Development of a Language-Guided Audio Source Separation System

Yueyan Pang
Department of Electrical Engineering
Columbia University
New York, US
yp2726@columbia.edu

Jiabao Shen
Department of Electrical Engineering
Columbia University
New York, US
js6626@columbia.edu

Yucheng Teng
Department of Electrical Engineering
Columbia University
New York, US
yt2911@columbia.edu

***Abstract*—This paper develops a system for separating audio sources from mixtures using natural language queries. The system combines audio signal processing with modern deep learning methods, including ResUNet and Transformer models, and uses contrastive pre-training (CLAP) to link language and audio features. The training and testing process uses a two-step transfer learning approach to fine-tune it for specific tasks. The results show strong performance for separating instruments like violins and trumpets after fine-tuning. While the system works well for some tasks, challenges remain, such as improving performance for underrepresented sounds and handling diverse language queries. This model can be applied in music production, sound editing, and assistive listening.**

***Index Terms*—Language-Guided Audio Separation, Deep Learning, Natural Language Queries**

I. Introduction

A. Background

The Short-Time Fourier Transform (STFT) converts audio signals into spectrograms, providing a time-frequency representation that is well-suited for deep learning. Spectrograms enable models to capture detailed audio patterns, making them particularly effective for separating overlapping audio sources, such as two-instrument duets.

Audio-Sep, a pre-trained audio separation model trained on large-scale datasets such as AudioSet, WavCaps, and VGGSound, leverages extensive learned features. This pre-training significantly reduces training time while enhancing performance on our specific task. This project focuses on fine-tuning Audio-Sep to achieve high-quality separation of two musical instruments.

We explore state-of-the-art deep learning architectures, including ResUNet and Transformer blocks, alongside signal processing methods like STFT and Mel-spectrograms. To address challenges such as language ambiguity and overlapping audio sources, we employ contrastive language-audio pre-training models (CLAP) and multi-modal learning strategies, tailoring the pre-trained model to meet the specific demands of our task.

B. Related Works

The task of language-guided audio source separation has gained significant attention in recent years, with researchers exploring various methods to bridge the gap between natural language understanding and audio signal processing. This section highlights key contributions in this domain.

Liu et al. introduce the task of Language-Queried Audio Source Separation (LASS) [1]. This study proposes LASS-Net, an end-to-end neural network designed to jointly process audio and language inputs. The network utilizes textual descriptions to extract target sources from audio mixtures, demonstrating its effectiveness in aligning language queries with specific audio components.

Building on the integration of multiple modalities, Tan et al. [2] explore a self-supervised approach to leverage audio, visual, and textual inputs. The authors propose a trimodal framework that enforces consistency across these modalities, enabling the system to accurately separate audio sources based on natural language queries and visual context. This work underscores the importance of multimodal alignment for robust source separation.

Expanding on open-domain source separation, Liu et al. further introduce AudioSep [3], a foundational model trained on a large-scale multimodal dataset. AudioSep supports zero-shot generalization for diverse tasks, including instrument separation, audio event isolation, and speech enhancement. This work highlights the potential of leveraging large-scale data and pre-trained models to achieve high-performance, language-guided separation.

II. Technical Approach

Before training the model, we examined the separation performance of a pre-trained model by directly loading and classifying by input text (instrument name). The model's performance was inadequate, particularly for the trumpet. To address these limitations, sliced and resampled data were utilized for transfer learning using the pre-trained model. This approach was designed to enhance the model's sensitivity to the specific characteristics of trumpet sounds, thereby improving separation performance for trumpet extraction.

A. Pretrained Model Download and Integration

To initiate the development process, two pretrained model weights were downloaded and saved into a specified directory. These models included a general audio separation model and a specialized model for music and speech separation. These pretrained weights formed the foundation for further model fine-tuning and customization.

B. Dataset Preparation

We chose three three-minute music segments, including the mixed audio and separate tracks for the violin and trumpet. The audio data was prepared to meet the input requirements of the pretrained model, which specified a slice duration of 5 seconds and a sampling rate of 32,000 Hz. To facilitate model training and fine-tuning, we constructed a dataset by dynamically linking mixed audio files with their respective target audio files and textual descriptions. This process included constructing paths for both mixed and target audio files based on a predefined directory structure, associating each target audio file with a corresponding textual description, and validating file existence to ensure data consistency. The resulting dataset comprised a collection of data samples ready for training, each containing a mixed audio file, a target audio file, and a textual description.

C. Transfer Learning and Fine-tuning

To adapt the pre-trained model for the specific task of instrument separation, a two-phase transfer learning approach was employed.

In the initial phase, the QueryNet module, responsible for language processing, was frozen to retain its pre-learned capabilities, ensuring that the language representation remained intact. During this stage, the audio separation model was fine-tuned on the target dataset to improve separation performance. The fine-tuning process involved a loss function based on the $L1$ norm, which measured the difference between the predicted and target audio signals. The optimization was carried out using the AdamW optimizer with a carefully tuned learning rate, coupled with a warm-up-based learning rate scheduler to stabilize training dynamics. The model was trained over 75 epochs, during which the loss was progressively reduced, ultimately reaching a final value of 0.0160, indicating effective convergence and improved separation performance.

As part of the adaptation process, the structure of the output layer was modified to incorporate depthwise separable convolutions. This adjustment aimed to enhance computational efficiency and improve the model's capacity to learn distinct audio features. Specifically, the output layer was replaced with a structure consisting of a 1×1 convolutional layer with 512 input and output channels, followed by another 1×1 convolutional layer to produce the final single-channel output. This modification not only reduced the number of parameters but also streamlined the separation process.

In the second phase, QueryNet was unfrozen to allow joint optimization of both the language and audio processing components. This step aimed to further refine the model's ability to process diverse language queries while maintaining effective audio separation. Importantly, despite unfreezing during this phase, the original QueryNet architecture remained unmodified throughout the process to preserve its functionality and integrity. Overall, QueryNet was utilized as a tool to facilitate the task, while the primary focus of this work was centered on achieving high-quality separation of two musical instruments.

This two-phase approach, combined with the modifications to the output layer, ensured both the preservation of pre-trained knowledge and the adaptation of the model to the specific demands of the target task.

D. Lightning Integration and Model Validation

The fine-tuned model was integrated into a PyTorch Lightning framework to streamline training and evaluation. This framework enabled modular, scalable, and efficient training pipelines. The integration validated that the model weights were correctly loaded and that it could perform high-fidelity audio separation tasks.

E. Test and Evaluation

To evaluate the effectiveness of the model, the fine-tuned version was tested using the same sample file and query as in the initial examination. The output file demonstrated, from a subjective auditory perspective, an improved ability to accurately separate the target sounds. In addition to this subjective assessment, a comprehensive evaluation was conducted on the separate tracks of the two instruments. This included a Separation Performance Evaluation and Spectrum Consistency Evaluation. These evaluations provided both qualitative and quantitative insight into the enhanced separation capabilities of the model after fine-tuning.

III. Experiments

A. Audio Separation Performance Evaluation

The evaluation of the performance of the system was conducted through two primary metrics: Signal-to-Distortion Ratio (SDR) and SDR improvement (SDRi), as shown in Table I. The results highlight the effectiveness of the AudioSep model in separating violin and trumpet sounds under different conditions, specifically comparing the performance of the pre-trained model with the transfer learning(fine-tuning) approach.

For violin separation, the Pre-trained Model achieved an SDR of 17.57 and an SDRi of 17.10, demonstrating strong performance in isolating violin sounds with minimal distortion. However, after applying fine-tuning, the SDR and SDRi slightly decreased to 16.87 and 15.88, respectively. The trumpet separation results highlight the advantages of transfer learning. The Pre-trained Model performed poorly, with an SDR of -9.91 and an SDRi of -5.59, indicating significant distortion and an inability to accurately separate trumpet sounds. After fine-tuning, the SDR improved to 10.09, and the SDRi rose to 9.01, reflecting a substantial enhancement in trumpet separation.

The high baseline performance on the violin indicates that the Pre-trained Model has already learned robust violin-specific features during pre-training, likely because violin data is abundant in the pre-training dataset. However, transfer learning introduces a new output layer and retrains the model, potentially disrupting the well-learned features, and leading to a slight performance decline. This suggests that violin separation might not benefit much from further tuning when using small, domain-specific datasets. In contrast, the Pre-trained Model struggles with trumpet separation, likely due to the insufficient representation of trumpet sounds in the pre-training dataset. Transfer learning helps the model adapt to trumpet-specific features by learning from the new dataset, which results in significant performance improvement. This demonstrates that transfer learning is particularly beneficial for underrepresented categories in the pre-training phase.

TABLE I: Comparison of SDR and SDRi for Violin and Trumpet Separation

Instrument	Model	SDR	SDRi
Violin	Pre-trained Model	17.57	17.10
	Fine-Tuned Model	16.87	15.88
Trumpet	Pre-trained Model	-9.91	-5.59
	Fine-Tuned Model	10.09	9.01

B. Spectrum Consistency Evaluation

The spectrum consistency evaluation (Table II) highlights similar patterns by assessing the correlation and mean squared error (MSE) between the target and separated signals. For violin, the Pre-trained Model achieved a high correlation of 0.9933 with a very low MSE of 0.0006, reinforcing the observation of its strong performance. After transfer learning, the correlation slightly dropped to 0.9898, with a minor increase in MSE to 0.0009. For the trumpet, the Pre-trained Model showed a correlation of 0.2872 and an MSE of 0.0272, reflecting poor separation quality. Transfer learning improved the correlation to 0.9575 and reduced the MSE to 0.0006, further confirming the effectiveness of the training process in addressing the deficiencies of the pre-trained model.

In addition to the learning bias inherent in the pre-trained model, several factors may explain the insufficient performance of trumpet separation. The relatively small amplitude and limited dynamic range of trumpet signals result in reduced prominence within the mixture, making it challenging for the model to effectively extract and separate the spectral information associated with trumpet sounds during feature extraction. Furthermore, the spectral characteristics of trumpet signals are primarily concentrated in the mid-to-low frequency range, while the model may inherently exhibit greater sensitivity to instruments with rich high-frequency components, such as violins. This discrepancy can reduce the model's ability to capture and utilize trumpet-specific features.

TABLE II: Comparison of SDR and SDRi for Violin and Trumpet Separation

Instrument	Model	Correlation	MSE
Violin	Pre-trained Model	0.9933	0.0006
	Fine-Tuned Model	0.9898	0.0009
Trumpet	Pre-trained Model	0.2872	0.0272
	Fine-Tuned Model	0.9575	0.0006

C. Visualization of separation results

We visualized the spectrograms of audio mixtures, target audio sources, and separated sources using the Fine-tuning Model, as violin shown in Fig. 1 and trumpet shown in Fig. 2 We observed that the spectrogram pattern of the separated sources closely matches that of the target sources, match our expected experimental results.

IV. Discussion

A. Accomplishments

For trumpet separation, the transfer learning approach significantly improved the separation performance for the trumpet, as the SDR improves from -9.91 to 10.09. The use of fine-tuning pre-trained models on domain-specific datasets is important in capturing under-represented features in the pre-training phase.

For violin separation, the pre-trained model achieved high baseline performance in minimal distortion, which may due to the sufficient representation of violin data in the pre-training dataset. This may highlight the robustness of pre-trained features for well-represented categories.

The visualization of the spectrograms of the separated, target and mix sources shows that the model generally has a effective performance in reconstructing the temporal and spectral characteristics in the separating task.

B. Limitations

The slight decrease in SDR for violin separation after transfer learning indicates that fine-tuning on small, domain-specific datasets can disrupt well-learned features, leading to diminished performance. Further refinement of the training strategy is necessary to preserve pre-trained knowledge for well-represented categories.

Despite significant improvements, the model initially struggled with trumpet separation due to the instrument's limited representation in the pre-training dataset and its spectral characteristics. Since trumpet signals are always concentrated in the mid-to-low frequency range, it is more difficult for the model to extract relevant features for trumpets.

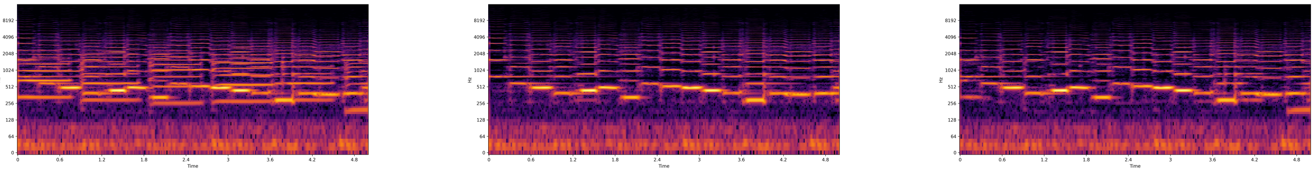
The reliance on pre-trained models may introduce biases based on the distribution of the pre-training data [4] [5]. Instruments with scarce representation like trumpets require more targeted interventions to achieve comparable performance.

C. Working Conditions and Domain of Applicability

The methods in this project are most effective when the target audio source is well-represented in the pre-training dataset and exhibits distinctive spectral characteristics. Instruments with overlapping frequency ranges are more difficult to separate, particularly without additional data augmentation or advanced feature engineering.

The results of this project show that language-guided audio separation systems are well-suited for applications involving:

- Music Production: pre-trained models can leverage abundant instrument-specific data.
- Environmental Sound Separation: diverse sound categories are typically well-represented.

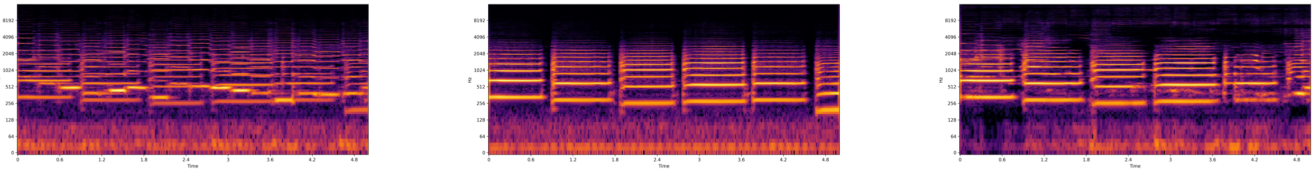


(a) Mixed Audio Spectrogram

(b) Target Violin Spectrogram

(c) Model-separated Violin Spectrogram

Fig. 1: Visualization of violin separation results



(a) Mixed Audio Spectrogram

(b) Target Trumpet Spectrogram

(c) Model-separated Trumpet Spectrogram

Fig. 2: Visualization of trumpet separation results

D. Future Improvements

To address the limitations observed, several improvements are proposed:

- Enhanced Pre-training: Incorporate more diverse and balanced datasets during the pre-training to reduce biases and improve generalization across under-represented categories.
- Mixed Training Strategies: Develop training pipelines that balance fine-tuning with feature preservation. For example, selectively freezing certain layers of the pre-trained model while fine-tuning others may help retain robust pre-trained features.
- Domain-Specific Data Augmentation: Use data augmentation techniques to artificially increase the representation of under-represented categories in the training data.

References

[1] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, “Separate what you describe: Language-queried audio source separation,” in *Interspeech*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247778595

[2] R. Tan, A. Ray, A. Burns, B. A. Plummer, J. Salamon, O. Nieto, B. Russell, and K. Saenko, “Language-guided audio-visual source separation via trimodal consistency,” 2023. [Online]. Available: https://arxiv.org/abs/2303.16342

[3] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, “Separate anything you describe,” 2024. [Online]. Available: https://arxiv.org/abs/2308.05037

[4] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, “Vggsound: A large-scale audio-visual dataset,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 721–725.

[5] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.