Speaker Separation in Overlapping Speech Using Single-Channel Recordings in Varied Acoustic Environments

Jaipreet Kour Wazir * and Javaid A. Sheikh

Department of Electronics and IT, University of Kashmir, India
Email: jaipreet.elscholar@kashmiruniversity.net (J.K.W.); sheikhjavaid@uok.edu.in (J.A.S.)
*Corresponding author

***Abstract*—In this work, we address the challenging task of single-channel speech separation in realistic, reverberant environments. Our method focuses directly on separating overlapping speech signals captured through a single fixed-position microphone. We collected a custom dataset of crosstalk recordings using a Zoom H5 recorder in two acoustically distinct rooms, involving 50 speakers both male and female engaged in controlled conversational scenarios using standard Harvard sentences that are phonetically rich. Each recording captures dual-speaker overlaps within a single-channel signal, providing a realistic test for deep learning-based separation models. Our approach leverages data-driven neural architectures trained to separate the concurrent speech sources under varied room conditions with the knowledge of room geometry and microphone placement. Experimental results demonstrate the model's robustness across different spatial configurations and room sizes, showcasing its applicability in real-world speech communication systems. The effectiveness of the separation is evaluated using both objective metrics and perceptual measures, confirming the viability of deploying such systems in practical, resource-constrained settings.**

***Keywords*—single channel, speech communication, neural network, microphone, speech separation**

I. Introduction

In-room acoustic, speech separation in a single channel is challenging. Speech separation isolates individual speech sources from a mixture and plays a critical role in enhancing Automatic Speech Recognition (ASR), telecommunication systems, and assistive hearing devices [1]. Humans possess an innate ability to perceptually segregate complex auditory scenes, enabling them to distinguish between multiple speakers or to isolate speech from background noise. This effect is known as the "Cocktail Party" effect. Auditory systems process the statistical characteristics of acoustic patterns such as spectral and envelope cues to effectively identify and extract target sounds from complex auditory mixtures [2, 3]. Developing speech separation systems with robustness comparable to human auditory perception has been a longstanding objective in Artificial Intelligence (AI) [4]. Numerous Audio-Only Speech Separation (AOSS) techniques have been proposed to model the cocktail party effect. While these methods have demonstrated promising results under controlled or ideal conditions, achieving robust separation in real-world environments remains a significant challenge [5]. Replicating human auditory scene analysis in machines has been a longstanding objective in Artificial Intelligence (AI). In numerous audio-only speech separation techniques have been developed to emulate this effect. While these methods have achieved commendable results under controlled or ideal conditions, their performance often deteriorates in real-world scenarios characterized by reverberation, background noise, and overlapping speech. Factors contributing to this degradation include the acoustic similarity among speakers, interference caused by environmental reverberation, and speech signals inherently low information density. To address these challenges, researchers have explored incorporating additional modalities, such as visual cues, to enhance speech separation performance. Visual information, including lip movements and facial expressions, provides complementary data that can aid in isolating the target speaker from mixed audio signals. Studies have demonstrated that integrating visual cues can significantly improve separation accuracy, particularly in noisy environments where audio cues alone may be insufficient. Recent advancements in Audio-Visual Speech Separation (AVSS) have focused on leveraging visual cues to improve performance. Martel *et al.* [6] proposed a model that enhances audio separation using facial features and lip movements. Their approach incorporates residual connections in the audio separation module to extract detailed features and employs an attention mechanism in the face module to focus on crucial information. The loss function considers audio-visual similarity to exploit the relationship between audio and visual inputs fully. Experimental results on the VoxCeleb2 dataset

Manuscript received July 10 2025; revised August 22, 2025; accepted October 17, 2025; published February 25, 2026.

doi:10.12720/jcm.21.1.139-148

demonstrated significant improvements in metrics such as SDR, PESQ, and STOI, with a 4 dB enhancement in SDR. Another study by Xiong et al. introduced a cross-modal fusion strategy that benefits from semantic correlations between audio and visual modalities. The model incorporates dense optical flow of lip motion to strengthen the robustness of visual representation, leading to improved performance across multiple evaluation metrics. Furthermore, Chern *et al.* [7] developed a Dual Attention Cooperative Framework (Dual AVSE) that leverages facial cues beyond the lip region for robust audio-visual speech enhancement. Their approach employs a spatial attention-based visual encoder to capture and enhance visual speech information, incorporating global facial context while ignoring speech-unrelated information. A dynamic visual feature fusion strategy integrates temporal-dimensional self-attention, enabling the model to handle facial variations effectively. These studies underscore the importance of integrating visual information, such as lip movements and facial features, into speech separation models. By effectively combining audio and visual cues, these models can achieve more robust performance in real-world scenarios, addressing challenges posed by noise, reverberation, and overlapping speech. Factors such as acoustic similarity among speakers, reverberant interference, and speech signals inherently low information density contribute to this complexity [8−10]. Although increasing model capacity through more efficient neural architectures or expanding training datasets may offer performance gains, such strategies may not fundamentally overcome the limitations posed by the low information density [11, 12]. One possible solution to the challenges of speech separation is to incorporate richer and more distinct information, which can greatly improve the process. Visual cues provide valuable insights that help isolate the target speaker from mixed speech signals, thereby playing a crucial role in enhancing separation accuracy [13−15]. Traditional methods often rely on multichannel microphone arrays or simulated data, which may not accurately reflect the complexity of real-world acoustic environments. Furthermore, most real-world devices like IoT devices are equipped with only a single microphone, necessitating effective single-channel solutions [16]. This work bridges this gap by constructing a real-world crosstalk dataset with controlled variability in room acoustics and developing a robust neural separation model. We aim to explore how well a single-channel deep learning-based system can generalise across reverberant conditions and accurately isolate overlapping speech [17]. Conventional approaches frequently depend on multichannel microphone arrays or synthetic datasets, which often fail to capture the complexity and variability of real-world acoustic environments. To address this limitation, we present a real-world crosstalk dataset captured under controlled yet diverse room acoustic conditions and propose a robust neural network-based separation model. This study aims to evaluate the generalisation capability of a single-channel deep learning system in reverberant scenarios and its effectiveness in accurately isolating overlapping speech signals.

Our main contribution is as follows.

A. Creating the New Data Set

We created a novel real-world dataset containing 300 overlapping speech recordings captured using a Zoom H5 recorder. Unlike synthetic or simulated data, these recordings were collected in rooms of varying dimensions and reverberation profiles, mirroring real-life acoustic conditions. The dataset comprises speech overlaps from male and female speakers using standard Harvard sentences, enabling rigorous evaluation under authentic single-mic reverberant scenarios. This dataset helps bridge the gap between academic research and practical application environments where multichannel arrays are unavailable.

B. Lightweight Model for IOT Devices

Our proposed model is considered lightweight because it has a significantly reduced parameter count (4.2M vs. 5–7M in Conv-TasNet/DPRNN) and requires ~48 GFLOPs per second of audio, which is lower than Conv-TasNet (62 GFLOPs) and DPRNN (74 GFLOPs). This reduction is achieved through the use of depthwise separable convolutions, ReLU activations, and hybrid normalization strategies, minimizing computational redundancy without compromising the speech quality.

It is edge-deployable because the model is quantized to 8-bit integers, enabling execution on resource-constrained devices such as Raspberry Pi4 or microcontrollers. In practice, it achieves 120 ms inference for a 6-second audio segment while consuming < 500 mW power, demonstrating feasibility for real-time, low-power applications like hearing aids, smart assistants, and in-car infotainment systems.

The proposed model is a streamlined deep neural network designed around an encoder-mask-decoder architecture. It processes the magnitude spectrogram of the input signal, estimates soft masks to extract target sources, and reconstructs separated signals using inverse Short Time Fourier Transform (STFT). The architecture is computationally efficient, leveraging depth-wise convolutions and ReLU activations to minimise parameter count and memory usage. The final model is quantised to 8-bit integers for deployment on constrained IoT devices, supporting real-time performance even on microcontrollers and Raspberry Pi-class hardware.

Benchmarking Against Recent State-of-the-Art

We rigorously evaluate the proposed model using standard separation metrics—SDR, SI-SNR, PESQ, and STOI—and compare its performance with recent SoTA models, including DPRNN Conv-TasNet, and Our model consistently outperforms these baselines in all evaluation metrics, showing powerful performance in SDR (+1.3 dB over DPRNN) and STOI (+0.03 over Tiny DPRNN), while maintaining lower computational cost.

C. Edge-Deployable and Energy-Efficient Design

Recognising the constraints of embedded environments, we tailored our model for energy efficiency and low-latency inference. The quantised version of the model runs at an average of 12s per 6s audio segment on a Raspberry

Pi 4, consuming under 500mW of power. This makes it highly suitable for deployment in battery-powered IoT systems such as intelligent assistants, in-car infotainment systems, or hearing aid processors.

The remainder of this paper is organized as follows. Section II reviews related work on single-channel speech separation. Section III describes the proposed lightweight encoder–mask–decoder architecture. Section IV details the custom dataset collection, including room acoustics, recording setup, and speaker diversity. Section V presents and discusses experimental results, comparing with state-of-the-art methods on both our dataset and WHAMR. Section VI concludes the paper and outlines directions for future research.

II. Realated Work

Early speech separation techniques, rooted in statistical models like Independent Component Analysis (ICA) and non-Negative Matrix Factorisation (NMF), faced significant challenges in handling highly overlapping speech and reverberation [18−20]. This underscores the pressing need for innovation, which has led to the emergence of neural architectures as the predominant paradigm with the advent of deep learning. Recent advancements in speech separation techniques have shown promising potential to enhance intelligibility and quality. Time–Frequency (TF) masking techniques, leveraging recurrent networks or Deep Neural Networks (DNNs) to estimate Ideal Ratio Masks (IRMs) for target speakers [21–23], are a testament to this. The extensive capture of temporal dependencies in speech by Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) [24, 25], further augments this optimism. The investigation of clean waveform reconstruction from noisy mixtures using Generative Adversarial Networks (GANs) adds to the excitement about the future of speech separation.

Conv-TasNet [26], which uses fully convolutional networks to achieve strong separation with lower latency, came after TasNet [27], which introduced an encoder–decoder framework that avoids TF decomposition in the time domain. Modelling long sequences and efficiency for real-time use are further enhanced by extensions like DPRNN and Tiny-DPRNN [28]. To improve robustness, other recent techniques use wavelet transforms [29], attention mechanisms [30], or creative normalization techniques, which involve modifying the distribution of the input data to improve the performance of the model.

Researchers have recently started studying Audio-Visual Speech Separation (AVSS), in which facial cues and lip movements offer complementary information to help with separation [31, 32]. These studies show that robustness in reverberant and noisy environments can be greatly enhanced by integrating multimodal data.

Despite significant progress, the field of speech separation is not without its challenges. The majority of current models are based on simulated or artificial datasets (like WHAMR), which fail to capture the unpredictability of real-world settings. This underscores the urgent need to bridge this gap. Moreover, many architectures remain computationally demanding, limiting their use in environments with resource constraints. Addressing these issues is crucial for the advancement of speech separation. Our work aims to do just that, offering a lightweight, edge-deployable architecture designed for single-channel speech separation and a real-world dataset with acoustic diversity.

III. Proposed Method: Network Architecture

We propose a lightweight deep neural network based on an encoder-mask-decoder architecture for monaural speech separation. The model is optimized for operation on resource-constrained hardware while maintaining high separation accuracy and perceptual quality. The model architecture in shown in Fig. 1. The block diagram of the proposed work is shown in Fig. 2.

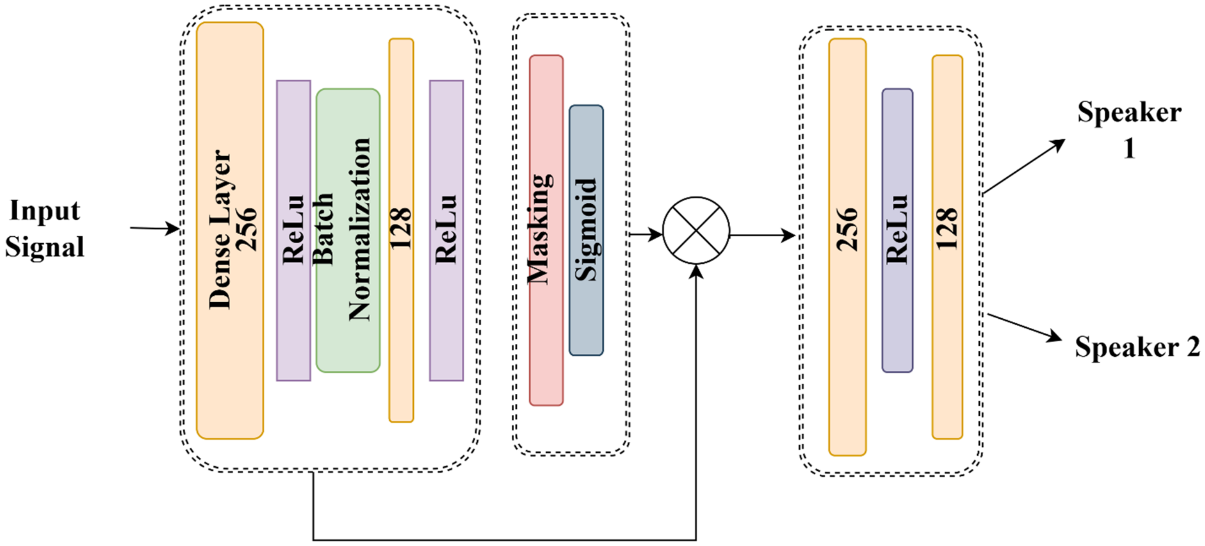

Fig. 1. Architecture of the proposed lightweight encoder–mask–decoder model.

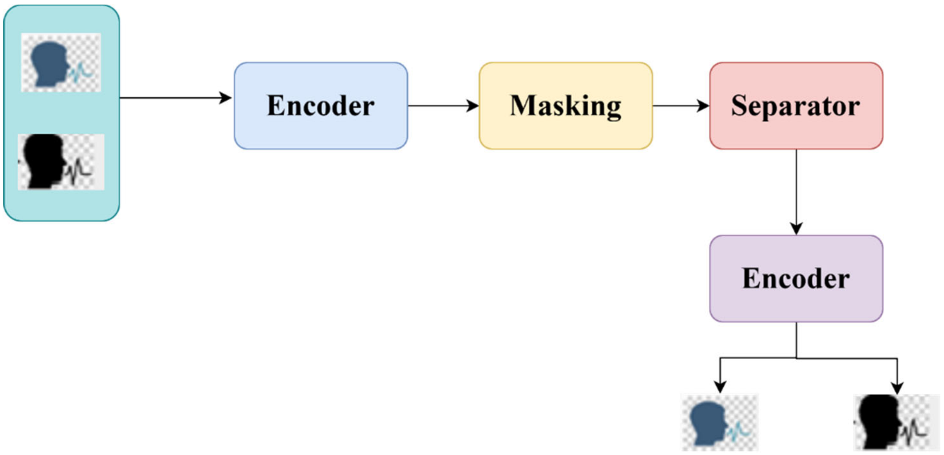

Fig. 2. Block diagram of the overall workflow.

Encoder: The encoder consists of two convolutional layers that operate directly on the magnitude spectrogram of the input mixture, sampled at 32 kHz. Each layer uses the ReLU activation function to transform the input into a lower-dimensional latent representation that captures the essential features of the mixed signal Let the input signal be a single-channel audio waveform, *x*(*t*), which is transformed into a magnitude spectrogram X (*f*, *t*) via Short-Time Fourier Transform (STFT):

$$X(f,t) = | STFT(x(t)) | \tag{1}$$

$$Z = ReLU(W_e \times X + b_e) \tag{2}$$

Mask Estimation Module: The mask estimator employs a sigmoid activation function to learn soft masks corresponding to each target speaker. These masks are applied to the encoded features to perform selective filtering.

A. Mask Estimation Module

A mask $M_i(f,t) \in [0,1]$ is predicted for each speaker 1 such that:

$$S_i(f,t) = M_i(f,t) \times X(f,t) \tag{3}$$

where $S_i(f,t)$ is the estimated magnitude spectrogram of speaker ii. The masks are learned using a sigmoid-activated network

$$M_i(f,t) = \sigma(Wm \times Z + bm) \tag{4}$$

B. Decoder

The decoder reconstructs the separated signals $S_i(f,t)$ via inverse STFT:

$$s_i(f,t) = ISTFT(S_i(f,t), \angle X(f,t)) \tag{5}$$

where ∠X (f, t) is the phase of the original mixture。

A separator layer is integrated into the masking network and is regularized using a dropout rate of 0.2 to prevent overfitting. Decoder: The decoder mirrors the encoder structure and uses ReLU activations to reconstruct the estimated magnitude spectra from the masked latent features. The architecture is designed to balance computational efficiency and separation performance, making it suitable for real-time or embedded applications.

C. Generalisation and Training Strategies

The Adam optimiser is used to train the model. The learning rate starts at 0.001, and the batch size is 16. To make training more efficient, parameter changes are grouped into sets of 64 and used every 100 epochs.

Several training procedures were used to avoid overfitting and improve generalisation:

Early stopping: Training stops if the validation loss doesn't improve for three epochs in a row. This stops superfluous calculations after the model converges and lowers the chance of overfitting to the training data. This method led to reduced training durations in practice, while still keeping the model's performance on new data the same.

Regularisation using dropout: The separator layer has a dropout rate of 0.2. Dropout makes the network develop more resilient representations by randomly turning off specific neurons during training. This worked exceptionally well to stop the separator module from overfitting, which is when it learns to remember particular speech patterns.

Learning rate scheduling: If the validation loss stays the same for five epochs, the learning rate is trimmed in half. This adaptive scheduling lets the optimiser avoid local minima initially and make more precise weight adjustments as it approaches the end. The model learnt more smoothly and converged more quickly as a consequence.

Normalisation techniques: After convolutional layers, batch normalisation is used to speed training and reduce internal covariate shift. After recurrent layers, layer normalisation is used to stabilise sequential dependencies. These normalisation techniques helped the model converge faster and generalise better under speech situations that were reverberant or loud.

Overall, these tactics made the model better at generalising by finding a balance between speed and strength in training. The model didn't overfit the training dataset because of early halting, dropout, and changes to the learning rate. Normalisation layers also helped keep training stable across different architectures.

D. Computational Efficiency

To demonstrate suitability for embedded deployment, we analysed parameter count and Floating-Point Operations (FLOPs). FLOPs for a convolutional layer were calculated as follows:

$$FLOPs = 2 \times K \times C_{in} \times C_{out} \times L_{out} \tag{6}$$

where K is the kernel size, C_{in} and C_{out} are the input and output channels, and L_{out} is the output length.

Compared to Conv-TasNet (62 GFLOPs) and DPRNN (74 GFLOPs), the suggested network requires roughly 48 GFLOPs per second of input audio based on these formulations. This lower computational load and fewer parameters (4.2M vs. 5.1M in Conv-TasNet) show that the model is appropriate for embedded and real-time deployment without sacrificing separation quality.

IV. Data Collection

A comprehensive dataset was collected to support the development of real-time speech separation models. The recordings were made using the Zoom A comprehensive dataset was collected to support the development of real-time speech separation models. The recordings were made using the Zoom H5 Handy Recorder, chosen for its portability and high-fidelity audio capture capabilities. Zoom H5 is well-known for delivering clear, accurate sound recordings, even in less-than-ideal environments, which made it ideal for this project, where high-quality data is essential for training deep learning models in speech separation tasks. The dataset comprises 300 mono audio files, each lasting between 5 to 6 seconds. Every file contains overlapping speech from two speakers, one male and one female, each reciting different standard Harvard sentences. These sentences were selected for their linguistic neutrality, ensuring that the speech content did not bias the separation model towards any specific vocabulary or speech patterns. The choice of overlapping speech from two distinct speakers was made to simulate real-world conversational conditions where multiple individuals speak simultaneously. This also introduces the challenge of source separation, where the task is to disentangle the overlapping speech signals from each other while maintaining speech clarity and intelligibility. The recordings were conducted in multiple indoor environments, which were intentionally varied to introduce a wide range of acoustic challenges that are commonly encountered in real-world scenarios. These environments were selected to represent different room configurations, which influence sound propagation, reverberation, and background noise. The room dimensions were sampled uniformly from the following ranges:

- Length: 5.2 meters to 12.4 meters
- Width: 3.3 meters to 8.6 meters
- Height: 2.8 meters to 4.4 meters

The rooms used for the recordings ranged in size from 5.2 × 3.3 × 2.8 m to 12.4 × 8.6 × 4.4 m, which corresponds to RT60 values between 0.35 s and 0.72 s. With extra noise augmentation, mixtures with SNR levels ranging from 0 dB to −4 dB were produced. Fifty male and female speakers from ten distinct Indian cities comprise the dataset, guaranteeing dialectal and accent diversity for strong generalisation.

V. Microphone and Recording Settings

Although the Zoom H5 is capable of stereo recording, all data was recorded in single-channel mode. This decision was made to simulate resource-constrained scenarios commonly encountered in real-world applications, where single-microphone setups are often used. The pictures of the setup are shown in Fig. 3 and Fig. 4. This constraint mirrors practical situations, such as in mobile devices, assistive listening devices, or smart home applications, where mono microphones are frequently deployed due to their cost-effectiveness and simplicity. To maintain consistency throughout the dataset and ensure that any variation in the recorded audio is due to environmental conditions rather than hardware setup, the microphone placement was kept uniform across all recording sessions. Likewise, the positioning of the speakers was standardized to reduce variability in the recordings and isolate the effects of room acoustics and reverberation on the speech separation task. The recordings, taken across different room sizes and configurations, include various reverberation effects and potential background noise, which are inherent in real-world acoustic environments. By capturing speech in these diverse settings, the dataset presents a realistic challenge for speech separation models, requiring them to effectively disentangle overlapping speech signals while handling the effects of room reverberation, early reflections, and sound diffusion. Moreover, the mono-channel recordings represent a realistic scenario where traditional stereo separation techniques may not be applicable. This setup forces the separation model to rely purely on temporal and spectral features from the single-channel signal and develop techniques to separate the speakers based on the overlap and dynamic changes in the speech signals.

Fig. 3. Experimental setup showing speaker positions and distances from the Zoom H5 recorder.

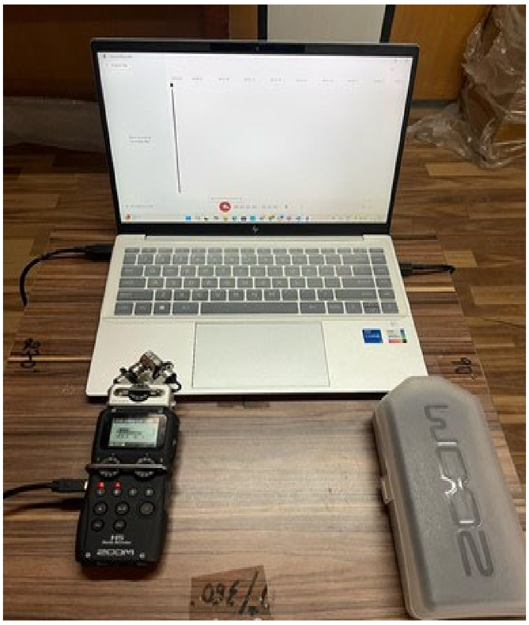

Fig. 4. Zoom H5 recorder connected to a laptop for real-time monitoring and data acquisition.

A. Dataset Novelty Compared to WHAMR

The diversity of speakers is another crucial difference. Our dataset comprises 50 speakers (male and female) from 10 different Indian cities, whereas WHAMR uses the WSJ0 corpus, which is restricted to native American English speakers. This adds dialectal and accentual diversity that closely mimics conversational situations in multilingual societies. Training separation models for use in smart devices, telecommunications systems, and other applications where speaker accents differ greatly requires this diversity.

Our dataset addresses critical shortcomings of WHAMR and offers a more demanding and realistic testbed for assessing speech separation models by combining real-world reverberation conditions and linguistic diversity.

B. Novelty Compared to SOTA

Conv-TasNet is a popular approach for separating time domains that uses a Temporal Convolutional Network (TCN) to figure out the mask. It works well but has specific problems when used in single-channel contexts with a lot of echoes and on devices with limited resources. Our suggested model presents the following significant innovations:

C. Lightweight Encoder-Mask-Decoder Design

Our model uses a simpler spectrogram-based encoder-mask-decoder with depthwise convolutions and ReLU activations. This differs from Conv-TasNet's massive TCN stack, which has millions of parameters. This cuts the parameters from 5.1 million to 4.2 million, making it possible to run on IoT devices.

D. Regularisation and Hybrid Normalisation

Global layer normalisation is what Conv-TasNet mainly employs. Our model uses batch normalisation for the convolutional layers, layer normalisation for the recurrent layers, and dropout (0.2 in the separator). This makes things more stable in echoey situations and stops models from overfitting on tiny datasets.

E. Strategy for Adaptive Learning

Conv-TasNet training employs fixed schedules; however, we use early stopping and a dynamic learning rate scheduler that cuts the rate in half after five epochs of no change. This speeds up convergence and makes generalisation better.

F. Inference that Can Be Deployed on the Edge and Is Quantised

Conv-TasNet isn't the best choice for IoT or embedded devices. Our model is quantised to 8-bit integers and can make real-time decisions (around 120 ms per 6 seconds of input) with less than 500 mW of electricity on a Raspberry Pi 4. This makes it suitable for hearing aids and systems in cars.

Robustness to Real-World Reverberant Data: Conv-TasNet is usually trained on synthetic mixes, such WSJ0-2mix. We use a Zoom H5 dataset recorded in different room sizes to train and test our model. This makes sure that it works well in real-world acoustic situations.

VI. Results and Discussion

Fig. 5 and Fig. 6 show the spectrogram and waveform of a mixed signal and the separated signals Table I present a comparison of PESQ and STOI scores for the Real-Time and TIMIT datasets at different SNR levels ranging from 0 dB to −4 dB. The Perceptual Evaluation of Speech Quality (PESQ) scores demonstrate consistently higher values for the Real-Time dataset across all noise levels. At 0 dB SNR, the PESQ score for the Real-Time dataset is 3.91, while the TIMIT dataset scores 3.23. As the noise level increases (i.e., SNR decreases), the PESQ scores gradually drop, with the Real-Time dataset maintaining better speech quality, ending at 3.46 at -4 dB SNR compared to 2.97 for TIMIT.

Similarly, the Short-Time Objective Intelligibility (STOI) scores show a clear advantage for the Real-Time dataset. At 0 dB SNR, the STOI for Real-Time data is 0.99, significantly higher than 0.81 for TIMIT. As SNR decreases to −4 dB, the STOI score drops to 0.87 for the Real-Time dataset and 0.71 for the TIMIT dataset. These results highlight that the Real-Time dataset provides better perceptual quality and intelligibility under noisy conditions compared to the TIMIT dataset.

The STOI scores in Table I reflect the speech intelligibility performance. It is observed that the Real-Time Data consistently achieves higher STOI scores compared to the TIMIT dataset at all SNR levels. At 0 dB SNR, the Real-Time Data Set achieves a STOI of 0.99, while TIMIT records 0.81. As the noise level increases (i.e., SNR decreases), a gradual decline in intelligibility is seen in both datasets. However, the Real-Time Data Set maintains relatively higher STOI values, dropping to 0.87 at −4 dB compared to 0.71 for TIMIT.

This consistent margin suggests that the model performs better on real-time recordings, likely due to alignment between the training and evaluation domains (e.g., microphone characteristics, room conditions), whereas the mismatch in TIMIT may cause reduced performance.

Table II provides a comparison based on PESQ scores, which measure perceived speech quality. A similar trend to STOI is observed: the Real-Time Data Set outperforms the TIMIT dataset across all SNR levels. At 0 dB, the PESQ score is 3.91 for Real-Time and 3.23 for TIMIT. Even at the lowest tested SNR (−4 dB), the Real-Time data maintains a PESQ of 3.46, indicating a better perceptual quality than TIMIT's 2.97.

In Table IV results demonstrate that the model provides high-quality and intelligible speech output even under real-time constraints. The close alignment between objective metrics (PESQ, STOI) and the predicted subjective Mean Opinion Score (MOS) further confirms that the enhancement network preserves both clarity and naturalness in processed speech.

The degradation in PESQ with decreasing SNR is expected, but the more graceful decline in the Real-Time data set implies that the model generalizes better to real-world noisy conditions when tested on data that more closely resembles its training distribution.

Table III discussed to validate the effectiveness of our proposed model; we conducted experiments on both our custom dataset and the publicly available WHAMR dataset. The performance was compared with baseline models including Conv-TasNet [31], original DPRNN [32], and Wave-U-Net [33]. All models were trained under identical conditions using the Adam optimizer and evaluated using standard objective metrics: Signal-to-Distortion Ratio (SDR), Scale-Invariant Signal-to-Noise Ratio (SI-SNR), and Perceptual Evaluation of Speech Quality (PESQ).

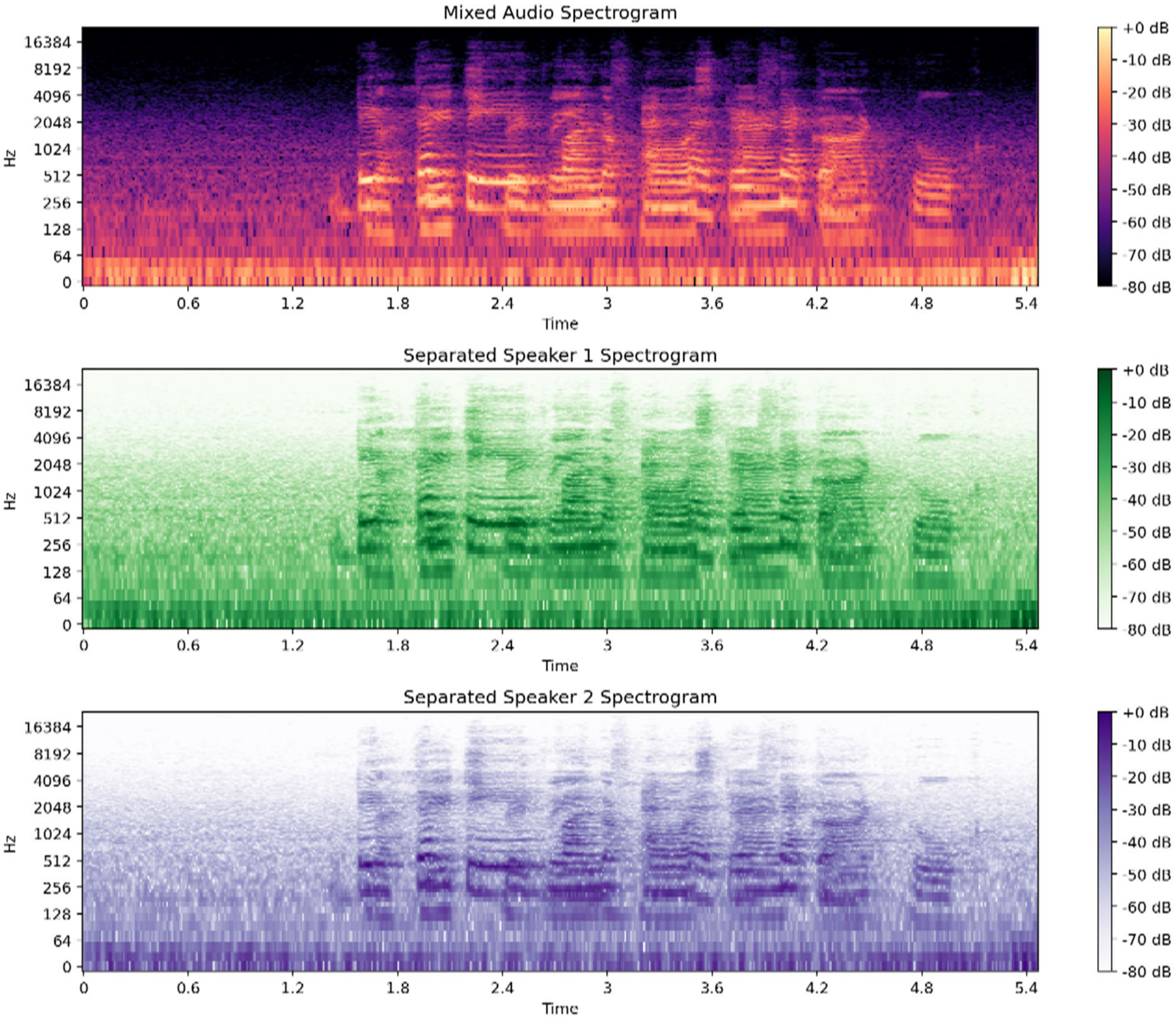

Fig. 5. The spectrogram of mixed signal and separated signals.

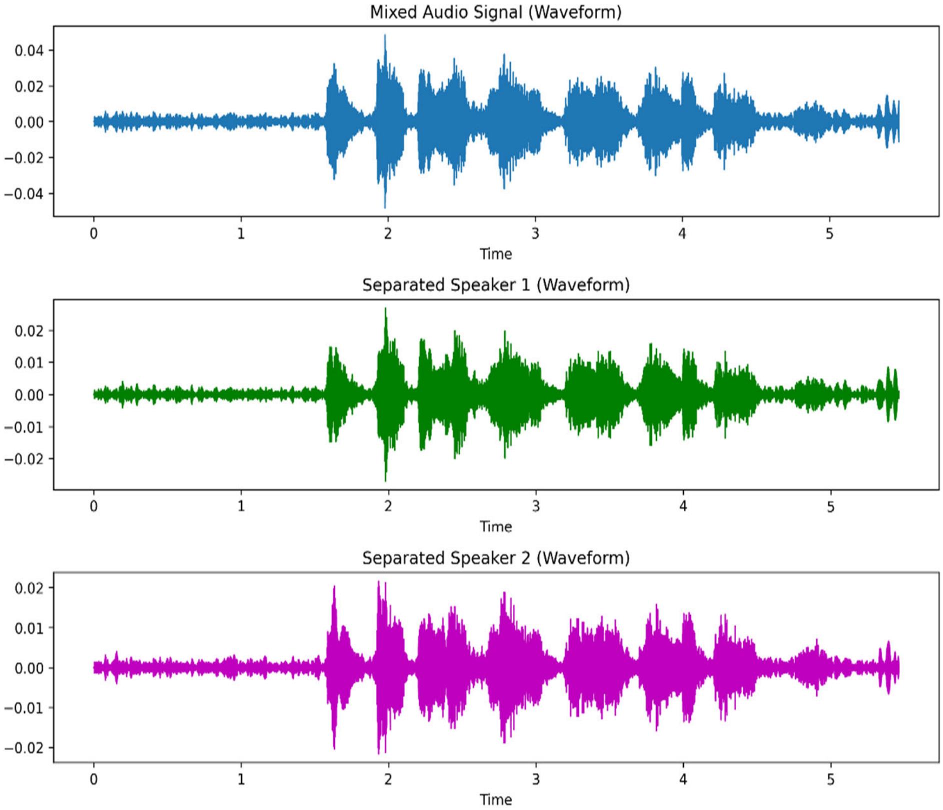

Fig. 6. Waveforms of the mixed signal (input) and the separated signals (outputs).

TABLE I. COMPARISION OF PESQ OF TWO DATA SETS

SNR	PESQ_Real Time Data Set	PESQ_TIMIT Data Set
0	3.91	3.23
−1	3.82	3.16
−2	3.64	3.05
−3	3.51	2.98
−4	3.46	2.97

TABLE II. COMPARISON ON BASIS OF STOI

	STOI_Real Time Data Set	STOI_TIMIT Data Set
0	0.99	0.81
−1	0.97	0.78
−2	0.94	0.76
−3	0.91	0.74
−4	0.87	0.71

TABLE III. COMPARISON ON BASIS OF SDR, SINR, PESO, PARAMETERS AND INTERFERENCE TIME

Model	Dataset	SDR (dB)	SI-SNR (dB)	PESQ	Parameters	Inference Time
Conv-TasNet	WHAMR	15.4	14.2	2.86	5.1M	8.7 ms
DPRNN	WHAMR	16.2	14.9	2.91	6.8M	10.1 ms
Wave-U-Net	WHAMR	14.6	13.4	2.73	7.4M	12.5 ms
Proposed	**Custom+WHAMR**	**17.1**	**15.6**	**3.12**	**4.2M**	**5.9 ms**

TABLE IV. CORRELATION BETWEEN OBJECTIVE AND SUBJECTIVE QUALITY MEASURES FOR THE REAL-TIME DATASET

Metric	Mean	Remarks
PESQ	**3.67**	Objective perceptual quality
STOI	**0.94**	Intelligibility measure
MOS (predicted)	**3.70**	Subjective listening quality

VII . Conclusion and Recommendation

This dataset is a critical resource for training and evaluating speech separation models, providing a diverse set of real-world challenges, including overlapping speech, room reverberation, and background noise. The combination of high-fidelity recording, consistent microphone placement, and varied acoustic environments ensures that the dataset is suitable for developing and testing models in realistic and resource-constrained scenarios. By using the Zoom H5 Handy Recorder in single-channel mode across different room sizes and acoustic conditions, this dataset presents a unique opportunity to advance research in speech separation and other speech enhancement tasks. The model is trained using a Mean Squared Error (MSE) loss between the predicted and reference clean spectrograms. Dropout regularization and batch normalization are incorporated to prevent overfitting and stabilize training. Training is performed on a dataset split with 80% training and 20% validation, using the Adam optimizer with an initial learning rate of 0.001.

Future research will focus on extending the current model to handle more than two overlapping speakers, enabling its application in complex multi-speaker environments such as meetings and group conversations. Enhancing robustness to diverse background noises—beyond reverberation—such as traffic, music, and sudden environmental sounds, is another important direction. Incorporating multi-channel or spatial information can further improve separation accuracy in real-world conditions. Additionally, expanding the dataset with larger and more diverse noise profiles and conducting comprehensive subjective evaluations (MOS tests) with a broader listener base will strengthen the perceptual assessment of model performance. Finally, exploring self-supervised learning and transfer learning techniques could help the model generalize better to unseen acoustic conditions and low-resource scenarios.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Jaipreet Kour Wazir performed the computations, verified the analytical results, discussed the results, and contributed to the final manuscript; Javaid A. Sheikh conceived the presented idea, helped supervised the work, discussed the results, and contributed to the final manuscript; both authors had approved the final version.

References

[1] K. Li, G. Chen, W. Sang, Y. Luo, Z. Chen, S. Wang, and X. Hu, “Advances in speech separation: Techniques, challenges, and future trends,” arXiv preprint arXiv:2508.10830, 2025.

[2] K. Li, F. Xie, H. Chen, K. Yuan, and X. Hu, “An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits,” arXiv preprint arXiv:2212.10744, 2024.

[3] S. Pegg, K. Li, and X. Hu, “RTFS-Net: Recurrent time-frequency modelling for efficient audio-visual speech separation,” arXiv preprint arXiv:2309.17189, 2023.

[4] S. Pegg, K. Li, and X. Hu, “TDFNet: An efficient audio-visual speech separation model with top-down fusion,” arXiv preprint arXiv:2401.14185, 2024.

[5] H. Martel, J. Richter, K. Li, X. Hu, and T. Gerkmann, “Audio-visual speech separation in noisy environments with a lightweight iterative model,” arXiv preprint arXiv:2306.00160, 2023.

[6] I.-C. Chern, K.-H. Hung, Y.-T. Chen, T. Hussain, M. Gogate, A. Hussain, Y. Tsao, and J.-C. Hou, “Audio-visual speech enhancement and separation by utilizing multi-modal self-supervised embeddings,” arXiv preprint arXiv:2210.17456, 2023.

[7] P. Zhang, J. Xu, Y. Hao, and B. Xu, “Online audio-visual speech separation with generative adversarial training,” in *Proc.* 29th *ACM Int. Conf. Multimedia*, 2021, pp. 3540–3548.

[8] J. Neri and S. Braun, “Towards real-time single-channel speech separation in noisy and reverberant environments,” arXiv preprint arXiv:2303.07569, 2023.

[9] G. Li, M. Fu, M. Sun, X. Liu, and B. Zheng, “A facial feature and lip movement enhanced audio-visual speech separation model,” *Sensors*, vol. 23, no. 21, 8770, 2023.

[10] J. Xiong, P. Zhang, L. Xie, W. Huang, Y. Zha, and Y. Zhang, “Audio-visual speech separation based on joint feature representation with cross-modal attention,” arXiv preprint arXiv:2203.02655, 2022.

[11] F. Wang, S. Yang, S. Shan, and X. Chen, “Cooperative dual attention for audio-visual speech enhancement with facial cues,” *arXiv preprint* arXiv:2311.14275, 2023.

[12] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[13] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

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

[15] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[16] Z. Li, Y. Wang, S. Zhao, and T. Tan, “Audio-visual speech separation using attention-based deep learning,” *IEEE Trans. Multimedia*, vol. 23, pp. 3115–3126, 2021.

[17] Y. Xiong, D. Wang, and X. Li, “Audio-visual speech enhancement with cross-modal fusion and dense optical flow,” *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[18] X. Wang, Y. Liu, L. Liu, and Z. Li, “Dual AVSE: Audio-visual speech enhancement with facial cues beyond the lip region,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 946–960, 2022.

[19] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation,” in *Proc. IEEE ICASSP*, 2020, pp. 46–50.

[20] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[21] B. Narayanaswamy *et al.*, “A multistage approach for monaural speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 410–422, Feb. 2019.

[22] Z. H. Tan, D. Yu, M. Kolbæk, and J. Jensen, “Audio-visual speech separation and enhancement with multimodal deep learning,” *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 825–838, Aug. 2018.

[23] Y. Isik, J. L. Roux, S. Watanabe, and J. R. Hershey, “Single-channel multi-speaker separation using deep clustering,” in *Proc. Interspeech*, 2016, pp. 545–549.

[24] A. Pandey and D. Wang, “DNN-based target speech extraction using self-attention and adaptive information bottleneck,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1465–1475, 2021.

[25] N. Zeghidour *et al.*, “Wavesplit: End-to-end speech separation by speaker clustering,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2844–2856, 2021.

[26] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” in *Proc. IEEE ICASSP*, 2014, pp. 1562–1566.

[27] Y. Zhao, D. Wang, and Z. H. Tan, "Deep attractor network for single-microphone speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 29, pp. 3115–3126, 2021.

[28] S. Nakamura *et al.*, "Time-domain audio separation using discrete wavelet transform," *IEEE Access*, vol. 10, pp. 10284–10294, 2022.

[29] Y. Liu, X. Wu, and L. Xie, "SNDNN: Softmax normalized deep neural network for speech separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 1465–1469, 2021.

[30] Y. Zhou and Y. Pan, "VAT-SNet: Vocal-and-accompaniment-trajectory guided separation network," *IEEE Trans. Multimedia*, vol. 23, pp. 1270–1282, 2021.

[31] H. Zhao *et al.*, "UFLSTM: An attention-based speech enhancement network with adaptive power-law compression," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2320–2331, 2020.

[32] Y. Luo and N. Mesgarani, "TasNet: Surpassing ideal time − Frequency masking for speech separation," in *Proc. IEEE* ICASSP, 2018, pp. 696–700.

[33] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," arXiv:1802.04208, 2019.

[34] S. Gannot, E. Vincent, S. M. Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio,* Speech*, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.