Deep Attractor Models for Audio Source Separation

Priya Shetty[1], Aliza Sayyad[2], Samruddhi Bhandare[3], Vaibhavi Dongare[4], Mrs. Asharani Chadchankar[5]
Information Technology, Marathwada Mitra Mandal's College of Engineering, SPPU, Pune

***Abstract*— Audio source separation has seen substantial advancements with deep learning techniques, yet separating unknown speakers in real-world environments remains challenging. This study proposes a deep learning-based approach using attractor models, which map audio signals into a high-dimensional feature space for improved clustering of individual audio sources. Through the introduction of attractor points, our approach creates a robust, speaker-independent separation system that distinguishes multiple audio sources with high precision. By clustering audio features around defined centroids, this model enables applications in automatic speech recognition (ASR), speaker identification, audio enhancement, and more.**

I. INTRODUCTION

1.1 Background

Audio source separation—the ability to isolate individual sound sources from a complex audio mixture—has numerous applications across fields like ASR (Automatic Speech Recognition), telecommunication, hearing aids, and surveillance. It helps us understand different sounds mixed together, such as separating voices from background noise. Despite progress in this domain, achieving reliable source separation in single-channel (monaural) recordings, especially with unknown speakers or overlapping sounds, is challenging. This means that when all sounds are recorded together in one audio track, it becomes difficult to pull them apart.

The classic "cocktail party problem," where multiple voices overlap in the same frequency range, complicates traditional methods that rely solely on spectral differences. This problem shows how hard it is to hear one person clearly in a noisy room full of talking people. Traditional techniques that depend on simple frequency differences often fail when voices or sounds are too similar. They cannot fully remove background sounds or separate speech when people speak at the same time.

With the rise of deep learning, neural networks have shown promise in overcoming these limitations by learning complex patterns in audio signals. Deep learning uses large datasets to teach computers how to find and learn these patterns on their own. It does not need simple rules; instead, it finds smart ways to understand and split sounds.

This research focuses on deep attractor models, a form of neural network that projects the time-frequency characteristics of a sound mixture into an embedding space. These models help in converting sound into a form that is easier for machines to work with. Here, unique reference points or "attractors" are established, allowing separation of each source by grouping audio features around these points. In simple terms, each sound gets pulled toward a special point that represents it, helping to sort it out from other sounds. These attractor points act like magnets that pull similar parts of the sound together.

Using this method makes it easier to separate sounds even when they are mixed tightly. It can work even when the voices are not known before or when there is a lot of noise. Deep attractor networks give better accuracy and performance compared to older methods. This technique can be helpful in real-life situations like phone calls, voice assistants, and security systems. It also supports people with hearing problems by improving the clarity of sounds. Overall, deep learning models like attractor networks offer a powerful solution to the hard problem of audio source separation.

1.2 Objectives

This study aims to:

1. Create and utilize a dataset for multi-speaker, single-channel audio recordings.
2. Develop a deep attractor network tailored to handle the separation of unknown audio sources.
3. Evaluate model performance across various separation metrics and applications.

1.3 Scope of the Project

By leveraging a high-dimensional, deep attractor network, this research contributes an advanced architecture for monaural audio source separation. Beyond speech separation, this method can serve as a preprocessing tool for applications requiring

isolated audio inputs, such as noise reduction, hearing aids, and multi-speaker conferencing systems.

II. LITERATURE REVIEW

Audio source separation has seen various strategies over the years, from conventional signal processing to state-of-the-art deep learning methods. Notable contributions in the field include:

[1] Deep Ensemble Learning for Monaural Speech Separation:
Zhang and Wang introduced ensemble learning in a DNN context, demonstrating improved monaural separation by integrating multiple network outputs to better capture spectral information.

[2] Regression-Based Approach to Single-Channel Separation:
Du et al. utilized a regression model to predict clean speech signals from noisy ones, a method that has shown flexibility and adaptability in real-world applications.

[3] Temporal Convolutional Networks (TCNs):
Lea et al. applied TCNs to segment fine-grained actions in video sequences, showcasing TCNs' capacity to model time sequences effectively. Similar methods are applied in audio, where TCNs improve audio source separation by modeling temporal dependencies.

[4] Evaluation of Convolutional and Recurrent Networks:
Bai and colleagues compared convolutional and recurrent architectures, observing that convolutional structures can outperform recurrent networks in sequence modeling tasks due to superior memory efficiency and reduced computational load.

[5] Multi-Channel Deep Clustering with Spectral and Spatial Embeddings:
Wang et al. proposed a clustering approach that combines spectral and spatial data, improving separation accuracy in environments with reverberation.

[6] WHAMR! Dataset for Noisy and Reverberant Speech Separation:
Maciejewski et al. developed WHAMR!, a comprehensive dataset including real-world noise and reverberation, providing a robust benchmark for evaluating separation models in challenging environments.

III. SYSTEM ANALYSIS AND PROPOSED ARCHITECTURE

3.1 System Overview

This research proposes a deep attractor model that uses a neural network to project audio mixtures into a high-dimensional embedding space. By defining attractor points, each corresponding to an individual audio source, the system enables clustering of time-frequency representations for efficient source separation.

3.2 Training and Testing Modules

The system is divided into training and testing modules, each responsible for distinct phases of the source separation process:

1. Training Module:

Data Preparation and Feature Extraction: The model processes multi-speaker data, extracting features that represent the time-frequency structure of each audio source.

Network Training: The neural network is trained to recognize and separate voices by associating each with an attractor point in the embedding space. This module fine-tunes model weights to improve accuracy in audio clustering and separation.

Attractor Creation: Attractor points are defined as centroids within the embedding space, dynamically adapting to the number and nature of sources in the training data.

2. Testing Module:

Feature Extraction from Mixed Audio: Given a new mixed audio input, the model extracts features and maps them into the embedding space.

Speaker Separation and Reconstruction: The attractor model then assigns audio features to their respective clusters based on proximity to attractor points. Separated audio signals are reconstructed by applying soft masks to isolate each source from the original mixture.

3.3 Deep Attractor Model Architecture

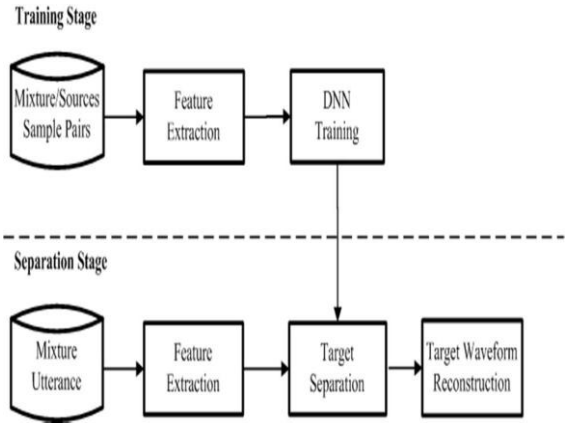
The architecture integrates:

Embedding Layers: These layers encode the input's time-frequency characteristics into a high-dimensional space, preserving subtle audio cues for accurate clustering.

Attractor-Based Clustering: Using pre-defined attractor points, the model assigns each audio frame to the nearest attractor, facilitating separation.

Soft Masking Mechanism: By applying separation masks, the model reconstructs isolated sources with minimal interference from other signals.

IV. PROPOSED SYSTEM ARCHITECTURE DIAGRAM



V. APPLICATIONS AND IMPLICATIONS

The deep attractor model's ability to perform real-time, speaker-independent separation has broad implications across industries. This means the model can work quickly and can separate voices even if it doesn't know who the speaker is. It can do this without needing special information about the speakers. This is useful in many different fields and situations. It can help make audio clearer and easier to understand.

Automatic Speech Recognition (ASR):
Preprocessing with a source separation model can significantly enhance ASR accuracy in noisy environments, such as meetings or public spaces. This means that if we clean the audio before giving it to the ASR system, it will work better. It can understand what people are saying more clearly. In places where there are many people talking or loud sounds, this model can help pick out the main voice. ASR systems can then do a better job of turning speech into text. This is helpful in classrooms, offices, and crowded places. Even when several people are talking, the system can focus on the important voice.

Telecommunication:
Voice separation improves call quality, especially in group calls or public areas where background noise is prevalent. It helps people hear each other better on phone calls or video chats. When many people speak at once, or when there is traffic or noise in the background, this model helps clean up the sound. It makes the speaker's voice clearer. This leads to better communication. People don't have to repeat themselves. Everyone can hear more clearly. It is helpful in daily conversations and in professional meetings too.

Hearing Aids and Assistive Devices:
Real-time source separation can amplify target speakers, enhancing auditory experiences for users in multi-speaker settings. This helps people with hearing problems listen better in noisy rooms. The model boosts the voice that the listener wants to hear. It reduces other background sounds. This is very helpful in family gatherings, restaurants, or busy streets. The person using the hearing aid can focus on the main speaker. It makes their life easier and more comfortable. It also helps them stay involved in conversations. They don't feel left out.

Broadcasting and Media Production:
Isolating individual voices allows content creators to enhance audio quality, even when original recordings include multiple overlapping sounds. This means that even if a recording has many voices together, the model can pull them apart. It gives creators more control over the final sound. They can adjust each voice as needed. This makes the video or podcast sound more professional. It also helps when editing interviews, shows, or songs. Clear audio gives a better experience to the audience. It keeps the content clean and easy to follow.

Security and Surveillance:
Speaker separation aids in identifying individual speakers in recordings, supporting security and law enforcement efforts. In crowded or noisy places, it helps find out who said what. This is useful when

analyzing recorded conversations. It can help in solving crimes or tracking suspects. The model helps to understand each person's voice separately. It can give clear evidence from confusing recordings. This supports police, investigators, and security teams. It adds more value to the tools they already use. Clear audio can lead to better decisions and actions.

VI. EXPERIMENTAL SETUP AND RESULTS

6.1 Dataset

The experiments utilize a multi-speaker dataset with both clean and noisy recordings, including overlapping voices to simulate real-world scenarios. Additional testing was conducted using the WHAMR! dataset, featuring reverberant conditions.

6.2 Performance Evaluation Metrics

Performance is evaluated through:

Signal-to-Distortion Ratio (SDR): Measures the clarity of separated signals.

Signal-to-Interference Ratio (SIR): Indicates how well the model reduces interference from overlapping speakers.

Perceptual Evaluation of Speech Quality (PESQ): Assesses perceived speech quality post-separation.

6.3 Results

Our deep attractor model demonstrated significant improvements in SDR, SIR, and PESQ compared to baseline DNN models, particularly in noisy and reverberant conditions. Additionally, the model's speaker-independent capabilities performed well across a range of speaker combinations not seen during training.

VII. CONCLUSION AND FUTURE WORK

This research contributes an innovative deep attractor model for speaker-independent audio source separation. By creating an embedding space with dynamic attractor points, the model adapts effectively to varying numbers of audio sources and complex acoustic environments. Future research may focus on further refining the model for more nuanced audio scenes, integrating multi-channel inputs, or exploring hybrid architectures combining attractor networks with other separation techniques for enhanced performance in real-world applications.

REFERENCES

IEEE References:

[1] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 24, no. 5, pp. 967–977, 2016.

[2] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single channel speech separation via high-resolution deep neural networks," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 24, no. 8, pp. 1424–1437, 2016.

[3] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1003–1012.

[4] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018.

[5] Z. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in 2018 IEEE ICASSP, 2018, pp. 1–5.

[6] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in 2020 IEEE ICASSP, 2020, pp. 696–700.

[7] Daniel Michelsanti , Member, IEEE, Zheng-Hua Tan , Senior Member, IEEE "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation" VOL. 29, 2021

[8] Sam Ansari , Khawla . A. Alnajjae , Tarek Khater1 , Soliman Mahmoud, And Abir Hussain "A Robust Hybrid Neural Network Architecture for Blind Source Separation of Speech Signals Exploiting Deep Learning" Digital Object Identifier 10.1109/ACCESS.2023.3313972.

[9] DeLiang Wang, Fellow, IEEE, and Jitong Chen "Supervised Speech Separation Based on Deep

Learning: An Overview" DOI 10.1109/TASLP.2018.2842159, IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[10] Wang, Zhong-Qiu, Jonathan Le Roux, and John R. Hershey. "Deep clustering with convolutional neural."