Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/30/20230085

An optimized approach to speech transcription using blind source separation and speech-to-text machine learning models

Chaoyang Yin

Grainger College of Engineering, University of Illinois at Urbana-Champaign, Champaign, Illinois, 61820, United States

cyin9@illinois.edu

Abstract. The use of speech-to-text transcription has a multitude of applications in various industries, including accessibility support, language processing, and automatic subtitling. In recent years, there has been greater interest in incorporating automatic speech source separation features to improve the accuracy and efficiency of transcription mechanisms. This paper aims to design a transcription mechanism that utilizes DUET algorithm to separate speech sources in a stereo setup. The separated sources are then transcribed into text using a machine learning model. The study evaluates the effectiveness of this approach using a dataset of speech recordings. The results of the study indicate high accuracy in speech separation and transcription, highlighting the potential of this approach for practical applications. However, the study also revealed potential issues with the mechanism, indicating the need for further exploration and refinement. These findings indicate the potential of the proposed approach for practical applications, and propose insight for further development and researches in this area.

Keywords: audio processing, blind source separation, speech recognition.

1. Introduction

The field of speech processing continues to be a topic of considerable interest in both academia and industry, with applications ranging from transcription and translation to speech enhancement and separation [1]. One of the challenges in speech processing is separating speech signals from noisy or overlapping sources, which can propose challenge to the next steps of voice transcription. This issue has become particularly important as more people work remotely and rely on online communication tools, which often struggle to capture clear speech signals in noisy home environments.

The importance and necessity of this study lie in its potential to improve the quality and intelligibility of speech signals in a range of applications, from online communication to automatic transcription and translation. By developing an effective and efficient solution for speech separation, this study could help improve the accuracy and reliability of speech processing tools and services. Additionally, the study contributes to the ongoing research efforts in speech processing and deep learning, demonstrating the potential of these technologies to address important challenges in the field.

Numerous academic approaches have been proposed to address the problem of speech-to-text transcription of multiple sources, including deep learning-based methods and topic modeling [2,3]. However, the objective of this current study is to explore a novel approach for the source separation problem in speech processing, and combine it with a machine learning (ML) model to construct a

DOI: 10.54254/2755-2721/30/20230085

pipelined workflow that allows for real-time speech-to-text transcription. This proposed approach aims to overcome the limitations of existing methods and enhance the performance of speech-to-text transcription by leveraging the benefits of both source separation and machine learning techniques.

2. Methods

The methodology employed by this study to conduct the blind source separation process is centered around the Degenerate Unmixing Estimation Technique (DUET), which is an example of unsupervised speech separation, aiming to separate speech signals from a mixture of them without any prior knowledge about the sources and the mixture [4]. This method was chosen due to its capability to successfully filter out noise and separate sources even in scenarios where the number of speakers is uncertain and there are more than two speakers present in the recording, assuming a stereo microphone array consisting of two microphones. This specialized approach is especially useful for human speech signal separation. DUET is particularly advantageous in handling degenerate situations, where the number of sources exceeds the number of mixtures [4]. By effectively filtering out unwanted sources, DUET's use in this study promises to facilitate a cleaner and more accurate separation of the desired signals.

The foundation of the DUET algorithm rests on the premise that each source of speech is located within a distinct Time-Frequency bin in the windowed Fast Fourier Transform. This essential assumption serves as the core mechanism behind DUET's capacity for successful source separation. Through the process of clustering Time-Frequency components according to each source, DUET can extract individual signals from a mixture of sources. The existence of unique amplitude attenuation and phase differences in the speech signals of each speaker is attributed to the microphone array and speaker geometry. These critical parameters, therefore, are utilized as key evaluative metrics in the clustering of Time-Frequency bins. By identifying and isolating the unique characteristics of each speaker's signal, DUET promises to deliver an accurate and efficient separation of multiple sources of speech in a variety of complex recording scenarios.

The actual implementation of this study involved performing Short Time Fourier Transform (STFT) on two signals corresponding to the received signal of each microphone. The Time-Frequency components were then scrutinized to detect clusters based on phase difference and amplitude attenuation [5]. Following the identification of clusters, the Time-Frequency bins corresponding to each cluster were retrieved by masking the original signal in the Fourier domain. To convert these isolated speech signals back into the time domain, the Inverse STFT was employed. Due to the limitations of computing time and resources, the implementation of the algorithm works to separate the stereo input as a .wav file and saves the separated speeches in multiple .wav files as output. It is important to note that the DUET algorithm's performance is highly dependent on the quality of the input signal and the clustering parameters chosen [6]. Moreover, optimal performance can be achieved by selecting appropriate threshold values to cluster the Time-Frequency bins.

In addition, to save computational time, the study reduced the metrics in bin separation to singularly cluster along phase difference. This approach provides a streamlined process for the algorithm to identify speaker signals by evaluating their unique phase and amplitude characteristics. Ultimately, the success of the implementation depended on the accuracy and efficiency of the DUET algorithm in identifying and separating distinct sources of speech from a complex mixture of signals in a recording scenario.

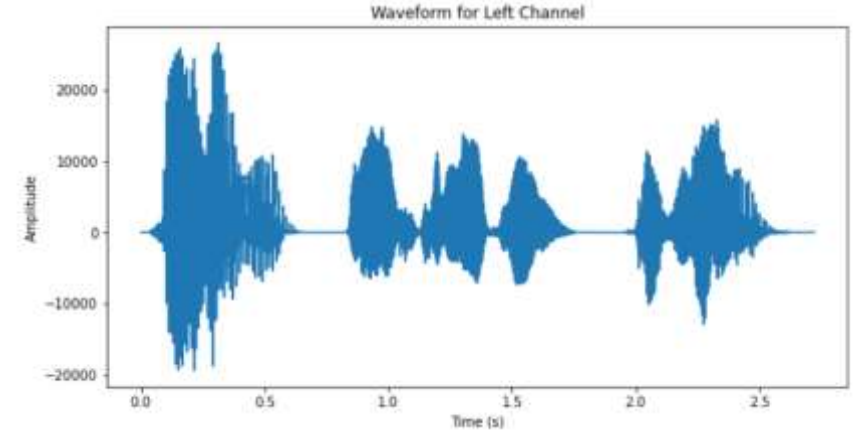
To achieve speech-to-text recognition, the present study adopts a machine learning approach and develops a transcriber with the Wav2Vec2 model, which is developed by Facebook AI Research. The Wav2Vec2 model is a transformer-based model pre-trained on a massive amount of unlabeled audio data, which includes a convolutional neural network (CNN) and a transformer architecture, both of which have exhibited impressive performance in natural language processing (NLP) tasks [7]. The Wav2Vec2 model can be fine-tuned on a smaller labeled dataset, optimizing its representations for specific speech recognition tasks, such as transcribing a particular language or domain, leading to

DOI: 10.54254/2755-2721/30/20230085

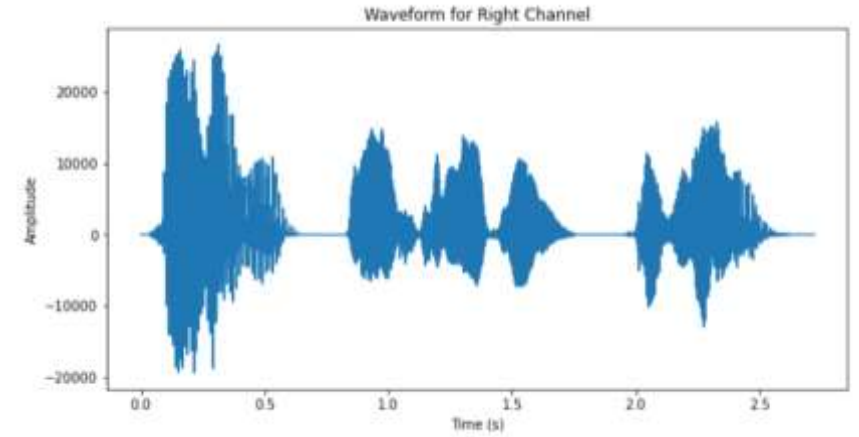
improved accuracy. This study implements the model using the transformers package in Python and utilizes the pre-trained model "facebook/wav2vec2-base-960h."

3. Results

To demonstrate the performance of the present implementation speech source separation, two datasets are tested. One simpler case is the mixtures from the recording of three time-separated, computer-generated speakers with little noise, and the other involves a complex setting of five human speakers speaking simultaneously under a noisy environment. Both stereo recordings are under the premise of being recorded in a two-microphone setup. Figure 1 shows the waveform of the mixtures, Figure 2 shows the waveform of outputs from the simple setup. The waveform of the output of the separation is shown in Figure 3, and the waveform of the mixtures of the complex dataset is shown in Figure 4.



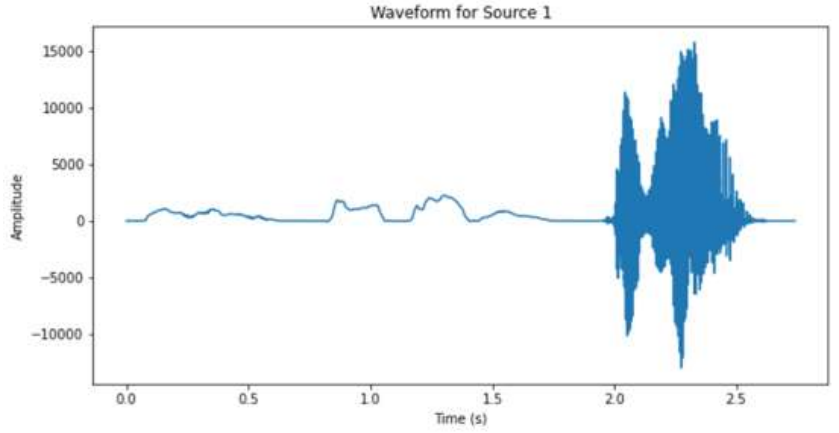
(a)



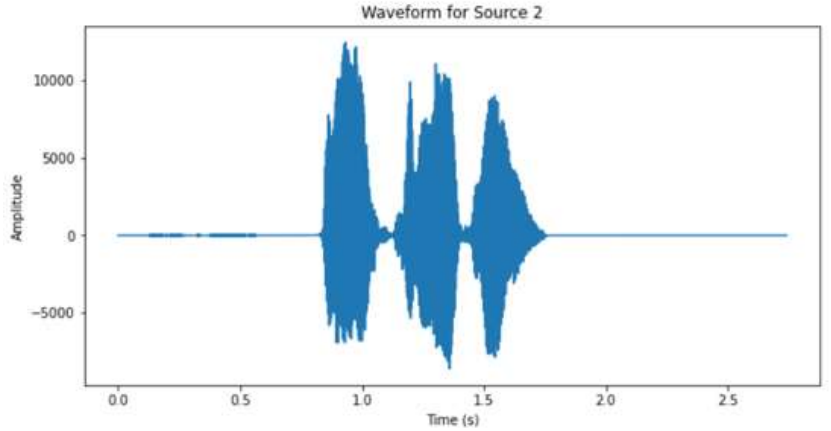
(b)

Figure 1. Waveform of the two mixtures (simple setup).

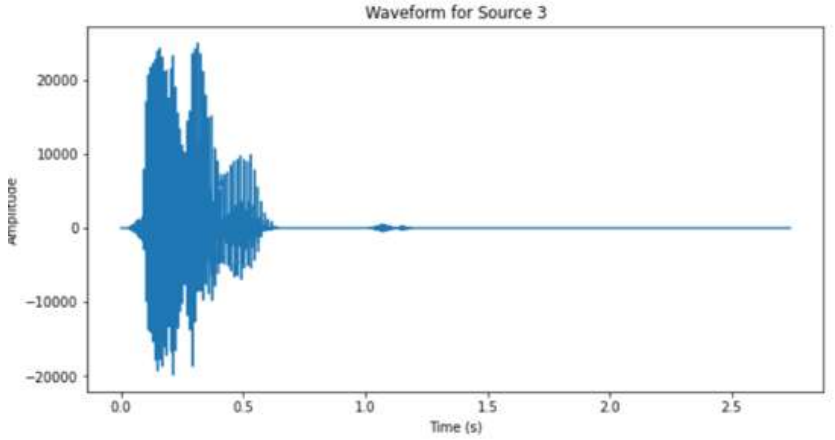
DOI: 10.54254/2755-2721/30/20230085



(a)



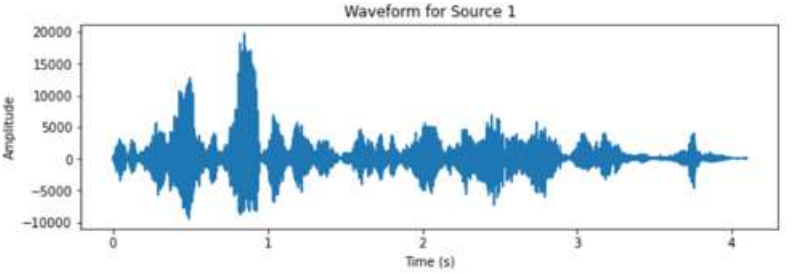
(b)



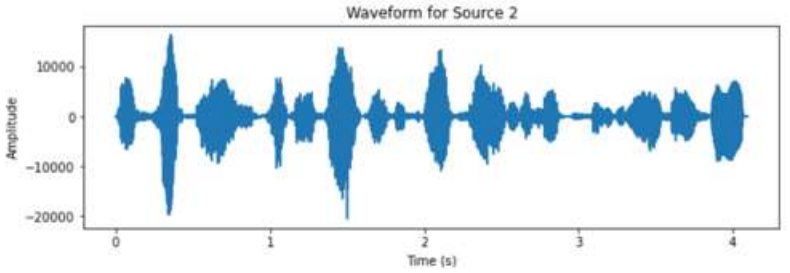
(c)

Figure 2. Waveform of separated sources, source number = 3 (simple setup).

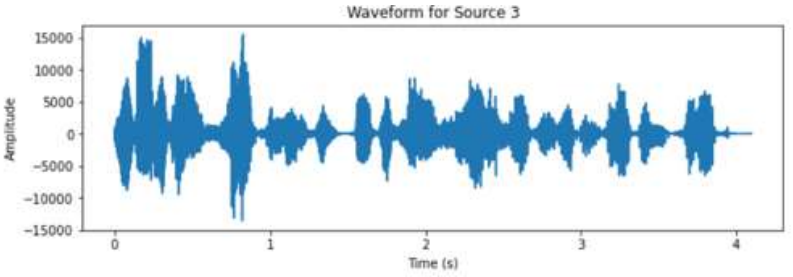
DOI: 10.54254/2755-2721/30/20230085



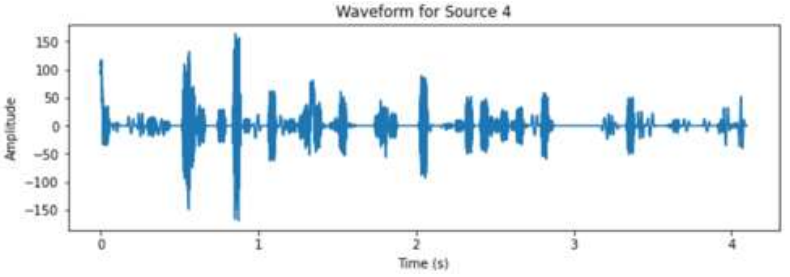
(a)



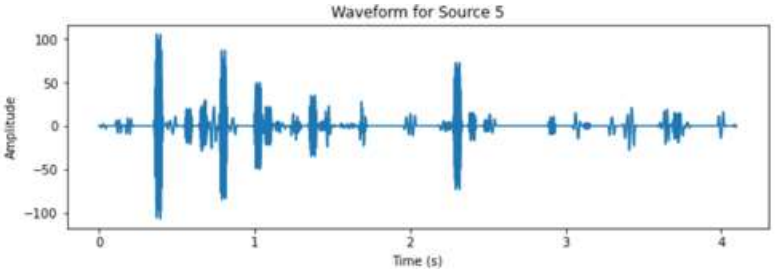
(b)



(c)



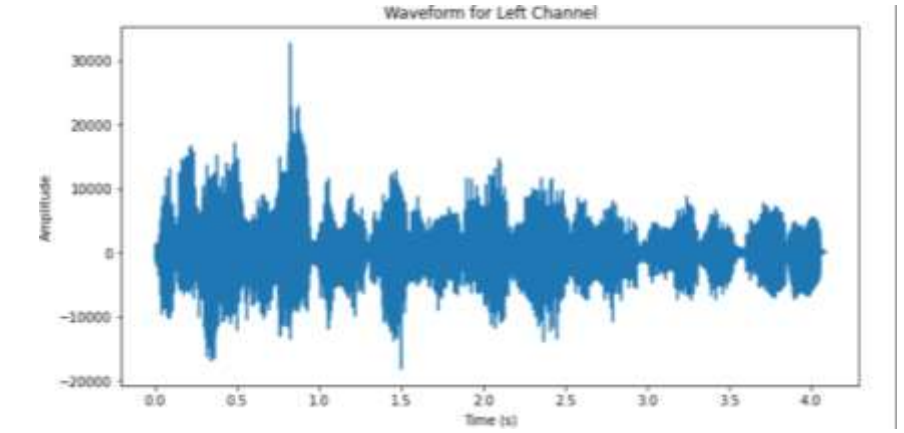
(d)



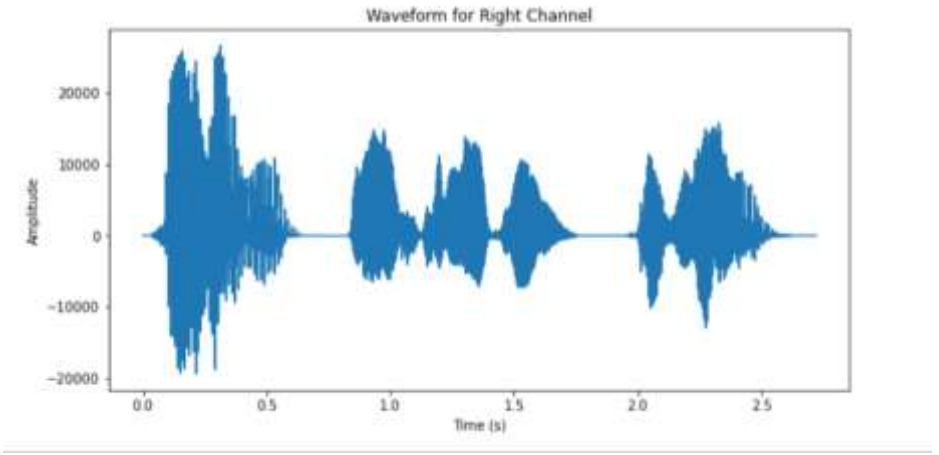
(e)

Figure 3. Waveform of separated sources, source number = 5 (complex setup).

DOI: 10.54254/2755-2721/30/20230085



(a)



(b)

Figure 4. Waveform of the two mixtures (complex setup).

From the waveform results of the two datasets, it can be observed that the DUET algorithm has significantly better performance on the dataset corresponding to the simpler recording setup. All three speakers' speech signal are clearly separated from the two mixtures, as the waveform of the outputs are separated in time slots of each speaker. The results from the complex setup, however, does not exhibit satisfactory output. For two of the output signals have very low amplitude relative to the other sources and similar outline of waveform. The audio playback of all output also revealed impaired quality across all five recovered signals.

The performance of our algorithm is tested under a variety of experimental conditions. One of the challenges is the computational time required for processing large signals with multiple sources. On a Jupyter notebook running on a 10th generation Intel Core i7 laptop, the time required to separate sources from signals of varying lengths and numbers of sources is measured.

Our results indicate that it takes approximately 25 seconds to separate three sources from a signal with 60,000 samples, and approximately 1 minute and 45 seconds to separate five sources. These computational requirements may limit the algorithm's applicability in real-time systems or applications with strict latency requirements.

Based on the limitations of the testing dataset, the performance of the machine learning model employed in this study cannot be directly measured. However, the model demonstrated high recognition accuracy for the simpler dataset used in the study, indicating potential for further investigation on more complex datasets. While detailed investigation of the accuracy of the wav2vec2 model is beyond the

DOI: 10.54254/2755-2721/30/20230085

scope of the current article, other studies have found that Wav2Vec2 achieves remarkable performance without requiring large amounts of labelled data for training, outperforming other speech recognition models on several standard benchmarks, while also requiring significantly less labelled data for training.

4. Discussion

The use of the DUET algorithm for speech separation tasks has shown promising results in conditions where the number of sources is low and the noise level is minimal. However, the algorithm's performance tends to degrade when applied to more complex settings with a high number of sources. One potential reason for this limitation is due to the W-disjoint Orthogonality of human speech. As the number of sources increases, the likelihood of sources colliding in Time-Frequency bins increases as well. To improve the DUET algorithm's performance, optimization of the parameters used in the algorithm is necessary [6]. The optimization to improve the algorithm's accuracy can include adopting a duo-metric standard for clustering with both amplitude and phase differences considered, selecting an appropriate window size, and tuning the parameters of the K-means clustering algorithm. Also, to save the existing issue of computational complexity, improvements can be achieved by tuning the window size of the STFT, tuning the parameters of the K-means, and switching to a more cost-efficient clustering algorithm.

Alternative blind source separation methods, such as FastICA, have shown promising results in separating signals in complex mixtures. FastICA utilizes a non-linear method based on maximizing the non-Gaussianity of the observed signals [8]. One advantage of FastICA is its ability to handle a larger number of sources, making it a potentially better solution for speech separation tasks in complex settings. Also, Non-negative Matrix Factorization (NMF) has emerged as a powerful tool for blind source separation that has shown promising results in separating speech signals. NMF has demonstrated improved performance in terms of separation quality, computational efficiency, and robustness to noise compared to traditional methods such as independent component analysis and principal component analysis [9]. Additionally, NMF has shown excellent performance in separating speech signals from different speakers and sources, making it a popular tool for speech separation tasks [10].

In terms of the machine learning-based transcription task, the Wav2Vec2 model has exhibited remarkable performance in transcribing audio data with a high degree of accuracy. However, its computational demands for processing large audio files pose a significant challenge. To mitigate this issue, the model can be deployed on GPU acceleration, and smaller batch sizes can be used. Optimization of the model's parameters is also necessary to enable real-time transcription. Furthermore, other machine learning-based transcription methods, such as those based on convolutional neural networks and other transformer-based architectures, can be identified and compared to the Wav2Vec2 model to determine the most effective solution for the specific task at hand. Such evaluations are essential to identify the most efficient and effective solutions for speech recognition tasks under various conditions.

This study's current implementation requires the number of sources as input, limiting its applicability to scenarios where the number of sources is known beforehand. In practice, it is often difficult to know the exact number of speakers involved in a mixture, making the solution limited in its application. To address this limitation, incorporating feedback from the machine learning model and performing trials to estimate the number of speakers involved in the mixture can be explored. After transcribing the results from multiple assumptions of the number of sources, an estimation of the likelihood of the true number of sources can be performed based on the transcription's likelihood to be human speech. This approach can be particularly useful in scenarios where the number of speakers involved in the mixture is unknown. However, such a feature's implementation would bring an increase in computational complexity and require additional models to analyze the similarity to human speech.

In addition, the application of machine learning to speech recognition tasks can be improved further by developing models that can better handle background noise, accents, and dialects. The introduction of adaptive models that can learn from and adapt to real-time feedback can further enhance the accuracy and robustness of the system. Exploring techniques for improving the model's interpretability can

DOI: 10.54254/2755-2721/30/20230085

provide insights into how the model makes decisions. This can be particularly useful in fields such as healthcare, where the interpretation of transcription results plays a crucial role in patient care.

5. Conclusion

The findings of this study demonstrate that the DUET algorithm and Wav2Vec2 model show potential for speech separation and transcription tasks, respectively. The algorithm performed well in simpler recording setups, while the machine learning model exhibited a high degree of accuracy in transcribing audio data. These results imply that employing both methods in a multi-source speech separation-transcription workflow can be beneficial for practical applications, such as in speech recognition systems. However, challenges such as reduced performance in complex settings, the need for the number of sources' prior knowledge, and computational complexity must be addressed. The results presented here suggest that further optimization of the algorithms, specifically parameter tuning, can improve the algorithms' efficacy and reduce computation time.

The impact of this study is in demonstrating the potential of using blind source separation and machine learning models in speech separation and transcription tasks. It can serve as a starting point for future research evaluating other blind source separation methods and exploring other machine learning-based transcription methods.

Future studies can address more works on this topic to tune and improve implementations of DUET that yields more accuracy and efficiency, also explore evaluating different blind source separation methods, such as FastICA and Non-negative Matrix Factorization, to identify the most effective solution for specific complex settings. Additionally, developing models that can handle different accents and dialects, and optimizing the interpretability of the models to gain insights into how they make decisions, can contribute to more effective and robust models. These efforts can ultimately lead to the implementation of more accurate and efficient speech separation and transcription technologies with broader applications across various fields.

References

[1] Li Y and Wang D 2021 Noise-Robust Text-Dependent Speaker Verification based on Deep Neural Networks IEEE Access 9 p 1-11

[2] Chen J and Wang D 2018 Speaker-Independent Speech Recognition System for Multi-Speaker Scenarios using Convolutional Neural Networks IEEE Access 6 p 78317-78324

[3] Zhu S, Li H, Chen D and Li Y 2018 Multi-Channel Meeting Recognition with Source Separation and Topic Modeling IEEE Signal Processing Lett. 25(5) p 695-698

[4] Rickard S, Yilmaz O and Herault M 2000 Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures Pro. IEEE Int. Conf. Acoustics Speech Signal Processing 6 p 3466-3469

[5] Han Y and Wang D 2020 Singing Voice Separation from Music Accompaniment for Monaural Recordings IEEE/ACM Trans. Audio Speech Language Processing 28 p 2178-2193

[6] Wang Z, Liao W and Liu M 2020 DuET-Net: A dual-branch network for speech separation with an application to noise-robust ASR IEEE/ACM Trans. Audio Speech Language Processing 28 p 2605-2618

[7] Zhang Y and Nguyen T 2020 Speech recognition using self-supervised learning and wav2vec2 Pro. IEEE Int. Conf. Acoustics Speech Signal Processing p 6904-6908

[8] Hyvärinen A and Oja E 1999 Fast and robust fixed-point algorithms for independent component analysis IEEE Trans. Neural Networks 10(3) p 626-634 doi:10.1109/72.761722

[9] Le Roux J, Hershey J R, and Wand M, 2014 Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups IEEE Signal Processing Magazine 31(3) p 82-97

[10] Kameoka H, Miyoshi M, Saruwatari H and Nakatani T 2005 Blind extraction of audiosources using non-negative matrix factorization with time-frequency masking IEEE Int. Conf. Acoustics Speech Signal Processing p 393-396