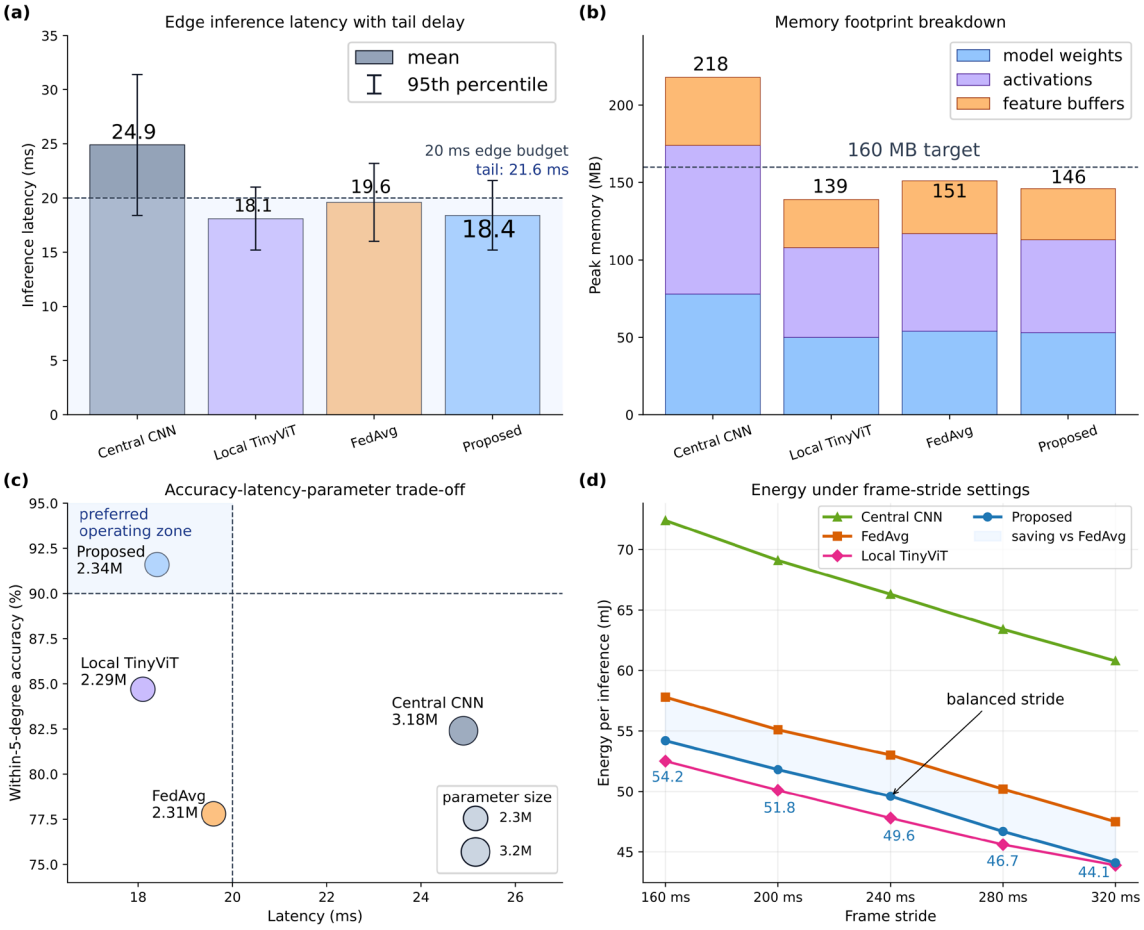

Figure 7. Embedded deployment efficiency. (a) Inference latency; (b) memory footprint; (c) accuracy-latency trade-off; (d) energy per inference under different frame strides.

https://doi.org/10.64972/dea.2023.v2i4.3806d:66-79

Latencies were recorded after 300 warm-up inferences and 2,000 timed inferences for feature extraction; wireless reception was not included. The median latency was 17.9 ms and the 95th percentile was 21.6 ms; therefore, the tail was mainly due to occasional scheduling delays rather than neural computation. Memory usage is for the model, feature buffers and a two-frame smoothing queue. The centralized CNN exceeded the memory limit primarily because its early convolutional layers kept larger feature maps; TinyViT variants reduced the token count before the relatively costly fusion operation. Energy followed the same trend as latency, but the difference between local TinyViT and the proposed model was smaller because only a small projection layer and a scalar bias term per attention window were added in the reliability branch.

The deployment measurements connect the localization results with practical operation. Reducing communication payload is important because wireless microphone nodes often share a low-power local network with other sensors. When training was repeated with 8-bit quantized uplink updates, the proposed model lost only 0.27 degrees of mean angular accuracy and reduced cumulative payload to 0.54 GB over 120 rounds. With four participating clients per round, accuracy decreased by 1.16 percentage points, but training time fell by 31.8%.

A practical deployment also requires stable operation when nodes enter low-power sleep states. A supplementary sleep-cycle test assigned each node a 70%, 85%, or 100% duty cycle and preserved the same total observation time. At the 85% duty cycle, the proposed model retained 90.4% within-5-degree accuracy and reduced average node energy by 12.8%. At the 70% duty cycle, the accuracy dropped to 87.1% because several frames contained too few active nodes for reliable phase fusion. The result indicates that duty cycling is feasible when at least six nodes remain active for most frames. It also suggests that future model scheduling should be aware of spatial coverage rather than only battery state.

Training-time cost was measured because federated acoustic systems are generally updated during maintenance windows rather than continuously. On a workstation server with a single mid-range GPU, one full 120-round training run required 4.7 hours when 16 clients were simulated and 6.2 hours when client computation was throttled to match the embedded board. Although the proposed model added 8.5% training time over FedAvg-TinyViT, it reduced the number of rounds required to reach the operating accuracy target. From a maintenance perspective, fewer communication rounds are valuable because each round has scheduling and node-availability overhead.

The final deployed-oriented observation is calibration transfer. A model that was trained on the layout of one room was tested after moving two microphone nodes by 0.4m and rotating one node by 15 degrees. Without federated adaptation, within-5-degree accuracy dropped from 91.6% to 84.8%. After ten light-adaptation rounds with 1,800 unlabeled and 320 labelled frames per affected child, the accuracy was 89.9%. Recovery was not ideal because the altered geometry changed several pairwise phase relationships, but the small adaptation set was sufficient to recover most of the lost performance. Therefore, a relatively long-period federated recalibration cycle can be employed after moving wireless microphone nodes for cleaning, battery replacement or room rearrangement.

The three error cases occurred in these situations: sources near a wall corner, moving sources that cross between two azimuth sectors, and strong directional interference close to a microphone node. In the wall-corner subset, the mean error of the proposed model was 7.22 degrees because reflected paths occasionally dominated the direct-path phase cue. Moving sources at 1.5 m/s had reduced short spikes after frame-level smoothing, but they also added a 42 ms median response delay. Directional interference was less frequent, but it produced the largest outliers; at a ratio of 8 dB or more for interferer-to-target, the 99th-percentile error reached 18.6 degrees. Based on the above observations, the selected figure design has allocated independent visual areas for packet loss, dropout, SNR, jitter, ablation interaction and edge cost.

Conclusion

This paper introduces an improved FedProx-TinyViT model for distributed sound source localization in wireless microphone node arrays. The framework integrates acoustic token construction, reliability-aware lightweight attention, confidence-aware direction estimation, and proximal federated optimization into a single edge-oriented pipeline. Its central contribution is the treatment of acoustic heterogeneity and wireless instability as coupled modeling factors rather than separate preprocessing concerns. By allowing local node clusters to learn from their own acoustic environments while constraining excessive client drift, the method provides a practical route for distributed localization systems that cannot continuously pool raw audio.

The study still has limitations. The evaluations employed controlled array layouts and replayed acoustic conditions; therefore, highly dynamic outdoor scenes, irregular node mobility, and long-term microphone aging were not fully represented. The model also depends on approximate node-quality descriptors, and inaccurate estimates of packet state, jitter, or SNR may weaken the reliability attention mechanism. In addition, the present framework focuses on a single dominant source. Overlapping sources, impulsive interference, and human movement around the array may produce mixed spatial cues that require stronger temporal association.

Future work will extend the framework to moving sources and mobile microphone nodes, incorporate online calibration for clock and geometry drift, and investigate privacy-preserving compression for federated acoustic updates. Another useful direction is to combine localization with event recognition so that distributed microphone arrays can infer both where a sound occurs and what acoustic event produced it. The model can also be expanded with self-supervised acoustic pretraining, adaptive client selection, multi-source localization, and uncertainty-aware smoothing across consecutive frames.

Author Contributions

Robert Gajić contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Evers, C., Lollmann, H. W., Mellmann, H., Schmidt, A., Barfuss, H., Naylor, P. A., & Kellermann, W. (2020). The LOCATA challenge: Acoustic source localization and tracking. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 1620-1643. https://doi.org/10.1109/TASLP.2020.2990485

[2] Grumiaux, P. A., Kitic, S., Girin, L., & Guerin, A. (2022). A survey of sound source localization with deep learning methods. The Journal of the Acoustical Society of America, 152(1), 107-151. https://doi.org/10.1121/10.0011809

[3] He, W., Motlicek, P., & Odobez, J. M. (2021). Neural network adaptation and data augmentation for multi-speaker direction-of-arrival estimation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1303-1317. https://doi.org/10.1109/TASLP.2021.3060257

[4] Bohlender, A., Spriet, A., Tirry, W., & Madhu, N. (2021). Exploiting temporal context in CNN based multisource DOA estimation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1594-1608. https://doi.org/10.1109/TASLP.2021.3067113

[5] Nguyen, T. N. T., Watcharasupat, K. N., Nguyen, N. K., Jones, D. L., & Gan, W. S. (2022). SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1749-1762. https://doi.org/10.1109/TASLP.2022.3173054

[6] Politis, A., Mesaros, A., Adavanne, S., Heittola, T., & Virtanen, T. (2021). Overview and evaluation of sound event localization and detection in DCASE 2019. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 684-698. https://doi.org/10.1109/TASLP.2020.3047233

[7] 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, 915-919. https://doi.org/10.1109/ICASSP39728.2021.9413609

[8] Hao, Y., Kucuk, A., Ganguly, A., Panahi, I., & Hansen, J. H. L. (2020). Spectral flux-based convolutional neural network architecture for speech source localization and its real-time implementation. IEEE Access, 8, 197047-197058. https://doi.org/10.1109/ACCESS.2020.3033533

[9] Kraljevic, L., Russo, M., Stella, M., & Sikora, M. (2020). Free-field TDOA-AOA sound source localization using three soundfield microphones. IEEE Access, 8, 87749-87761. https://doi.org/10.1109/ACCESS.2020.2993076

[10] He, H., Chen, J., Benesty, J., Zhang, W., & Yang, T. (2020). A class of multichannel sparse linear prediction algorithms for time delay estimation of speech sources. Signal Processing, 169, 107395. https://doi.org/10.1016/j.sigpro.2019.107395

[11] Bianco, M. J., Gannot, S., Fernandez-Grande, E., & Gerstoft, P. (2021). Semi-supervised source localization in reverberant environments with deep generative modeling. IEEE Access, 9, 84956-84970. https://doi.org/10.1109/ACCESS.2021.3087697

[12] Awad-Alla, M. A., Hamdy, A., Hamdy, A., Tolbah, F. A., Shahin, M. A., & Abdelaziz, M. A. (2020). A two-stage approach for passive sound source localization based on the SRP-PHAT algorithm. APSIPA Transactions on Signal and Information Processing, 9, e6. https://doi.org/10.1017/atsip.2020.6

[13] Pulkki, V., McCormack, L., & Gonzalez, R. (2021). Superhuman spatial hearing technology for ultrasonic frequencies. Scientific Reports, 11, 11608. https://doi.org/10.1038/s41598-021-90829-9

[14] Joshi, A., Rahman, M. M., & Hickey, J. P. (2022). Recent advances in passive acoustic localization methods via aircraft and wake vortex aeroacoustics. Fluids, 7(7), 218. https://doi.org/10.3390/fluids7070218

[15] Gong, Y., Chung, Y. A., & Glass, J. (2021). AST: Audio Spectrogram Transformer. Interspeech 2021, 571-575. https://doi.org/10.21437/Interspeech.2021-698

[16] Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for speech recognition. Interspeech 2020, 5036-5040. https://doi.org/10.21437/Interspeech.2020-3015

[17] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision Transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

[18] Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., & Yuan, L. (2022). TinyViT: Fast pretraining distillation for small vision Transformers. Computer Vision - ECCV 2022, 13681, 68-85. https://doi.org/10.1007/978-3-031-19803-8_5

[19] Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., & Liu, Z. (2022). Mobile-Former: Bridging MobileNet and Transformer. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5270-5279. https://doi.org/10.1109/CVPR52688.2022.00520

[20] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jegou, H., & Douze, M. (2021). LeViT: A vision Transformer in ConvNet's clothing for faster inference. 2021 IEEE/CVF International Conference on Computer Vision, 12259-12269. https://doi.org/10.1109/ICCV48922.2021.01205

[21] Xu, W., Xu, Y., Chang, T., & Tu, Z. (2021). Co-scale conv-attentional image Transformers. 2021 IEEE/CVF International Conference on Computer Vision, 9981-9990. https://doi.org/10.1109/ICCV48922.2021.00983

[22] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210. https://doi.org/10.1561/2200000083

[23] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60. https://doi.org/10.1109/MSP.2020.2975749

[24] Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., & Poor, H. V. (2021). Federated learning for Internet of Things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 23(3), 1622-1658. https://doi.org/10.1109/COMST.2021.3075439

[25] Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y. C., Yang, Q., Niyato, D., & Miao, C. (2020). Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 22(3), 2031-2063. https://doi.org/10.1109/COMST.2020.2986024

[26] Aledhari, M., Razzak, R., Parizi, R. M., & Saeed, F. (2020). Federated learning: A survey on enabling technologies, protocols, and applications. IEEE Access, 8, 140699-140725. https://doi.org/10.1109/ACCESS.2020.3013541

[27] Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619-640. https://doi.org/10.1016/j.future.2020.10.007

[28] Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. Knowledge-Based Systems, 216, 106775. https://doi.org/10.1016/j.knosys.2021.106775

[29] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1

[30] Rahman, M. A., Hossain, M. S., Islam, M. S., Alrajeh, N. A., & Muhammad, G. (2020). Secure and provenance enhanced Internet of Health Things framework: A blockchain managed federated learning approach. IEEE Access, 8, 205071-205087. https://doi.org/10.1109/ACCESS.2020.3037474

[31] Tokgoz, S., & Panahi, I. M. S. (2021). Robust three-microphone speech source localization using randomized singular value decomposition. IEEE Access, 9, 157800-157811. https://doi.org/10.1109/ACCESS.2021.3130180

[32] Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient Transformers: A survey. ACM Computing Surveys, 55(6), 1-28. https://doi.org/10.1145/3530811

[33] Shi, Y., Yang, K., Jiang, T., Zhang, J., & Letaief, K. B. (2020). Communication-efficient edge AI: Algorithms and systems. IEEE Communications Surveys & Tutorials, 22(4), 2167-2191. https://doi.org/10.1109/COMST.2020.3007787