Speech Separation: Addressing the Cocktail Party Problem with Deep Learning

Ravi Saxena(P23CS0006) , Aditi Baheti(M23CSA001)

Speech Understanding : Assignment 1

1 Introduction

In real-world environments, multiple people often speak simultaneously, making it difficult for machines to isolate individual speech signals. This challenge, known as the *Cocktail Party Problem*, is critical for applications such as automatic speech recognition (ASR), telecommunication systems, hearing aids, and voice-controlled assistants. Traditional methods like Independent Component Analysis (ICA) and Non-Negative Matrix Factorization (NMF) rely on signal processing techniques but struggle in complex, noisy scenarios with overlapping speech.

Deep learning-based approaches have significantly improved speech separation by directly learning representations from raw audio. This report explores three advanced models: **Conv-TasNet**, a fully convolutional time-domain model; **SepReformer**, an asymmetric encoder-decoder approach; and **SepTDA**, a transformer-based model capable of handling an unknown number of speakers. These models are evaluated on benchmark datasets such as *WSJ0-2Mix, WSJ0-3Mix, WHAM!, and Libri2Mix*, demonstrating superior performance in real-world speech separation tasks.

2 Datasets

Evaluating speech separation models requires datasets with mixtures of multiple speakers. This report focuses on benchmark datasets widely used in deep-learning-based speech separation research. These datasets provide controlled multi-speaker mixtures with ground truth source signals, enabling model training and evaluation.

2.1 WSJ0-2Mix and WSJ0-3Mix

WSJ0-2Mix and WSJ0-3Mix are constructed from the Wall Street Journal (WSJ0) corpus, containing recordings of professional speakers. **WSJ0-2Mix** includes two-speaker mixtures, while **WSJ0-3Mix** extends this to three-speaker scenarios. Speech segments are randomly selected and mixed at signal-to-noise ratios (SNRs) ranging from -5 dB to 5 dB.

2.2 WHAM! and WHAMR!

WHAM! extends WSJ0-2Mix by introducing real-world background noise recorded from cafes, restaurants, and other urban environments. WHAMR! further adds reverberation effects, making separation more challenging by simulating realistic acoustic conditions.

2.3 Libri2Mix

Libri2Mix is derived from the LibriSpeech dataset, containing speech from audiobook recordings. It provides clean two-speaker mixtures and serves as an alternative benchmark to WSJ0-2Mix. The dataset is created by normalizing loudness levels and randomly combining speakers.

These datasets provide diverse testing environments, enabling robust evaluation of speech separation models under clean, noisy, and reverberant conditions.

3 State-of-the-Art Speech Separation Models

3.1 SepReformer: Asymmetric Encoder-Decoder for Speech Separation

3.1.1 Overview

SepReformer [1] introduces an *early feature separation strategy*, enhancing speaker differentiation at the encoding stage. It employs **Global and Local Transformer Blocks** to efficiently process long speech sequences while maintaining computational efficiency.

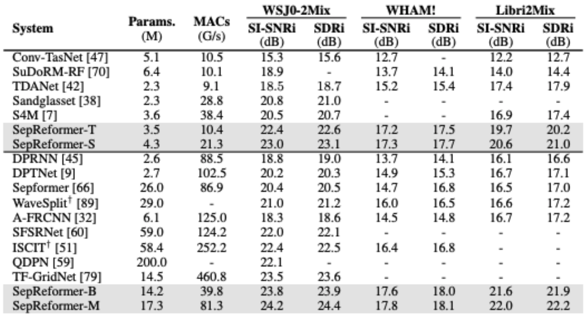
3.1.2 Architecture

Key components include:

- **Separation Encoder**: Splits input speech into speaker-specific latent representations.
- **Speaker Split Module**: Uses *Gated Linear Units (GLU)* for distinct speaker feature extraction.
- **Reconstruction Decoder**: A weight-sharing structure that refines separated waveforms.
- **Transformer Blocks**: *Efficient Global Attention (EGA)* for long-range dependencies and *Convolutional Local Attention (CLA)* for fine-grained features.

3.1.3 Evaluation and Results

Tested on **WSJ0-2Mix, WHAM!/WHAMR!, and Libri2Mix**, SepReformer achieves high **SI-SNRi** and **SDRi**, surpassing prior methods with reduced computational cost.



System	Params. (M)	MACs (G/s)	WSJ0-2Mix SI-SNRi (dB)	WSJ0-2Mix SDRi (dB)	WHAM! SI-SNRi (dB)	WHAM! SDRi (dB)	Libri2Mix SI-SNRi (dB)	Libri2Mix SDRi (dB)
Conv-TasNet [47]	5.1	10.5	15.3	15.6	12.7	-	12.2	12.7
SuDoRM-RF [70]	6.4	10.1	18.9	-	13.7	14.1	14.0	14.4
TDANet [42]	2.3	9.1	18.5	18.7	15.2	15.4	17.4	17.9
Sandglasset [38]	2.3	28.8	20.8	21.0	-	-	-	-
S4M [7]	3.6	38.4	20.5	20.7	-	-	16.9	17.4
SepReformer-T	3.5	10.4	22.4	22.6	17.2	17.5	19.7	20.2
SepReformer-S	4.3	21.3	23.0	23.1	17.3	17.7	20.6	21.0
DPRNN [45]	2.6	88.5	18.8	19.0	13.7	14.1	16.1	16.6
DPTNet [9]	2.7	102.5	20.2	20.3	14.9	15.3	16.7	17.1
Sepformer [66]	26.0	86.9	20.4	20.5	14.7	16.8	16.5	17.0
WaveSplit† [89]	29.0	-	21.0	21.2	16.0	16.5	16.6	17.2
A-FRCNN [32]	6.1	125.0	18.3	18.6	14.5	14.8	16.7	17.2
SFSRNet [60]	59.0	124.2	22.0	22.1	-	-	-	-
ISCIT† [51]	58.4	252.2	22.4	22.5	16.4	16.8	-	-
QDPN [59]	200.0	-	22.1	-	-	-	-	-
TF-GridNet [79]	14.5	460.8	23.5	23.6	-	-	-	-
SepReformer-B	14.2	39.8	23.8	23.9	17.6	18.0	21.6	21.9
SepReformer-M	17.3	81.3	24.2	24.4	17.8	18.1	22.0	22.2

Figure 1: Performance comparison of SepReformer on standard datasets.

3.2 SepTDA: Transformer Decoder-Based Attractor for Unknown Speaker Count

3.2.1 Overview

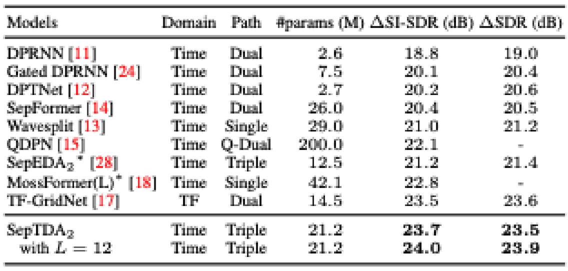
SepTDA [2] addresses *unknown-number speaker separation* by integrating a **Transformer Decoder-Based Attractor (TDA)**, dynamically estimating speaker embeddings.

3.2.2 Architecture

- **Encoder**: 1D CNN with *GELU activation* for feature extraction.
- **Separation Module**: *Dual-Path and Triple-Path Processing* improve speaker interaction.
- **TDA**: Learns speaker queries to detect and separate an unknown number of sources.

3.2.3 Evaluation and Results

Evaluated on **WSJ0-2Mix, WSJ0-3Mix, WSJ0-4Mix, and WSJ0-5Mix**, it achieves **24.0 dB SI-SDRi** and outperforms models like SepFormer and Wavesplit.



Models	Domain	Path	#params (M)	ΔSI-SDR (dB)	ΔSDR (dB)
DPRNN [11]	Time	Dual	2.6	18.8	19.0
Gated DPRNN [24]	Time	Dual	7.5	20.1	20.4
DPTNet [12]	Time	Dual	2.7	20.2	20.6
SepFormer [14]	Time	Dual	26.0	20.4	20.5
Wavesplit [13]	Time	Single	29.0	21.0	21.2
QDPN [15]	Time	Q-Dual	200.0	22.1	-
$\text{SepEDA}_2{}^*$ [28]	Time	Triple	12.5	21.2	21.4
MossFormer(L)* [18]	Time	Single	42.1	22.8	-
TF-GridNet [17]	TF	Dual	14.5	23.5	23.6
SepTDA_2	Time	Triple	21.2	**23.7**	**23.5**
with $L = 12$	Time	Triple	21.2	**24.0**	**23.9**

Figure 2: Performance comparison of SepTDA on datasets with varying speaker counts.

3.3 Conv-TasNet: Time-Domain Speech Separation

3.3.1 Overview

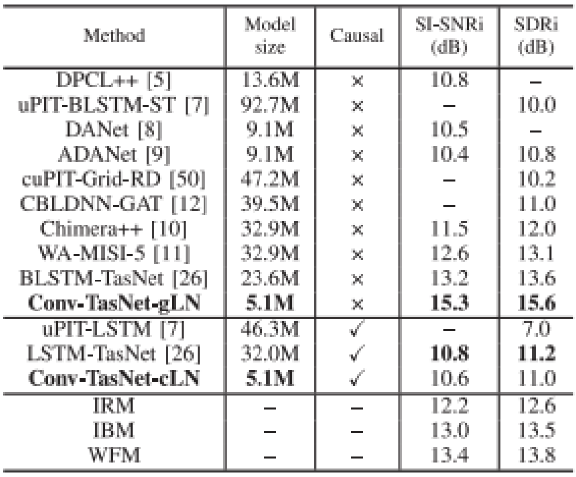
Conv-TasNet [3] operates directly in the **time domain**, eliminating STFT dependency and improving real-time separation efficiency.

3.3.2 Architecture

- **Encoder**: 1D CNN transforms waveform into latent representations.
- **Temporal Convolutional Network (TCN)**: *Stacked 1D dilated convolutional blocks* capture long-range dependencies.
- **Mask Estimation**: Predicts speaker-specific masks for separation.

3.3.3 Evaluation and Results

Achieves superior **SI-SNRi** and **SDRi** on **WSJ0-2Mix** and **WSJ0-3Mix**, outperforming traditional spectrogram-based methods.



Method	Model size	Causal	SI-SNRi (dB)	SDRi (dB)
DPCL++ [5]	13.6M	×	10.8	–
uPIT-BLSTM-ST [7]	92.7M	×	–	10.0
DANet [8]	9.1M	×	10.5	–
ADANet [9]	9.1M	×	10.4	10.8
cuPIT-Grid-RD [50]	47.2M	×	–	10.2
CBLDNN-GAT [12]	39.5M	×	–	11.0
Chimera++ [10]	32.9M	×	11.5	12.0
WA-MISI-5 [11]	32.9M	×	12.6	13.1
BLSTM-TasNet [26]	23.6M	×	13.2	13.6
Conv-TasNet-gLN	**5.1M**	×	**15.3**	**15.6**
uPIT-LSTM [7]	46.3M	✓	–	7.0
LSTM-TasNet [26]	32.0M	✓	**10.8**	**11.2**
Conv-TasNet-cLN	**5.1M**	✓	10.6	11.0
IRM	–	–	12.2	12.6
IBM	–	–	13.0	13.5
WFM	–	–	13.4	13.8

Figure 3: Performance comparison of Conv-TasNet on WSJ0 datasets.

4 Open Challenges and Research Opportunities

4.1 Open Challenges

4.1.1 Lack of Selective Voice Forgetting

Existing speech separation models excel at isolating individual speakers but lack the ability to *selectively forget* certain speakers from a given mixture. Once a voice is learned, current models do not provide mechanisms to erase speaker-specific features while preserving general speech intelligibility.

4.1.2 Accent and Speaker Bias in Speech AI

Speech separation and recognition models often exhibit **bias toward dominant accents and speaker demographics** due to imbalanced training datasets. This leads to unfair performance disparities, particularly in underrepresented languages and accents, making the models less robust in real-world applications.

4.1.3 Trade-off Between Unlearning and Speech Quality

Machine unlearning in speech AI poses a *fundamental trade-off*—removing speaker characteristics may inadvertently degrade speech intelligibility. Achieving an optimal balance between effective voice removal and natural-sounding speech reconstruction remains a major challenge.

4.2 Research Opportunities

4.2.1 Combining Cocktail Party Speech Separation with Machine Unlearning

Integrating speech separation techniques like **Conv-TasNet** with **Machine Unlearning** methods such as **AmnesiacML** [4] can enable selective removal of speaker identities from audio mixtures. This approach would allow systems to retain intelligible speech while eliminating targeted speaker attributes, addressing privacy concerns in voice-based AI applications.

4.2.2 Unified Loss Function for Separation and Unlearning

A promising direction is the development of a **joint optimization framework** that simultaneously minimizes:

- **Separation Loss** (for isolating speech sources effectively).
- **Feature Suppression Loss** (for unlearning speaker characteristics).

By balancing these objectives, the model can ensure **privacy-preserving speech separation** while maintaining high intelligibility.

4.2.3 Bias Mitigation Through Adaptive Unlearning

Unlearning techniques can be extended to **accent normalization and bias correction** by selectively reducing the model's reliance on speaker-specific traits. By dynamically adjusting training representations, future models can improve fairness in speech AI, reducing bias toward dominant language groups.

5 Conclusion

Speech separation has evolved significantly with deep-learning-based approaches, addressing the long-standing *Cocktail Party Problem.* Models such as **SepReformer**, **SepTDA**, and **Conv-TasNet** have demonstrated state-of-the-art (SOTA) performance in isolating individual speech signals from multi-speaker mixtures. These advancements have led to improved **signal-to-noise ratio (SI-SNRi)**, lower computational costs, and enhanced robustness across diverse datasets such as *WSJ0-2Mix, WHAM!, and Libri2Mix.*

Despite these improvements, several challenges remain. Current speech separation models do not support **selective speaker unlearning**, posing concerns for privacy and ethical AI deployment. Additionally, **bias in speaker accents** and the **trade-off between separation quality and computational efficiency** present open research problems.

In conclusion, deep learning has revolutionized speech separation, but the future lies in **privacy-preserving, unbiased, and interpretable AI models**. Addressing these challenges will enable more secure, fair, and high-quality voice-based applications in real-world settings.

References

[1] U. H. Shin, S. Lee, T. Kim, and H. M. Park, "Separate and reconstruct: Asymmetric encoder-decoder for speech separation," *arXiv preprint*, vol. arXiv:2406.05983, 2024. [Online]. Available: https://arxiv.org/abs/2406.05983

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

[3] Y. Lee, S. Choi, B. Y. Kim, Z. Q. Wang, and S. Watanabe, "Boosting unknown-number speaker separation with transformer decoder-based attractor," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2024, pp. 446–450.

[4] L. Graves, V. Nagisetty, and V. Ganesh, "Amnesiac machine learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, May 2021, pp. 11 516–11 524.