International Journal of Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2024.1.5

Volume 10, Issue 1
January - 2024

A Review: Speech Separation Techniques and Its Applications

Ahmed Abd Ali Abdulkadhim[1], Zainab Mohammed Essa[2], Muntaha Abdulzahra[3]

Department of Computer Science

Collage of education,

Mustansiriyah University Baghdad

Iraq

ABSTRACT

A fundamental task in signal processing, speech separation has many practical applications. For example, it can be used to improve the accuracy of automatic speech recognition by separating clear speech from noisy speech signals. When all that is available is a monaural recording of mixed speech, the task of extracting distinct sources from a complex mixture appears to be challenging for automatic calculation systems, in contrast to the remarkable ability of the human auditory system to do so. Monaural speech separation poses numerous challenges, but over the years, numerous attempts have been made in prior works to address this issue. Within the study that is being presented, we offer an extensive analysis of current research on speech separation and its uses.

Keywords: Automatic speech recognition, Clear speech, Noisy speech signals, Speech Separation.

1. INTRODUCTION

Voice communication and voice information are becoming more and more important in our lives. Examples of voice communication include using voice commands to operate programs for mobile phones, sending voice messages via chat software, making voice calls, identifying the singers from musical compositions, and figuring out the details, song lyrics, and style [1-3]. Splitting mixed speech into its two original speech enunciations is the aim of speech separation. Speech separation is a fundamental task in signal processing with many applications, such as song separation, speaker recognition, and mobile communication. There are numerous possible values to distinguish between mixed speech. These days, speech separation is becoming more and more crucial to speech processing, and an A growing number of devices need to perform speech recognition tasks. While speech separation comes naturally to humans, creating an automated system that can match the human auditory system is extremely difficult. [4] Speech separation has therefore always been a crucial area for research. The way speech separation operates is by dissecting the combined signals into their component parts. For this separation, researchers employ a range of methods and algorithms. Among the methods that are frequently employed is the blind source separation (BSS) method [5-7]. Using this method, the sources are distinguished from one another by first analyzing the mixed signals for statistical properties. Speech separation is also accomplished through the use of additional techniques like the independent component analysis (ICA) method and deep neural network (DNN) method. Over the past few years, It has put forth a number of speech separation techniques. We thoroughly examine speech separation and its uses in this work. This research article's remaining sections are arranged as follows: A review of some of the schemes put forth in the past ten years is covered in part 2. The comparative evaluation of the techniques covered in Part 2 is presented in Part 3. Part 4 presents conclusions at the end.

2. LITERATURE SURVEY

Numerous technologies are utilized in the investigation of speech separation, and some of this research will be covered in this section.

The work carried out by ZHANG Meng et al., [8], for speech separation, a multi-joint function and three joint constraint loss functions based on the dual-output deep neural network were presented. In addition to limiting The

DNN separation model's multiple joint-constraint loss function, or ideal ratio mask (IRM) errors of the two outputs also limits the relationships between the estimated IRMs and the mixed signal magnitude spectrogram, the estimated IRMs of the two outputs, and the clean speech signal magnitude spectrogram. Three parameters are used to adjust the constraint strength in order to improve the accuracy of speech separation.

In this work, a dual-path hybrid attention network (DPHA-Net) based on time-domain separation of monaural speech was proposed by WENBO QIU AND YING HU [9]. The DPHA module, which is an essential part of DPHA-Net, is made up of several attentions and is intended to capture the dependencies between short- and long-term context information. The adaptive feature fusion (AFF), element-wise attention (EA), and multi-head self-attention (MHSA) units make up the DPHA module. During the training, we suggested an enhanced multi-stage aggregation training approach.

Yigitcan Özer and Meinard Müller, propose a node encoder for squash standard deep pooling (ESDC) by integrating node coding, squash standard and deep pooling (DC) as a reinforcement discriminative learning framework. To start, intermediate features are created using a node encoder. A learning approach for graph representations based on matrix factorization is used to develop node encoders. It produces recognizable intermediate features that are crucial for enhancing functionality. The separation block then uses these distinguished intermediate features as input features. In the end, the decoder block uses the clustering method to build the estimation mask and reconstructs the estimated signal for every source. Specifically, we normalize the input and output vectors using Squash norm in order to improve the differentiation between high-dimensional embedding vectors [10].

The research of JAE-MO LEE1 JOO-HYUN [11], in order to define the NAS search space for Conv-TasNet, YANG proposed possible operations. Also, we present a low computational cost NAS to overcome the large training memory requirements of the backbone model. Use two search techniques based on gradient descent and reinforcement learning to identify optimal separation unit structures to avoid imbalance when applying NAS. To avoid imbalance when applying NAS, they introduced a suitable auxiliary loss method for the Conv-TasNet architecture.

The work, The Single-Path Global Modulation (SPGM) block was proposed by Shengkui Zhao and Jia Qi Yip [12] as a replacement for inter-blocks. The structure of SPGM—which consists of a parameter-free global pooling module and a modulation module that contains only 2% of the model's total parameters—gives rise to its name. The SPGM block makes the model single-path overall by enabling all transformer layers to be devoted to local feature modeling. With up to eight times fewer parameters than previous SOTA models, SPGM surpasses Sepformer's performance by 0.5 and 0.3 dB, respectively, and achieves 22.1 dB SI-SDRi on WSJ0-2Mix and 20.4 dB SI-SDRi on Libri2Mix.

The work carried out Chun-Miao Yuan, Xue-Mei Sun, and Hu Zhao [13] They proposed a speech separation model that makes use of attention mechanisms and convolutional neural networks. The input's high dimensionality is seen in the mixed speech signals' magnitude spectrum. Through an examination of the features of both the CNN and the attention mechanism, it is possible to determine that while the attention mechanism can lessen the loss of sequence information, the CNNis more effective at mining the spatiotemporal structure information and extracting low-dimensional features from speech signals. By merging two mechanisms, speech separation accuracy can be effectively increased.

For end-to-end speech separation, Jingjing Chen, Qirong Mao, and Dong Liu [14] introduced the dual-path transformer network (DPTNet). Context awareness is directly integrated into speech sequence modeling by this network. By implementing an enhanced transformer that permits direct interaction between elements in the speech sequences, DPTNet is able to model speech sequences with direct context-awareness. Our improved transformer can learn the speech sequence order information without positional encodings by incorporating a recurrent neural network into the original transformer. Moreover, our dual path structure model works well for modeling very long speech sequences.

DOI: 10.31695/IJASRE.2024.1.5

Changsheng Quan and Xiaofei Li [15] described an end-to-end narrow-band network in this work that outputs separated signals of the same frequency and takes in multi-channel mixture signals of a single frequency as input. The inter-channel difference, or spatial information, is a potent tool for narrow-band speaker position discrimination. This information is critical to many narrow-band speech separation methods, such as beamforming and spatial vector clustering. To make use of this information and automatically distinguish speech, a rule is learned by the proposed network. All frequencies share the network since all frequencies should be subject to this rule. Furthermore, a criterion for invariant training for full-band permutations is presented to address the frequency permutation issue that the majority of narrow-band techniques encounter.

3. COMPARISON AND ANALYSIS OF SCHEMES

Table 1 provides a comparison of the previously discussed systems.

Table 1 comparison table of different systems.

Ref	methods	gender	SIR(dB)	SDR(dB)		SAR(dB)	SNR
1	DNN	F-F	9.045	5.9041		7.6908	--------
		M-M	9.994	6.3496		8.1946	--------
		F-M	13.8387	8.6556		9.3879	--------
2	DPHA-Net	-------	--------	20.9		--------	20.7
3	ESDC	-------	20.67	12.85		13.84	
		M-M	--------	8.25		--------	7.24
		F-F	--------	90.5		--------	8.54
4	NAS-TasNets	-------	--------	17.50		--------	--------
5	SPGM	-------	--------	WSJ02Mix	Libri2Mix	-------	--------
				22.1	20.4		
6		-------	0.51	0.27		--------	--------
7	DPTNet	-------	--------	20.6		20.2	--------
8	prop	----	--------	13.89		--------	--------

4. CONCLUSION

In order to identify the key elements of speech separation performance and to gather and compare data, we examined a timeline of various speech separation techniques for the years 2019–2023.

REFERENCE

[1]. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021, June). Attention is all you need in speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 21-25). IEEE.

[2]. Luo, Y., & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing, 27(8), 1256-1266.

[3]. Chen, Z., Yoshioka, T., Lu, L., Zhou, T., Meng, Z., Luo, Y., ... & Li, J. (2020, May). Continuous speech separation: Dataset and analysis. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7284-7288). IEEE.

[4]. Zeghidour, N., & Grangier, D. (2021). Wavesplit: End-to-end speech separation by speaker clustering. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 2840-2849.

[5]. Luo, Y., Chen, Z., & Yoshioka, T. (2020, May). Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 46-50). IEEE.

[6]. Wang, Z. Q., Wichern, G., & Le Roux, J. (2021). On the compensation between magnitude and phase in speech separation. IEEE Signal Processing Letters, 28, 2018-2022.

DOI: 10.31695/IJASRE.2024.1.5

[7]. Gu, R., Zhang, S. X., Xu, Y., Chen, L., Zou, Y., & Yu, D. (2020). Multi-modal multi-channel target speech separation. IEEE Journal of Selected Topics in Signal Processing, 14(3), 530-541.

[8]. Linhui, S., Wenqing, L., Meng, Z., & Ping'an, L. (2023). Monaural Speech Separation Using Dual-Output Deep Neural Network with Multiple Joint Constraint. Chinese Journal of Electronics, 32(3), 493-506.

[9]. Qiu, W., & Hu, Y. (2022). Dual-path hybrid attention network for monaural speech separation. IEEE Access, 10, 78754-78763.

[10]. Özer, Y., & Müller, M. (2024). Source Separation of Piano Concertos Using Musically Motivated Augmentation Techniques. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[11]. Lee, J. H., Chang, J. H., Yang, J. M., & Moon, H. G. (2022). NAS-TasNet: Neural architecture search for time-domain speech separation. IEEE Access, 10, 56031-56043.

[12]. Yip, J. Q., Zhao, S., Ma, Y., Ni, C., Zhang, C., Wang, H., ... & Ma, B. (2023). SPGM: Prioritizing Local Features for enhanced speech separation performance. arXiv preprint arXiv:2309.12608.

[13] Yuan, C. M., Sun, X. M., & Zhao, H. (2020). Speech separation using convolutional neural network and attention mechanism. Discrete Dynamics in Nature and Society, 2020, 1-10.

[14]. Chen, J., Mao, Q., & Liu, D. (2020). Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. arXiv preprint arXiv:2007.13975.

[15]. Quan, C., & Li, X. (2022, May). Multi-channel narrow-band deep speech separation with full-band permutation invariant training. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 541-545). IEEE.

DOI: 10.31695/IJASRE.2024.1.5