JQCM
Journal Of Al-Qadisiyah For Computer Science And Mathematics
University Of AL-Qadisiyah

Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Voice separation and recognition using machine learning and deep learning a review paper

Zaineb h. ibrahemma [a*], ***Ammar I. Shihab*** [b]

[a] Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq.Email:zainebas553@gmail.com*

[b]Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq.Email: ammarste@gmail.com

ARTICLE INFO

Article history:
Received: 25 /05/2023
Rrevised form: 05 /07/2023
Accepted : 09 /07/2023
Available online: 30 /09/2023

Keywords:

Voice Isolation,

Deep Neural Networks,

Speech Recognition,

Speaker Identification,

Frequency Domain,

Time Domain

ABSTRACT

Voice isolation, a prominent research area in the field of speech processing, has garnered a great deal of attention due to its prospective implications in numerous domains. Deep neural networks (DNNs) have emerged as a potent instrument for addressing the challenges associated with vocal isolation. This paper presents a comprehensive study on the use of DNNs for voice isolation, focusing on speech recognition and speaker identification tasks. The proposed method uses frequency domain and time domain techniques to improve the separation of target utterances from background noise. The experimental results demonstrate the efficacy of the proposed method, revealing substantial improvements in voice isolation precision and robustness. This study's findings contribute to the increasing corpus of research on voice isolation techniques and provide valuable insights into the application of DNNs to improve speech processing tasks.

MSC..

https://doi.org/10.29304/jqcm.2023.15.3.1262

1-Introduction

Separation and recognition of speech are fundamental tasks in speech processing, with applications including automatic speech recognition, speaker identification, and hearing aids. Due to advancements

∗Corresponding author

Email addresses:

Communicated by ‘sub etitor’

in machine learning algorithms and the availability of large speech datasets, the efficacy of speech separation and recognition systems has increased substantially in recent years. In this review paper, we provide an exhaustive overview of the most recent techniques and approaches for speech separation and recognition using machine learning. Distinction of Speech The practice of extracting individual speech sources from a mixture of speech signals is known as speech separation. The issue at hand is a complex one, particularly in environments with high levels of noise, where speech signals are subject to corruption from various sources such as background noise, reverberation, and interference from multiple speakers. Over the last few years, there has been a notable advancement in speech separation through the utilization of deep learning-based techniques. This progress can be attributed to the emergence of sophisticated deep neural network (DNN) structures, including the Convolution Neural Network (CNN) and the Recurrent Neural Network (RNN), the topic under consideration is the Transformer. Hershey et al. (2016) [1] proposed the Deep Clustering method, which has gained significant popularity as a deep learning-based technique for speech separation. The methodology employed in this study involves the utilization of a neural network for the purpose of acquiring knowledge on the correlation between mixed speech and a high-dimensional embedding space. The resulting embedding are then clustered together based on their corresponding speech sources. Subsequently, the clustered embedding is employed to approximate the distinct speech sources. According to this method, a neural network is used to train a mapping from mixed speech to a high-dimensional embedding space, where the embedding belonging to the same voice source are grouped together. Once the individual voice sources have been estimated, the clustered embedding are employed. A other well-liked strategy is the Permutation Invariant Training (PIT) method developed by Kolbaek et al. (2017) [2], The proposed approach involves the utilization of a neural network for the purpose of generating a comprehensive set of permutations for each speech source, followed by the selection of the permutation that exhibits the least reconstruction error. Apart from deep learning-based techniques, alternative methods for speech separation include the Non-Negative Matrix Factorization (NMF) and Independent Component Analysis (ICA) approaches. The aforementioned techniques are predicated on the principles of signal processing and have been extensively employed in the domain of speech separation over an extended period.in the section of technology of speech recognition the process of transforming spoken language into textual or other symbolic forms is commonly referred to as speech recognition. This task presents a challenge owing to the diverse range of speech patterns and accents, as well as the potential interference of ambient noise. The prevalence of deep learning-based techniques has increased significantly in the field of speech recognition due to recent advancements in machine learning algorithms. These approaches have demonstrated superior performance on a range of speech recognition benchmarks. The Connectionist Temporal Classification (CTC) method, proposed by Graves et al. (2006) [3], is widely recognized as a prominent deep learning-based approach for speech recognition. The proposed methodology employs a neural network to acquire knowledge of a direct mapping from speech's acoustic features to text transcriptions, obviating the requirement for explicit alignment between the two. The Listen, Attend and Spell (LAS) technique, introduced by Chan et al. (2016) [4], is a commonly employed methodology that employs an attention mechanism to concentrate on distinct segments of the input speech signal while decoding. Apart from deep learning-based methodologies, alternative techniques for speech recognition include Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). The aforementioned techniques have been extensively employed in the field of speech recognition for numerous years and continue to be utilized in certain applications. While in the field of utilization of machine learning for the purpose of speech separation and recognition is a topic of interest in the academic community. The segregation and identification of speech signals are frequently treated as distinct undertakings, notwithstanding their interdependence, and can be simultaneously enhanced through the application of machine learning techniques. A prevalent approach to incorporating speech separation and recognition involves utilizing the separation system as a preliminary procedure for the recognition system. The elimination of interference from other speakers or background noise has the potential to enhance recognition accuracy. An alternative methodology involves the utilization of a joint model that is capable of executing speech separation and recognition concurrently. The aforementioned objective can be attained through the implementation of multi-task learning techniques or through the amalgamation of distinct models for speech separation and recognition, thereby producing combined outputs. The End-to-End (E2E) has been a topic of recent interest. The performance of speech separation systems has been notably enhanced by the latest developments in deep learning-based techniques. Huang et al. (2021) [5] present a thorough examination of contemporary methodologies and strategies for speech separation utilizing deep learning in their review article. Furthermore, Wang and colleagues (2018) [6]introduced an end-to-end approach for speech separation, which enables the direct mapping of mixed speech to individual speech sources, without the need for any intermediate processing stages. As Li et al. (2021) [7] describe in a survey paper, joint speech separation and recognition is another expanding area of research. More study is required to further enhance the performance of these systems because the field of using deep learning for voice separation and recognition is one that is continuously expanding (Luo et al., 2021) [8].

2- Speech Separation Datasets

The isolation of speech is a crucial undertaking within the realm of speech processing and artificial intelligence (AI) systems. This task is integral to various processes, such as speech recognition, speaker identification, and speech synthesis. In order to develop precise models, scholars and practitioners depend on speech isolation data sets of superior quality that capture uncontaminated and segregated speech samples. This article aims to present a comprehensive overview the most notable speech isolation datasets and their potential contributions to the advancement of AI technologies related to speech. The aforementioned datasets are extensively utilized and have made noteworthy contributions to the domain of speech processing. The first one is The LibriSpeech data set that is a widely utilized publicly accessible resource for research and development in the field of speech. The corpus is composed of roughly 1,000 hours of spoken English language extracted from audiobooks. The aforementioned dataset exhibits a wide array of speakers, recording circumstances, and linguistic material, rendering it exceptionally appropriate for the development of resilient speech isolation models. LibriSpeech divides data by kind and amount. "Train-clean-100" has 100 hours of high-quality, clear speech from 460 speakers, whereas "Train-other-500" offers 500 hours of diverse, loud conversation from 2,496 persons. Researchers can test models on "dev-clean," "dev-other," "test-clean," and "test-other" subsets. ASR, speech synthesis, and speaker recognition employ LibriSpeech. Benchmarking speech processing methods and methodologies is possible due to its availability and scalability. LibriSpeech is used for ASR, speech synthesis, and speaker recognition. Its availability and wide scale make it a significant resource for benchmarking speech processing algorithms and approaches [9].another speech isolation data set is the WHAM!, comprising of The Wall Street Journal and WSJ0 Audio Mixtures, endeavors to tackle a significant obstacle in speech isolation, namely the separation of a desired speaker's voice from the interference of overlapping background noise. The WHAM! Dataset was intentionally created to support scholarly investigations on speech separation in single-channel scenarios, where the availability of only one microphone recording is limited. WHAM! Simulates varied acoustic settings by mixing heterogeneous speech samples from the WSJ0 corpus with artificial room impulse responses. The system enables numerous training and assessment subsets with varied SNRs and reverberation durations. This feature helps create resilient models that can handle real-world challenges. Two training subsets—"wham_noise" and "wham"—make up the dataset. 20,000 mixed recordings and 40,000 mixed recordings with simulated reverberation make up the two subsets. WHAM! Also supplies a subset for assessment, "wham_test," comprising 10,000 mixed-source recordings. This subset tests the model on new data. WHAM! has been widely used to build speech separation and enhancement algorithms, improving audio restoration, teleconferencing, and voice assistants [10].in addition the MUSAN (Music, Speech, and Noise) dataset is a significant asset for the purpose of training and assessing speech isolation models amidst a wide range of acoustic backgrounds. In contrast to preceding data sets, MUSAN is designed to encompass a diverse array of non-verbal audio sources, such as music and assorted forms of ambient sound. The MUSAN dataset comprises a composite of licensed musical pieces, recordings of ambient sounds, and artificially produced noise samples. The dataset provides a degree of adaptability with regards to the categories and degrees of disruption that can be incorporated into speech recordings. The variability inherent in the data allows researchers to replicate authentic situations in which speech signals necessitate separation from diverse categories of background sources. The MUlti-Source Audio Network (MUSAN) dataset comprises various subsets, namely "music," "speech," and "noise," each of which encompasses unique audio clips that represent the respective category. Furthermore, it offers diverse subcategories, including "music_speech" and "music_noise," in which speech signals are superimposed with music or noise, correspondingly. The utilization of mixed subsets enables researchers to assess the efficacy of their models in discriminating speech from diverse sources of interference. The MUlti-Source Audio-visual recordings for Sound Analysis (MUSAN) dataset has been extensively employed in the advancement of speech separation techniques, as well as in associated domains such as audio event detection, noise resilience, and audio source localization. The incorporation of music and a variety of noise sources renders it a valuable instrument for the purpose of training models that are capable of effectively processing intricate acoustic environments [11].the last dataset that will be mentioned is the VoxCeleb dataset that is comprises a vast compilation of speech recordings featuring numerous celebrities. The objective is to furnish a heterogeneous and inclusive group of presenters, facilitating investigations on tasks that rely on speaker-specific characteristics, such as speaker recognition, speaker authentication, and speaker segmentation. The VoxCeleb dataset comprises a vast collection of more than one million spoken utterances, featuring a diverse range of speakers exhibiting a multitude of accents, languages, and speaking modalities. The dataset comprises audio excerpts sourced from interviews, YouTube videos, and other publicly available sources, thereby providing a diverse and varied collection of speech samples. The substantial and varied VoxCeleb dataset has made a noteworthy contribution to the progress of research related to speakers. The advancement of speaker recognition models has been instrumental in enabling precise identification and verification of individuals through their speech characteristics. VoxCeleb has played a significant role in addressing obstacles such as speaker recognition across different languages and domains [12].

3- Different Voice Separation Methods

3.1. Mask-based Voice Separation

The technique of mask-based voice separation is founded on the concept of approximating a binary or soft mask, which denotes the existence or non-existence of the intended speech signal at every time-frequency bin. The application of a mask to the spectrogram or time-frequency representation of a mixture serves to mitigate or eliminate the presence of interfering sources, thereby augmenting the lucidity and comprehensibility of the intended speech signal. Deep learning methods like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) are frequently used to estimate the voice separation mask. The mixed spectrogram serves as the input for these models' training, and the optimal binary or soft mask that best captures the target speech serves as the output. The process of training entails the reduction of the difference between the predicted mask and the ground truth mask. This is achieved through the application of techniques such as supervised learning or optimization methods based on deep learning. The models acquire the ability to comprehend the spectral and temporal attributes of speech and noise sources, thereby facilitating the production of precise masks for voice separation. The utilization of mask-based voice separation has been found to have various applications.

The utilization of mask-based voice separation has been observed in diverse speech processing domains, leading to enhanced performance in multiple areas:

1- The quality of input for automatic speech recognition (ASR) systems can be improved through mask-based voice separation, which involves isolating the target speech from interfering sources. The process of isolating the speech signal from ambient noise or overlapping speakers results in enhanced precision and resilience of automatic speech recognition (ASR) models, thereby facilitating superior conversion of speech to text.

2- The utilization of mask-based voice separation has proven to be beneficial in the process of speaker diarization, which involves the identification and distinction of speakers within an audio recording. The utilization of separation masks allows for the isolation of distinct speakers' voices, which in turn enhances the precision of speech segment segmentation and clustering. This process is based on the identification of individual speakers and ultimately leads to an improvement in the performance of diarization systems.

3- The utilization of voice separation masks is a viable approach for enhancing speech in various applications, particularly in scenarios where the objective is to mitigate ambient noise or ameliorate the quality of a deteriorated speech signal. Mask-based voice separation algorithms aid in improving the clarity and perceptual fidelity of isolated speech by selectively attenuating or eliminating undesired sources [13].

3.2. Machine Learning and Deep Learning Algorithms

The progress made in machine learning and deep learning algorithms has considerably propelled the domain of speech separation, facilitating the isolation of individual speech signals from intricate acoustic surroundings. The present article delves into various notable algorithms employed in the domain of speech separation.

1- Non-Negative Matrix Factorization (NMF):

Non-Negative Matrix Factorization (NMF) is a conventional machine learning technique that has been extensively employed for speech separation applications. The aforementioned technique offers a potent mechanism for breaking down an amalgamated audio signal into a composite of non-negative fundamental vectors that embody distinct sources. The fundamental concept underlying Non-negative Matrix Factorization (NMF) is to represent the provided audio mixture as a linear combination of non-negative constituents, with each constituent corresponding to a distinct source signal. The Non-negative Matrix Factorization (NMF) methodology posits that the observed mixture is a product of a linear combination of sources. The objective of NMF is to estimate the non-negative basis vectors and their corresponding weights to recover the individual sources. NMF is utilized in the domain of speech separation, where it functions on the magnitude spectrogram of the audio mixture. The spectrogram of magnitude portrays the varying frequency content of the mixed signal over time. Non-negative matrix factorization (NMF) aims to decompose the magnitude spectrogram into two matrices that are non-negative: the basis matrix, which represents the spectral patterns or basis vectors, and the activation matrix, which represents the weights or activations associated with each basis vector [14].

2- Independent Component Analysis (ICA):

The Independent Component Analysis (ICA) is a widely employed signal processing methodology that is utilized for the purpose of speech separation and source separation in a general sense. The objective is to decompose a collection of amalgamated signals into their constituent source components by relying on the premise that the sources are statistically independent from one another. In the domain of speech separation, Independent Component Analysis (ICA) regards the perceived amalgamation as a linear amalgamation of autonomous source signals. The objective is to derive an estimation of a mixing matrix that is capable of restoring the initial sources by reversing the process of mixing. Independent Component Analysis (ICA) is predicated on the assumption that the statistical characteristics of the sources are disparate, and endeavors to identify a collection of autonomous components that effectively encapsulate the fundamental sources. The iterative process of estimating the independent components is the fundamental mechanism underlying the functioning of the ICA algorithm. The objective is to optimize the statistical independence of said components through the reduction of mutual information between them [15].

3- Deep Clustering:

The application of Deep Clustering involves the integration of clustering algorithms with deep neural networks for the purpose of speech separation. The clustering structure is acquired through the process of mapping time-frequency bins of an audio mixture to an embedding space, wherein the proximity of bins from the same source is ensured. Subsequently, clustering algorithms are utilized to allocate the bins to particular sources, thereby segregating the amalgamated speech signals. The Deep Clustering technique has the ability to effectively manage intricate mixtures that contain overlapping sources. This approach has been further improved through the utilization of advanced methodologies such as deep attractor networks and mask estimation. The effectiveness of this approach is contingent upon the presence of annotated training data, and the accuracy of the outcomes is subject to the quality of the annotations that are accessible. In general, Deep Clustering represents a potent methodology for speech separation, facilitating the isolation and differentiation of sources through their spectral patterns [16].

4- Deep Attractor Network (DANet):

The Deep Attractor Network (DANet) is a sophisticated deep learning algorithm that is employed for the purpose of speech separation. The methodology employed involves the utilization of deep neural networks to approximate attractor points in the time-frequency domain. These points correspond to the desired sources present in a given mixture. By utilizing the spatial information inherent in multi-channel audio signals, DANet effectively merges the benefits of deep learning and source localization. Through the process of approximating the attractor points, the DANet methodology facilitates the disentanglement of distinct sources from a composite mixture. The findings have demonstrated favorable outcomes in situations where there are intersecting sources, and possess the capability to enhance the caliber of segregated speech signals. The utilization of DANet has been observed in diverse applications, such as music source separation and speech enhancement tasks, thereby making noteworthy contributions to the progress of speech separation technology [17].

5- Wave-U-Net:

The Wave-U-Net is a deep learning framework that has been tailored to perform source separation in audio signals. The operation is performed at the waveform level, eliminating the necessity for spectral representations. The Wave-U-Net architecture utilizes a structure similar to U-Net, comprising of both encoder and decoder pathways, which facilitates the extraction of hierarchical features from the input mixture. Wave-U-Net is capable of efficiently capturing local and global dependencies in audio signals through the utilization of dilated convolutions and skip connections. The technology has exhibited exceptional efficacy in segregating sources from amalgamations of music and speech, providing a potent instrument for professionals in the audio engineering, research, and music production domains, for activities like audio restoration, vocal isolation, and remixing. Wave-U-Net's adaptability and multi functionality render it a significant addition to the domain of audio source separation [18].

6- Permutation Invariant Training (PIT):

The Permutation Invariant Training (PIT) approach is a methodology employed in source separation assignments to tackle the issue of permutation ambiguity. The objective is to facilitate the training of a deep learning model that can generate source estimates that remain unchanged regardless of the sequence of sources in the mixture. The Permutation Invariant Training (PIT) technique involves the examination of every conceivable arrangement of the approximated sources and the actual sources, with the aim of identifying the arrangement that results in the lowest cost function. The cost function in question may be the mean squared error or the signal-to-distortion ratio. The aforementioned process facilitates the congruence between the approximated sources and the veritable sources, thereby successfully resolving the ambiguity of permutation. The efficacy of PIT has been demonstrated in enhancing the separation quality of source separation systems based on deep learning. This has facilitated the production of precise and well-coordinated approximations of the primary sources present in audio mixtures [19].

4- Brief Comparison of Speech Separation Studies

The objective of this review article is to furnish a concise evaluation of the aforementioned speech separation investigations, emphasizing their fundamental features, advantages, and drawbacks. The objective of this study is to analyze and contrast these methodologies in order to acquire knowledge about the most advanced techniques for speech separation and to pinpoint possible avenues for future investigations. The review paper provides a thorough examination that can assist researchers, practitioners, and system designers in the selection of appropriate methods for their particular speech separation tasks. This contribution can potentially advance the field as a whole. The below table "table 1" is represents the current studies in the field of speech separation and recognition.

Table 1: a brief comparison of the current studies

paper	Year	method	Data set	Result
[20]	2021	SepFormer, a novel RNN-free Transformer-based neural network	WSJ0-2/3mix datasets	SI-SNRi of 22.3 dB on WSJ0-2mix and an SI-SNRi of 19.5 dB on WSJ0-3mix
[21]	2020	two-step training procedure: first	(WSJ0) corpus	Si-sdr for time: 15.4

		learn transform (and it's inverse) then train a separation module		Si-sdr for latent: 16.1
[22]	2018	Deep clustering (DC) and deep attractor networks (DANs)	WSJ	sdr :10.1 sir :17.4 sar:11.4
[23]	2020	TRANSFORMER WITH TIME-RESTRICTED SELF-ATTENTION	wsj1-2mix	wer 12.08
[24]	2019	Probabilistic PIT	TIMIT and Chime datasets	Sdr:11 Sir:8
[25]	2020	single GAN model	Produced audio	Snr: 3.06 Rms: 0.10 Sir:11.70
[26]	2018	Time-domain Audio Separation Network (TasNet)	WSJ0-2mix dataset	SI-SNRi :10.8 SDRi :11.1
[17]	2017	DEEP ATTRACTOR NETWORK	Wall Street Journal dataset	GNSDR: 10.5 GSAR: 11.1 GSIR:22.2
[27]	2021	End to end modular system	LibriCSS	Sdr: 14.1
[28]	2021	spatio-temporal recurrent neural network based beam former (RNN-BF)	mandarin audiovisual corpus	PESQ: 3.56 Si-SNR: 15.84 SDR: 16.38 WER: 11.36
[29]	2018	Wave-U-Net	VCTK dataset	SSNR : 9.98
[30]	2022	Paraformer mode	AISHELL1,2	CER:6.19

[31]	2018	END-TO-END MULTI-SPEAKER JOINT CTC/ATTENTION-BASED ENCODER-DECODER	WSJ corpus and the wsj0-2mix	Cer: 10.93 Wer: 18.44
[32]	2023	Improved Transformer-Based Dual-Path Network	Bank + DEMAND dataset	Pesq:3.14 CSIG:4.45 CBAK:3.84
[33]	2020	joint training structure of two deep neural networks (DNNs)	TIMIT	6.93 on SDR
[34]	2020	end-to-end ASR objective	(WSJ) corpus	7.45 on SDR
[35]	2020	Tasnet-Blstm	Wsj0-2mix	11.9 of SI-SDR
[36]	2021	Oracle Masks	Wsj0-2mix	12.8 of SI-SDR
[37]	2018	Convolutional Non-Negative Matrix Factorization (CNMF)	TIMIT	0.95% STOI
[38]	2021	Complex Domain with Long Short-Term Memory Neural Network	TIMIT	83.52% STOI

5-Conclusion

This review paper delves into an examination and comparison of diverse speech separation investigations, with a particular emphasis on their datasets, methodologies, and efficacy. The findings of our analysis indicate that the selection of a data set is a pivotal factor in the assessment of various techniques' efficacy. It has been observed that the utilization of extensive and varied data sets in studies tends to result in superior separation quality and resilience. Furthermore, deep learning techniques, including deep clustering, Wave-U-Net, and DANet, have exhibited exceptional efficacy in isolating sources from intricate mixtures, outperforming conventional methodologies such as NMF and ICA. The utilization of neural networks in deep learning techniques enables the extraction of complex features and the identification of underlying structures within audio signals.

Furthermore, the studies that were reviewed have brought attention to the difficulties that are linked with speech separation. These challenges include the problem of permutation ambiguity, the requirement for labelled training data, and the susceptibility to noise and reverberation. Tackling these obstacles continues to be a thriving field of study. Additionally, the examination of various techniques exposes compromises among computational intricacy, segregation excellence, and exigencies of real-time processing.

To summarize, the present review article offers significant perspectives on the contemporary scenario of speech separation methodologies, emphasizing the significance of datasets, the effectiveness of deep learning-oriented methodologies, and the extant challenges. Subsequent investigations ought to priorities the advancement of algorithms that are more resilient and effective, the examination of supplementary evaluation metrics, and the integration of domain expertise to augment the performance of speech separation. The progression of speech separation technology has the potential to facilitate various applications, such as the enhancement of speech recognition systems, the augmentation of audio communications, and the customization of audio experiences.

6. REFERENCES

[1] Hershey, J. R., Chen, Z., Le Roux, J. and Watanabe, S., "Deep clustering: Discriminative embeddings for segmentation and separation," vol. 1, no. 2, pp. 31-35, 2016.

[2] Kolbæk, M., Yu, D., Tan, Z. H. and Jensen, J., "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 10, pp. 1901-1913, 2017.

[3] Graves, A., Fernández, S., Gomez, F., Schmidhuber and J., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *In Proceedings of the 23rd international conference on Machine learning,* vol. 2, no. 11, pp. 369-376, 2006.

[4] Chan, W., Jaitly, N., Le, Q., Vinyals and O. , "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP),* vol. 3, no. 5, pp. 4960-4964, 2017.

[5] Michelsanti, D., Tan, Z. H., Zhang, S. X., Xu, Y., Yu, M., Jensen, J. and Yu, D., "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 29, no. 3, pp. 1368-1396, 2021.

[6] Subramanian, A. S., Weng, C., Yu, M., Zhang, S. X., Xu, Y., Watanabe, S. and Yu, D. , "Far-field location guided target speech extraction using end-to-end speech recognition objectives," *In ICASSP 2020-2020*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 19, no. 7, pp. 7299-7303, 2019.

[7] Malik, M., Malik, M. K., Mehmood, K. and Makhdoom, I., "Automatic speech recognition: a survey," *Multimedia Tools and Applications,* vol. 80, no. 3, pp. 9411-9457, 2021.

[8] Koteswararao, Y. V. and Rao, C. R., "Multichannel speech separation using hybrid GOMF and enthalpy-based deep neural networks," *Multimedia Systems,* vol. 27, no. 2, pp. 271-286, 2021.

[9] Panayotov, V., Chen, G., Povey, D. and Khudanpur, S., "Librispeech: an asr corpus based on public domain audio books," *In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP),* vol. 2, no. 9, pp. 5206-5210, 2016.

[10] Wichern, G., Antognini, J., Flynn, M., Zhu, L. R., McQuinn, E., Crow, D., ... and Roux, J. L., "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv,* pp. 1907-1913, 2020.

[11] Snyder, D., Chen, G. and Povey, D., "Musan: A music, speech, and noise corpus," *arXiv preprint,* pp. 105-112, 2015.

[12] Nagrani, A., Chung, J. S. and Zisserman, A., "Voxceleb: a large-scale speaker identification dataset," *arXiv,* vol. 3, no. 11, pp. 1706-1711, 2017.

[13] Zhang, Y., Liu, Y. and Wang, D., " Complex ratio masking for singing voice separation," *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* vol. 12, no. 3, pp. 41-45, 2021.

[14] Lee, D. and Seung, H. S., "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems,* vol. 13, no. 1, pp. 22-27, 2001.

[15] Hyvärinen, A. and Oja, E. , "Independent component analysis: algorithms and applications," *Neural networks,* vol. 13, no. 4, pp. 411-430, 2000.

[16] Luo, Y., Chen, Z., Hershey, J. R., Le Roux and J., & Mesgarani, N., "Deep clustering and conventional networks for music separation: Stronger together.," *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP),* vol. 2, no. 3, pp. 61-65, 2018.

[17] Luo, Y. and Mesgarani, N., "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.

[18] Lin, K. W. E., , Balamurali, B. T., Koh, E., Lui, S. and Herremans, D., "Singing voice separation using a deep convolutional neural network trained by ideal binary mask and cross entropy," *Neural Computing and Applications,* vol. 32, no. 4, pp. 1037-1050, 2021.

[19] Yu, D., Kolbæk, M., Tan, Z. H. and Jensen, J. , "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* vol. 21, no. 5, pp. 241-245, 2017.

[20] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M. and Zhong, J., "Attention is all you need in speech separation," *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* vol. 3, no. 12, pp. 21-25, 2021.

[21] Tzinis, E., Venkataramani, S., Wang, Z., Subakan, C. and Smaragdis, P. , "Two-step sound source separation: Training on learned latent targets," *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* vol. 16, no. 3, pp. 31-35, 2010.

[22] Drude, L., von Neumann and T., Haeb-Umbach, R., "Deep attractor networks for speaker re-identification and blind source separation," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* vol. 13, no. 8, pp. 11-15, 2018.

[23] Chang, X., Zhang, W., Qian, Y., Le Roux, J. and Watanabe, S., "End-to-end multi-speaker speech recognition with transformer," *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) ,* vol. 13, no. 2, pp. 6134-6138, 2020.

[24] Yousefi, M., Khorram, S. and Hansen, J. H. , "Probabilistic permutation invariant training for speech separation," *arvix,* vol. 12, no. 5, pp. 132-139, 2019.

[25] Narayanaswamy, V., Thiagarajan, J. J., Anirudh, R. and Anirudh, R., " Unsupervised audio source separation using generative priors," *Unsupervised audio source separation using generative priors.,* vol. 11, no. 5, pp. 1367-1372, 2020.

[26] Luo, Y. and Mesgarani, N. , "TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* vol. 12, no. 3, pp. 696-700, 2018.

[27] Jiang, Y., Qiu, Y., Shen, X., Shen, X. and Liu, H. , "SuperFormer: Enhanced Multi-Speaker Speech Separation Network Combining Channel and Spatial Adaptability," *Applied Sciences,* pp. 112-118, 2022.

[28] Xu, Y., Zhang, Z., Yu, M., Zhang, S. X., Chen, L. and Yu, D., " Generalized RNN beamformer for target speech separation," *CoRR,* vol. 2, pp. 112-117, 2021.

[29] Macartney, C. and Weyde, T., "Improved Speech Enhancement with the Wave-U-Net," *ArXiv,* vol. 11, no. 7, pp. 1811-1817, 2018.

[30] Gao, Z., Zhang, S., Mcloughlin and I., Yan, Z. , "Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition," *Interspeech,* vol. 4, no. 2, pp. 167-173, 2022.

[31] Chang, X., Qian, Y., Yu, K. and Watanabe, S., "End-to-end monaural multi-speaker ASR system without pretraining," *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) ,* vol. 34, no. 10, pp. 6256-6260, 2019.

[32] Ye, M. and Wan, H., "Improved Transformer-Based Dual-Path Network with Amplitude and Complex Domain Feature Fusion for Speech Enhancement," *Entropy,* vol. 25, no. 2, pp. 228-232, 2023.

[33]

[34] Hassan, H. S., "Hybrid Filter for Enhancing Input Microphone-Based Discriminative Model," *Iraqi Journal of Science,* pp. 2434-2439, 2020.

[35] Togami, M., "Joint training of deep neural networks for multi-channel dereverberation and speech source separation," *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) ,* vol. 23, no. 1, pp. 3032-3036, 2020.

[36] Maciejewski, M., Shi, J., Watanabe, S. and Khudanpur, S., "Training noisy single-channel speech separation with noisy oracle sources: A large gap and a small step," *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* vol. 12, no. 6, pp. 5774-5778, 2021.

[37] Wang, Z., Le Roux, J., Wang, D., Hershey and J.R., "End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction," *arvix,* vol. 6, no. 10, pp. 1804-1809, 2018.

[38] Zhu, C., Huang, D., Huang, D., Chen, Y., Lin, J. and Jiang, D., " A Robust Unsupervised Method for the Single Channel Speech Separation," *2019 15th International Conference on Computational Intelligence and Security (CIS),* vol. 35, no. 5, pp. 387-390, 2019.