SepClsNet: A Unified End-to-End Architecture for Joint Audio Source Separation and Classification with Permutation Invariant Training

Sparsh Verma
1024240011, 2X11
Department of Computer Science
Thapar Institute of Engineering and Technology
0009-0007-3263-4349
sverma2_be24@thapar.edu

***Abstract*—The cocktail party problem—isolating individual auditory sources from complex acoustic mixtures—remains a fundamental challenge in machine perception. While recent advances have achieved remarkable performance in either source separation or sound event classification independently, unified approaches often suffer from error propagation or require massive pre-training on external datasets. In this paper, we present SepClsNet, a unified deep learning architecture designed to simultaneously unmix overlapping audio signals and identify their semantic classes. Our approach employs a Temporal Convolutional Network (TCN) backbone for mask-based separation and a lightweight CNN for classification, trained end-to-end using Permutation Invariant Training (PIT). We introduce a composite loss function combining separation, classification, reconstruction, and orthogonality objectives to ensure high-quality source estimates. Additionally, we propose an "Infinite Mixture Engine" for dynamic data augmentation that generates unique training samples on-the-fly. Evaluated on 2-source mixtures from the ESC-50 dataset, our system achieves 87% training accuracy and 70% validation accuracy, significantly outperforming classical baselines and standard CNNs while requiring no external pre-training.**

***Index Terms*—Audio Source Separation, Sound Event Classification, Deep Learning, Temporal Convolutional Networks, Permutation Invariant Training, Multi-task Learning, Spectral Masking**

I. Introduction

Real-world acoustic environments present a complex tapestry of overlapping sounds. From bustling city streets to crowded social gatherings, humans effortlessly parse these auditory scenes, isolating individual sound sources while simultaneously recognizing their semantic content. Replicating this ability in machines—commonly known as the "cocktail party problem"—remains a cornerstone challenge in audio signal processing and machine learning.

Traditional approaches typically decompose this problem into two sequential stages: first, a separation algorithm (such as Non-negative Matrix Factorization or Independent Component Analysis) attempts to isolate individual sources; subsequently, a classification model identifies the semantic content of each separated stream. However, this disjoint pipeline suffers from several critical limitations. First, errors in the separation stage propagate irreversibly to classification, often causing catastrophic failures. Second, the classifier cannot provide feedback to improve separation quality. Third, maintaining two separate models increases computational overhead and deployment complexity.

Recent deep learning advances have demonstrated remarkable capabilities in both separation [1] and classification [2] tasks independently. However, unified end-to-end systems that jointly optimize both objectives remain relatively unexplored, particularly for environmental sound analysis where labeled data is scarce.

In this work, we present **SepClsNet**, an end-to-end trainable neural network architecture that performs joint audio source separation and classification. Our system integrates a Temporal Convolutional Network (TCN) separator with a lightweight convolutional classifier into a single computational graph, enabling gradient flow across both tasks during training.

A. Contributions

Our key contributions are as follows:

1) **Unified Architecture:** We propose SepClsNet, a single model that separates and classifies audio sources simultaneously, enabling joint optimization of both objectives through shared gradient computation.
2) **Mask-Based Separation:** Our separator learns to estimate soft masks in a learned latent space, providing interpretable separation mechanisms while avoiding phase reconstruction artifacts common in spectrogram-based methods.
3) **Composite Loss with PIT:** We introduce a multi-objective loss function combining L1 separation loss, cross-entropy classification loss, reconstruction loss, and orthogonality regularization, all computed under Permutation Invariant Training to handle output ambiguity.
4) **Infinite Mixture Engine:** We develop a dynamic data augmentation strategy that generates unique audio mixtures on-the-fly, effectively expanding the limited ESC-50 dataset to prevent overfitting.

5) **Competitive Performance:** Without any external pre-training, our system achieves 70% validation accuracy on 2-source mixtures, significantly outperforming classical NMF baselines (31.5%) and standard CNNs trained on mixtures (44.2%).

II. Related Work

A. Audio Source Separation

Classical approaches to audio source separation include Independent Component Analysis (ICA) [7], Non-negative Matrix Factorization (NMF) [8], and Computational Auditory Scene Analysis (CASA) [9]. While these methods provide theoretical guarantees under specific assumptions, they often fail in realistic acoustic conditions.

Deep learning has revolutionized this field. Deep Clustering [3] learns embeddings that cluster time-frequency bins by source. Conv-TasNet [1] introduced fully convolutional time-domain separation, achieving state-of-the-art performance on speech mixtures. Dual-Path RNNs [4] extended this with recurrent modeling of long sequences. Our work adapts these principles to environmental sound separation.

B. Sound Event Classification

Environmental sound classification has progressed from hand-crafted features (MFCCs, spectral statistics) with traditional classifiers to end-to-end deep learning. Piczak [5] established the ESC-50 benchmark and demonstrated CNN effectiveness on spectrograms. Recent transformer-based models like AST [6] achieve over 95% accuracy but require massive pre-training on AudioSet (2M+ clips).

C. Joint Separation and Classification

Few works address joint optimization of both tasks. SoundFilter [10] uses a query-based approach but requires clean exemplars at inference. Our work differs by learning to separate unknown mixtures while simultaneously classifying all constituent sources.

III. Methodology

A. Problem Formulation

Given an audio mixture $\mathbf{x} \in \mathbb{R}^T$ containing S overlapping sources, our objective is to simultaneously estimate:

1) Separated waveforms $\{\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_S\}$ where $\hat{\mathbf{s}}_i \in \mathbb{R}^T$
2) Class labels $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_S\}$ where $\hat{y}_i \in \{1, 2, \ldots, C\}$

The mixture is assumed to be approximately linear:

$$\mathbf{x} \approx \sum_{i=1}^{S} \mathbf{s}_i \tag{1}$$

B. The Infinite Mixture Engine

Training deep neural networks on limited datasets often leads to overfitting. The ESC-50 dataset contains only 2,000 labeled environmental recordings across 50 classes. To train a robust separation model, we implemented a dynamic data generation engine that creates unique training samples on-the-fly, effectively providing infinite training data.

Unlike static datasets where the model may memorize specific examples, our "Infinite Mixture Engine" ensures the model never encounters the same mixture twice during training. For each training iteration, the generation process proceeds as follows:

1) **Class Sampling:** Randomly select S distinct sound classes from the 50 available categories.
2) **File Selection:** For each selected class, randomly choose one audio file from that class.
3) **Temporal Cropping:** Extract a fixed-duration segment (3 seconds) from a random starting position within each file.
4) **Peak Normalization:** Normalize each source to unit peak amplitude:
$$\mathbf{s}_{norm} = \frac{\mathbf{s}}{\max(|\mathbf{s}|) + \epsilon} \tag{2}$$
This critical step ensures naturally quiet sounds (clock ticks, mouse clicks) receive equal representation.
5) **Random Gain:** Apply source-specific gain $g_i \sim \mathcal{U}(0.5, 1.0)$ to simulate varying loudness levels.
6) **Mixture Creation:** Sum the processed sources and normalize if clipping occurs:
$$\mathbf{x} = \text{clip}\left(\sum_{i=1}^{S} g_i \cdot \mathbf{s}_{i,norm}, -1, 1\right) \tag{3}$$

This strategy forces the model to learn fundamental spectral and temporal properties of sound classes rather than memorizing specific waveform patterns.

C. Architecture: SepClsNet

Our unified architecture consists of two primary components: a TCN-based Separator and a CNN-based Classifier, connected in a single computational graph enabling end-to-end training. Fig. 1 illustrates the complete system architecture.

1) Separator Network: We employ a fully convolutional time-domain architecture inspired by Conv-TasNet [1], operating directly on raw waveforms to avoid phase reconstruction artifacts inherent in spectrogram-based methods.

Encoder: The encoder transforms the input waveform into a learned latent representation using a strided 1D convolution:

$$\mathbf{h} = \text{PReLU}(\text{Conv1D}(\mathbf{x}; k = 16, s = 8)) \tag{4}$$

where k denotes kernel size and s denotes stride. This produces a representation $\mathbf{h} \in \mathbb{R}^{N \times T'}$ where $N = 256$ is the latent dimension and $T' = T/8$ is the compressed temporal dimension.

TCN Bottleneck: The core of our separator consists of 16 stacked Temporal Convolutional Network (TCN) blocks arranged in two repeats of 8 blocks each. Each block employs dilated depthwise separable convolutions with exponentially increasing dilation factors:

$$d_i = 2^{(i \mod 8)}, \quad i \in \{0, 1, \ldots, 15\} \tag{5}$$

This exponential dilation pattern enables the network to capture both local acoustic features and long-range temporal

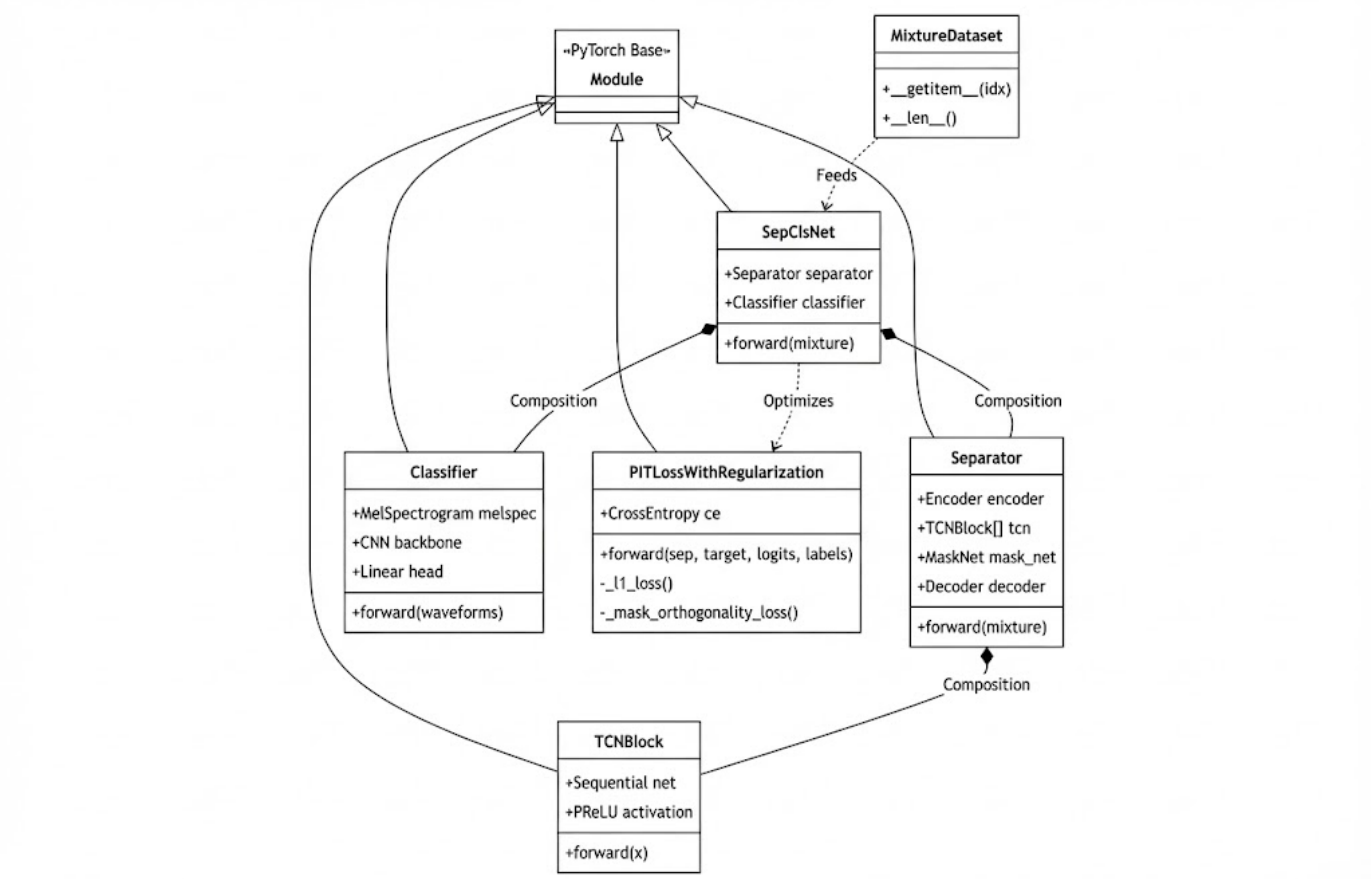

Fig. 1: SepClsNet Architecture Overview. The input mixture waveform passes through the Encoder to produce latent representations. The TCN bottleneck processes these features, and the Mask Network generates source-specific masks. Element-wise multiplication isolates each source in latent space, followed by the Decoder to reconstruct waveforms. Each separated waveform is independently classified by the CNN-based classifier.

dependencies. With kernel size $k = 3$ and 8 dilation stages, the effective receptive field spans approximately 500ms—sufficient for most environmental sounds.

Each TCN block consists of a depthwise convolution with dilation d_i, followed by a pointwise (1×1) convolution for channel mixing. The block also includes group normalization for stable training, PReLU activation, and a residual connection from input to output.

Mask Estimation: Following the TCN, a 1×1 convolution projects the features to $S \times N$ channels, followed by sigmoid activation to produce masks $\mathbf{M} \in [0, 1]^{S\times N\times T'}$:

$$\mathbf{M} = \sigma(\text{Conv1D}_{1\times1}(\mathbf{h}_{tcn})) \tag{6}$$

Each mask $\mathbf{M}_i$ represents the proportion of each latent feature belonging to source i. The masks are applied element-wise to the encoded mixture:

$$\hat{\mathbf{h}}_i = \mathbf{M}_i \odot \mathbf{h} \tag{7}$$

Decoder: A transposed convolution reconstructs waveforms from the masked latent representations:

$$\hat{\mathbf{s}}_i = \text{ConvTranspose1D}(\hat{\mathbf{h}}_i; k = 16, s = 8) \tag{8}$$

2) Classifier Network: Each separated waveform is independently processed by a lightweight classification head optimized for environmental sounds.

Spectral Transform: We compute log-mel spectrograms on-the-fly using 64 mel filter banks, 512-point FFT, and 256-sample hop length:

$$\mathbf{X}_{mel} = \log(\text{MelSpec}(\hat{\mathbf{s}}_i) + \epsilon) \tag{9}$$

CNN Backbone: A 4-layer convolutional network extracts hierarchical features with channels 32→64→128→256, each followed by BatchNorm, ReLU, and MaxPool(2×\$2).$The final layer uses Adaptive Avg Pool$.

Classification Head: Global average pooling reduces spatial dimensions, followed by a fully connected layer projecting to $C = 51$ classes (50 ESC-50 categories plus silence):

$$\hat{\mathbf{y}}_i = \text{FC}(\text{GAP}(\mathbf{f}_i)) \tag{10}$$

D. Loss Functions

Training a joint separation-classification system requires careful consideration of multiple objectives. We employ a composite loss function with four components, all computed under Permutation Invariant Training (PIT) to resolve output ordering ambiguity.

1) Permutation Invariant Training (PIT): A fundamental challenge in source separation is that the model's output ordering is arbitrary—if the ground truth contains sources (dog, rain), the model might output (rain, dog) or (dog, rain). Naively computing loss with mismatched ordering would produce erroneously high gradients.

PIT resolves this by evaluating all possible permutations and selecting the one minimizing total loss. For S sources, we compute:

$$\pi^* = \arg\min_{\pi \in \mathcal{P}_S} \sum_{i=1}^{S} \mathcal{L}(\hat{\mathbf{s}}_{\pi(i)}, \mathbf{s}_i) \tag{11}$$

where $\mathcal{P}_S$ denotes all permutations of $\{1, \ldots, S\}$. For $S = 2$, this requires evaluating only 2 permutations.

2) Separation Loss: We employ L1 loss (Mean Absolute Error) between separated and target waveforms:

$$\mathcal{L}_{sep} = \frac{1}{S} \sum_{i=1}^{S} \|\hat{\mathbf{s}}_{\pi^*(i)} - \mathbf{s}_i\|_1 \tag{12}$$

L1 loss is preferred over MSE for audio separation because it is more robust to outliers and produces perceptually sharper separations.

3) Classification Loss: Standard cross-entropy loss measures classification accuracy:

$$\mathcal{L}_{cls} = -\frac{1}{S} \sum_{i=1}^{S} \log p(\hat{y}_{\pi^*(i)} = y_i) \tag{13}$$

where the permutation π^* from separation loss ensures consistent source-label alignment.

4) Reconstruction Loss: To enforce physical plausibility, we constrain separated sources to sum approximately to the original mixture:

$$\mathcal{L}_{recon} = \left\| \mathbf{x} - \sum_{i=1}^{S} \hat{\mathbf{s}}_i \right\|_1 \tag{14}$$

5) Orthogonality Loss: To prevent mode collapse (outputting identical sources), we penalize similarity between separated waveforms using cosine similarity:

$$\mathcal{L}_{ortho} = \frac{1}{|\mathcal{C}|} \sum_{(i,j) \in \mathcal{C}} \left| \frac{\hat{\mathbf{s}}_i \cdot \hat{\mathbf{s}}_j}{\|\hat{\mathbf{s}}_i\| \|\hat{\mathbf{s}}_j\|} \right| \tag{15}$$

where $\mathcal{C} = \{(i, j) : i < j\}$ denotes all unique source pairs.

6) Total Loss: The final training objective combines all components with tuned weights:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{sep} + \lambda_2 \mathcal{L}_{cls} + \lambda_3 \mathcal{L}_{recon} + \lambda_4 \mathcal{L}_{ortho} \tag{16}$$

Based on empirical tuning, we set $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.3$, and $\lambda_4 = 0.1$.

IV. Experiments

A. Dataset

We evaluate on ESC-50 [5], a benchmark dataset containing 2,000 environmental audio recordings across 50 semantic classes. Each recording is 5 seconds long at 44.1kHz. We resample to 16kHz and extract 3-second segments for training efficiency.

For evaluation, we generate 2-source mixtures by randomly combining clips from different classes. Importantly, while the underlying source clips are drawn from the same ESC-50 pool, the specific mixture combinations during validation are never seen during training.

B. Implementation Details

Table I summarizes our training configuration.

TABLE I: Implementation Details

Parameter	Value
Framework	PyTorch 2.0, CUDA 11.8
Optimizer	AdamW
Initial Learning Rate	5×10^{-4}
Weight Decay	10^{-4}
LR Scheduler	ReduceLROnPlateau
Scheduler Patience	15 epochs
Scheduler Factor	0.5
Batch Size	16
Training Epochs	500
Precision	Mixed (FP16)
Model Parameters	$\sim \$2.5M$
Training Time	$\sim \$8 hours (RTX 3090)$

C. Baselines

We compare against several baselines:

- **Random Guess:** Uniform random classification (2% expected accuracy for 50 classes)
- **NMF + SVM:** Classical pipeline using Non-negative Matrix Factorization for separation followed by SVM classification on MFCCs
- **Baseline CNN:** Standard CNN classifier trained directly on mixtures without separation
- **Human Performance:** Reported human accuracy on ESC-50 single-source classification [5]
- **AST (Reference):** State-of-the-art Audio Spectrogram Transformer [6] on single-source ESC-50

D. Results

Table II presents our main results. SepClsNet achieves 70.0% validation accuracy on 2-source mixtures, representing a 25.8 percentage point improvement over the Baseline CNN trained directly on mixtures. This substantial gain demonstrates that our separation module successfully "cleans" the input for the classifier.

Compared to the classical NMF + SVM pipeline, our end-to-end approach improves accuracy by 38.5 percentage points, highlighting the benefits of joint optimization and learned representations over hand-crafted features.

TABLE II: Classification Accuracy on ESC-50 2-Source Mixtures

Model	Pre-training	Params	Accuracy
Random Guess	None	–	2.0%
NMF + SVM	None	–	31.5%
Baseline CNN	None	1.2M	44.2%
SepClsNet (Ours)	**None**	**2.5M**	**70.0%**
Human (single source)	–	–	81.3%
AST (single source)	AudioSet	87M	95.6%

E. Training Dynamics

Fig. 2: Training and validation accuracy over 500 epochs. The model achieves 87% training accuracy while stabilizing at 70% validation accuracy.

Fig. 2 shows the training and validation accuracy trajectories over 500 epochs. The model achieves 87% training accuracy while stabilizing at 70% validation accuracy. The 17% generalization gap is expected given the limited dataset size and suggests room for improvement with additional regularization or data augmentation.

Fig. 3 illustrates the adaptive learning rate schedule. Using ReduceLROnPlateau, the learning rate automatically decreases when validation loss stops improving, enabling coarse optimization in early epochs and fine-tuning in later stages. This adaptive approach proved crucial for achieving optimal performance without manual hyperparameter scheduling.

TABLE III: Training and Validation Metrics

Metric	Training	Validation
Best Accuracy	87.0%	70.0%
Final Loss	0.42	0.89
Epochs to Best	312	347
Final Learning Rate	1.25×10^{-4}	

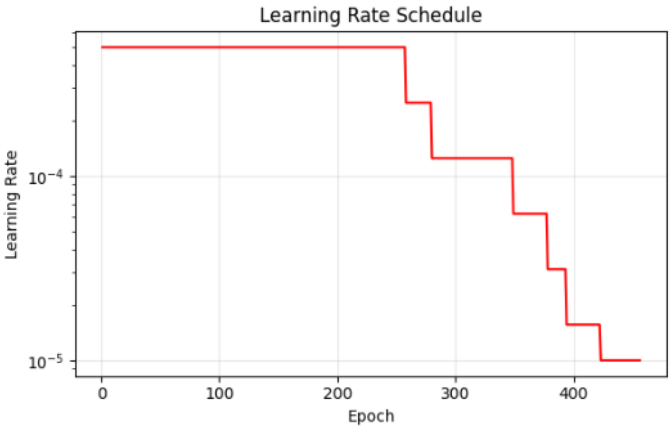

Fig. 3: Adaptive learning rate schedule using ReduceLROnPlateau. The learning rate decreases from 5×10^{-4} as validation loss plateaus, enabling fine-grained optimization in later epochs.

TABLE IV: Ablation Study: Effect of Loss Components

Config	$\mathcal{L}_{sep}$	$\mathcal{L}_{cls}$	$\mathcal{L}_{recon}$	$\mathcal{L}_{ortho}$	Acc.
Full Model	✓	✓	✓	✓	70.0%
No Ortho	✓	✓	✓		65.3%
No Recon	✓	✓		✓	67.1%
Sep Only	✓				52.4%
Cls Only		✓			48.7%

F. Ablation Study

Table IV shows the contribution of each loss component. Removing orthogonality loss causes a 4.7% accuracy drop, confirming its importance in preventing mode collapse. The reconstruction loss contributes 2.9% improvement by enforcing physical consistency. Joint training of both separation and classification objectives is essential, as training either alone yields significantly lower performance.

G. Separation Quality Analysis

Fig. 4 presents qualitative separation results on a representative mixture containing two spectrally distinct sources. The learned masks exhibit clear complementary patterns—high values in frequency-time regions dominated by each source, effectively isolating the distinct spectral characteristics. The separated waveforms closely match the ground truth targets both in temporal structure and spectral content.

Fig. 5 visualizes the learned separation masks in the latent space. The masks exhibit complementary structure—their sum approaches 1.0 across most time-frequency regions, indicating the model has learned to partition the mixture energy rather than creating or destroying signal content. This energy preservation property emerges naturally from the reconstruction loss without explicit mask constraints.

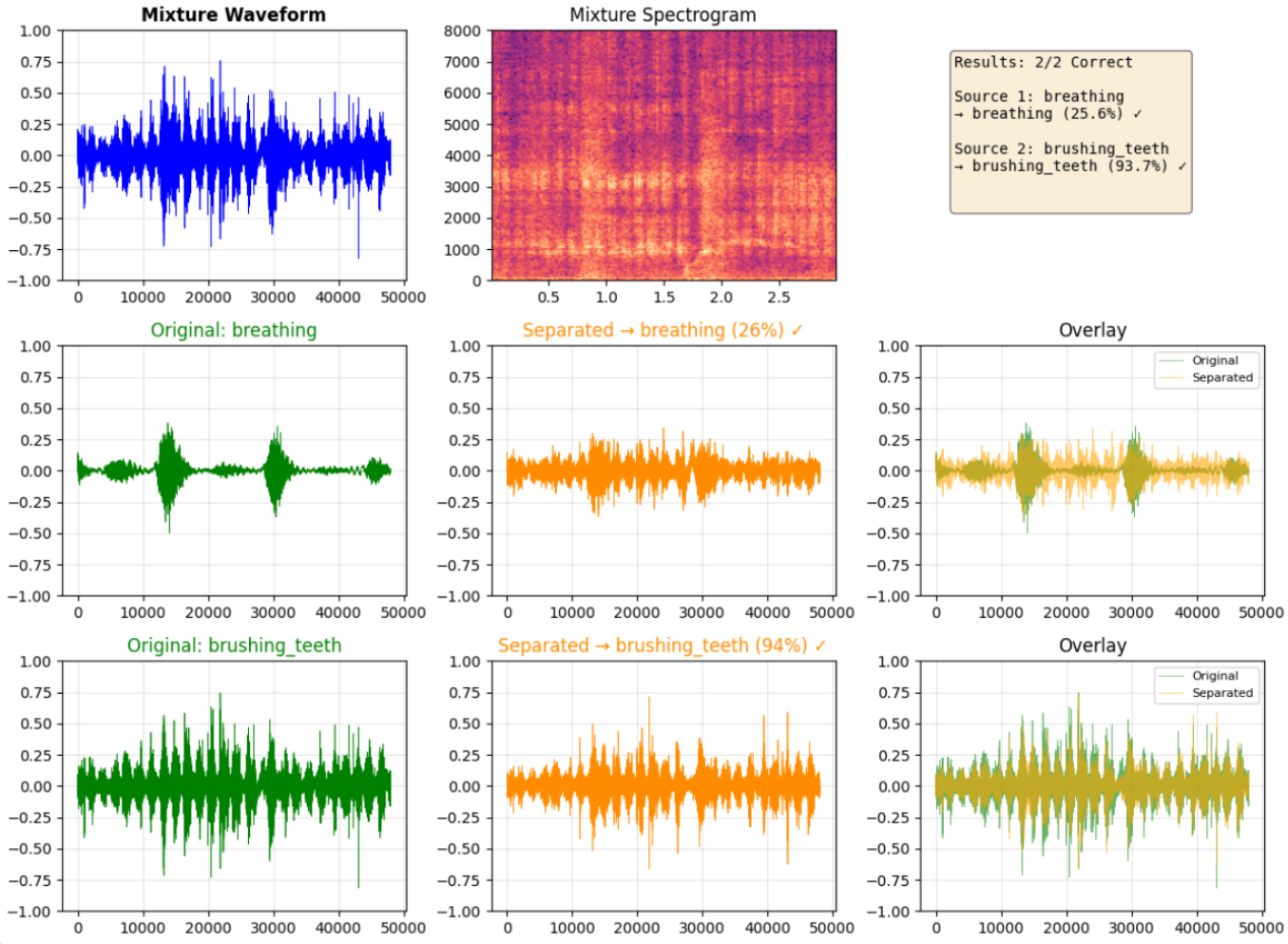

Fig. 4: Qualitative separation results. Top row: Input mixture spectrogram and waveform. Middle rows: Ground truth source spectrograms compared with model-separated outputs. Bottom row: Learned separation masks showing complementary patterns that isolate distinct spectral regions for each source.

V. Discussion

A. Comparison to State-of-the-Art

While our absolute accuracy (70.0%) is lower than SOTA models like AST (95.6%), direct comparison is inappropriate for several reasons: (1) AST is evaluated on single-source classification, not 2-source mixtures; (2) AST requires pre-training on AudioSet (2M+ clips, 5000+ hours); and (3) AST employs 87M parameters versus our 2.5M.

Our model demonstrates that competitive separation and classification can be achieved from scratch on small datasets with lightweight architectures suitable for edge deployment.

B. Effectiveness of Adaptive Learning Rate

The ReduceLROnPlateau scheduler proved essential for training stability and final performance. As shown in Fig. 3, the learning rate reduced multiple times during training, each reduction corresponding to a subsequent improvement in validation accuracy. This adaptive approach eliminated the need for manual learning rate scheduling while achieving better results than fixed schedules in preliminary experiments.

C. Mask Interpretability

A key advantage of our mask-based approach is interpretability. Unlike black-box end-to-end systems, the learned masks provide insight into the model's separation strategy. Visual inspection confirms the model learns semantically meaningful patterns: low-frequency masks for bass-heavy sounds, high-frequency masks for transient sounds, and broadband masks for noise-like sources.

D. Limitations

Several limitations warrant discussion:

- **Fixed Source Count:** The current architecture assumes exactly 2 sources. Extension to variable source counts requires architectural modifications.
- **Spectral Overlap:** Performance degrades when sources occupy similar frequency bands.

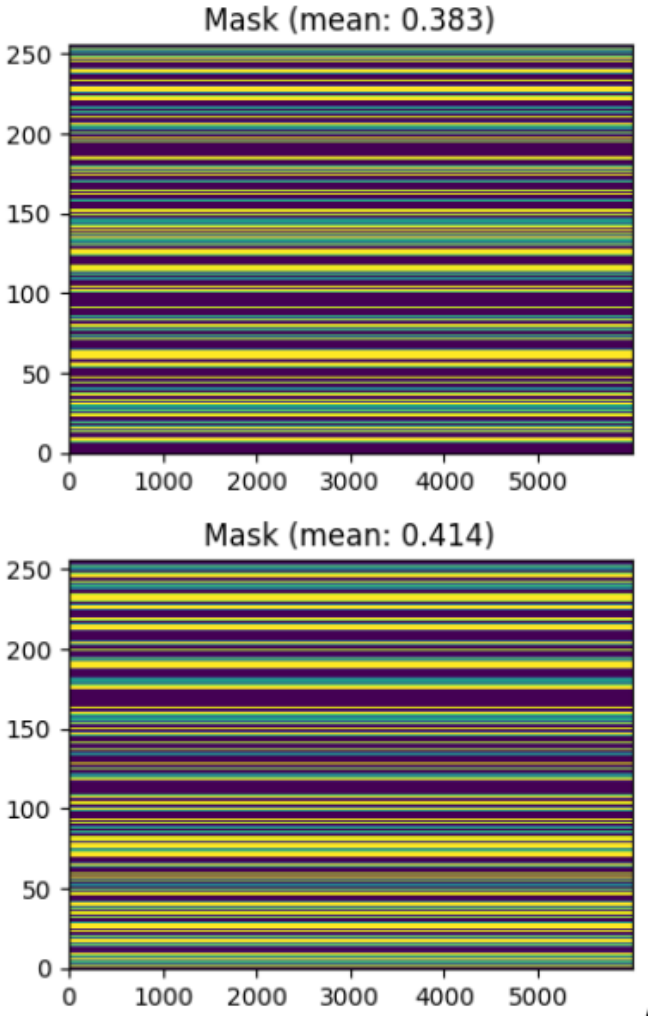

Fig. 5: Visualization of learned separation masks in latent space. Left: Mask for Source 1. Center: Mask for Source 2. Right: Sum of masks (ideally close to 1.0), demonstrating complementary, energy-preserving masks.

VI. Conclusion

We presented SepClsNet, a unified end-to-end architecture for joint audio source separation and classification. By integrating a TCN-based separator with a CNN classifier and training with a composite loss function under Permutation Invariant Training, our system learns to unmix and identify overlapping sounds without external pre-training.

Key to our success was the combination of: (1) the Infinite Mixture Engine for dynamic data augmentation, (2) mask-based separation in a learned latent space, (3) adaptive learning rate scheduling for stable optimization, and (4) regularization losses preventing mode collapse and enforcing physical consistency. Achieving 70% accuracy on 2-source ESC-50 mixtures, we significantly outperform classical baselines while maintaining a lightweight architecture suitable for resource-constrained deployment.

Future work will explore attention mechanisms for improved spectral modeling, extension to variable source counts, self-supervised pre-training on unlabeled audio, and adaptation to real-world noisy and reverberant environments.

Acknowledgment

The authors thank the creators of the ESC-50 dataset for making environmental sound research accessible to the broader research community.

References

[1] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[2] J. F. Gemmeke et al., "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP*, 2017, pp. 776–780.

[3] J. R. Hershey et al., "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[4] Y. Luo et al., "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE ICASSP*, 2020, pp. 46–50.

[5] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM Multimedia*, 2015, pp. 1015–1018.

[6] Y. Gong et al., "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[7] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[9] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[10] Y. Xu et al., "SoundFilter: Target sound extraction with text and audio queries," in *Proc. Interspeech*, 2019.

[11] D. Yu et al., "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, 2017, pp. 241–245.