ADVANCES IN CYBER-PHYSICAL SYSTEMS

Vol. 9, No. 2, 2024

MODELS AND METHODS FOR SPEECH SEPARATION IN DIGITAL SYSTEMS

Andrii Tsemko[1,2]*, Ivan Karbovnyk*[1]

[1]*Ivan Franko National University of Lviv, 50, Drahomanov Str, Lviv, 79005, Ukraine,*
[2]*Infineon Technologies, 20, Luhanska Str, Lviv, 79000, Ukraine.*
Authors' e-mails: *andrii.tsemko@lnu.edu.ua, ivan.karbovnyk@lnu.edu.ua*

https://doi.org/10.23939/acps2024.02.

Submitted on 23.09.2024

***Abstract*: The main purpose of the article is to describe state-of-the-art approaches to speech separation and demonstrate the structures and challenges of building and training such systems. Designing efficient optimized neural network model for speech recognition requires using encoder-decoder model structure with masks estimation flow. The fully-convolutinoal SuDoRM-Rf model demonstrates the high efficiency with relatively small number of parameters and can be boosted with accelerators, that supports convolutional operations. The highest separation performance has been shown by the SepTDA model with 24 db in SI-SNR with 21.2 million of trainable parameters, while SuDoRM-Rf with only 2.66 million has demonsrated 12.02 db. Another transformer-based neural network approaches has demonstrated almost the same performance as SepTDA model but requires more trainable parameters.**

***Index Terms*: Speech Separation, Speech Enhancement, Audio Processing, Neural Networks**

I.INTRODUCTION

Design of automated speech recognition (ASR) systems considered two types: pipeline or end-to-end architectures [1]. Pipeline ASR architecture consists of two main parts: speech enhancement and actual speech recognition. The idea behind the pipeline architecture is to process audio focusing on different acoustic effects to improve the overall speech intelligibility of the original signal. Various algorithms are used for audio processing for noise suppression, dereverberation, etc. Moreover, training and applying different neural networks for noise suppression and dereverberation is more efficient than training a single neural network model for the same audio processing. Separate models combined required fewer parameters and calculation cycles than a single neural network. Also, training such a model is much harder and requires more data and training epochs, as the task of simultaneous noise suppression and dereverberation is quite a complex task. Likewise, an end-to-end ASR system requires more parameters and operation for the same level of performance compared to a pipeline system.

Designing of noise suppression or dereverberation system is deeply investigated and solved by neural networks, as such systems mainly focus on suppressing non-speech audio components, while the speech separation part of speech enhancement focuses on dealing with two or more speech components. Similar to noise suppression neural network architectures poorly solve speech separation problems, as such neural network kernels are trained to find speech components of the audio and suppress others. Another fundamental difference of neural network model architectures for speech separation is that they generate more outputs than they process as input. It creates another branch of speech separation task as focusing on constant number of speakers or dynamic and requires more complex approaches.

The speech separation problem is hard to solve with the classical programming paradigm, as we have interfered with signals that have the same properties, frequencies, etc. Such approaches focused on spectral separation by classifying speakers as low or high-tone speech, but even for low-male and high-female voices, there are too many interfered frequencies. Another problem is that speech characteristics can vary over time, such as speech speed, tone, pauses, etc. Even in the simplest case, where the speaker's main frequency ranges are pre-defined, the overlapping of the words and syllables is a complex case that appears frequently.

On the other hand, the human ability to focus on a single speaker even in a complex environment with a lot of different speakers and noise sources is highly efficient, that people can understand what a person is saying to them at parties and other noisy conditions. It is important to note, that people's own experiences are not the same as lab experiments of the speech separation systems, as the human brain also processes additional sources of information that help us understand one speaker at a time with information from our eyes, etc. However, the human brain still demonstrates high speech separation in conditions where we use only audio information. People can easily focus and understand one speaker and change their perception to another at any time.

Genuility, speech separation with neural networks can be grouped into deep clustering [2], fully-convolutional [3], reccurent [4], transformer-based [5] approaches. Different applications of ASR systems require different architectures, resources, and latency. The main purpose of this article is to analyze different approaches for the speech separation system for single-channel audio. Describe the idea behind each neural network model, comparison of the performance, and resource usage.

II. LITERATURE REVIEW AND PROBLEM STATEMENT

The speech separation problem is a branch of the source separation task. Before neural network approaches, only the source separation problem was investigated, and solutions that usually work for multi-channel experiments were proposed [6]. Single-channel source separation was not solved until the machine learning paradigm was introduced. Designing algorithms for single-channel source separation is problematic as signals of different sources overlap in time and can be placed in the same frequency ranges of the signal spectrogram. Using deep neural network demonstrates high performance for single-channel source separation and has started the branch of source separation tasks solved with neural networks [7].

Single-channel speech separation is a more complex task as it needs to separate signals for several signals of the same characteristics as speeches of different people still contain features. Another efficient neural network of the encoder-decoder (Fig. 1) architecture with a separator part was proposed that demonstrates the high efficiency of the separation for a single-channel two-speaker experiment [8]. This architecture is using the masks estimation of each speaker, that should be applied to features representation of original signal.

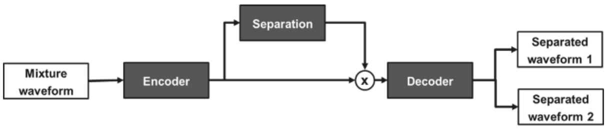

Fig. 1. Encoder-decoder architecture of neural networks for speech separation

Another type of architecture for speech separation is direct processing of the signal, without estimating masks. This approach also reuses the encoder-decoder concept to transfer time-domain input signal into feature maps, but instead of estimating masks and apply them onto the encoder output, the model uses straight processing layer-by-layer to get two or more speakers singals [9]. Direct processing of the feature does not provide an increase in separation of the speech but demonstrates insignificant improvement for non-speech separation.

Both architectures described above are not flexible for dynamic number of speakers. They are designed for a constant number of competitive speakers. Such approachs demonstrate degradation of the performance of the separation, when the number of actual speakers is bigger than expected. Completely different approach with model arechitecture of deep clustering is used to deal with such complex environments. Deep clustering approach contains at least two parts of the system, where one part tries to split the audio into segments, while second part does clusterization of such segments to combine them based on the speaker [10]. Later, such clusterized segments can be combined to generate separated speech.

The experiment setup consists of one microphone. Recorded audio is a mixture x(t) of original speeches $s_i(t)$ of N sources with convolved room impulse response $h_i(t)$:

$$x(t) = \sum_{i=1}^{N} s_i(t) * h_i(t). \qquad (1)$$

The neural network model is designed to process x(t) mixture signal and estimate $\hat{s}_1(t), \ldots, \hat{s}_N(t)$, which should minimize the difference to $s_1(t), \ldots, s_N(t)$.

III. SCOPE OF WORK AND OBJECTIVES

The main goal of the work is to compare and describe different neural network models and approach for single-channel speech separation tasks. This article should be useful to understand state-of-the-art approaches, their main features, and required resources that can help obtain desirable results in designing own speech separation system for strict system requirements, such as memory, latency, and complexity of the model.

This article focuses on the case of two interfered speakers, as most articles focus on it, but generally investigated methods and models can be extended to work with more complex cases with more than 2 speakers. Most of the methods or neural network methods work for a constant number of speakers.

IV. EXPERIMENTS AND EVALUATION METRICS

Evaluation of methods or neural network models requires separate original clean speeches without room impulse responses. Obtaining original clean speeches in real-world conditions is quite a complex task, so algorithm evaluations usually use artificially generated data with pre-recorded speeches and room impulse responses. It is important to use true clean speech datasets without any background noises because it can introduce false bias into evaluation metrics or neural network training.

Generating audio data for evaluation can be done in two ways with or without applying room impulse characteristics. The usage of audio signals with room impulse characteristics can affect training neural networks to do additional tasks such as dereverberation, but for our research, this part of audio processing is lying outside of the scope. For a better comparison of the methods and neural network models will be used generated dataset of the mixture of speeches without applying room impulse responses.

The most popular metric for evaluating source or speech separation approaches is the Scale-Invariant Signal-To-Distortion Ratio (SI-SDR) [11]. This metric is used to compare audio signals with different levels of the signals. It is important to use such metrics as speakers in the mixture can have different levels of loudness, while some of the methods are trained to obtain separate speech normalizes or of the same loudness. Based on that, such metric also can be used as a loss function for training the neural network models. Usage of SI-SDR metric as loss function demonstrates high separation for trained model than standard source-to-distoruin ratio (SDR). In some articles, researchers also used the SI-SNR name interchangeably for SI-SDR. SI-SDR is defined as

$$SI-SDR = 10\log_{10}\left(\frac{\|e_{target}\|^2}{\|e_{residual}\|^2}\right), \quad (2)$$

$$SI-SDR = 10\log_{10}\left(\frac{\left\|\frac{\hat{s}^T s}{\|s\|^2}s\right\|^2}{\left\|\frac{\hat{s}^T s}{\|s\|^2}s-\hat{s}\right\|^2}\right)\ldots \quad (3)$$

It is important to note, that applying SI-SDR or other metrics requires comparing several output channels as each channel contains only one speaker of the original mixture, which requires including some of the cases when original and estimated channels of speakers are permutated. For such cases, as a result of SI-SDR metrics provides maximum values of SI-SDR from all possible permutations of $\hat{s}_1(t)$ and $s_1(t)$. For training of speech separation models frequently used utterance-level permutation invariant training (uPIT) technique that used to calculate loss function for training invariant separation models.

Two popular datasets are used for evaluation of neural network models for speech separation: WSJ0-2mix and Libri2Mix. WSJ0-2mix dataset is generated based on the Wall Street Journal (WSJ0) corpus. A training set constains of 1800 minutes of a mixture of two speakers by randomly selecting speakers and applying different gaining to generate various signal-to-noise ratios from 0 db up to 5 db [12]. LibriMix or Libri2Mix is an open-source dataset, generated based on the LibriSpeech dataset, that contains 3480 minutes of two speakers [13].

V. OVERVIEW OF STATE-OF-THE-ART MODEL FOR SPEECH SEPARATION

The first significant step further in speech separation was done with TasNet neural network architecture, which introduced an efficient encode-decoder structure, that mimicked the ideas from the language modeling [14]. Encoder and decoder blocks are designed to transform the time-sequence input audio signal into a features map, that can be interpreted as a short-time Fourier transformation and vice versa. The TasNet model uses the separation part which estimates two masks of each speaker that should be applied to the encoder output and later transformed back to a time-series signal by decoder block. This architecture demonstrates high efficiency in separation tasks as it uses mask estimation instead of direct signal estimations. For the separation part, the TasNet architecture uses LSTM layers of the recurrent neurons.

The next efficient model is the Con-TasNet (fully-convolutional time-domain audio separation network) model, which is a successor of the TasNet model, that inherited the idea of encoder-decoder structure and mask estimations target [15]. Instead of TasNet, Conv-TasNet is a fully-convolutional model that consists of convolution and transposed convolution operators. Convolution layer is used for encoder and transposed convolution for decoder mimicking the STFT and iSTFT operators – for transforming time-domain signal into frequency-domain. The number of convolution filters, their length, and their dilation factor are similar for spectrogram calculation with STFT as several frequencies and overlap steps. The structure of the ConvBlock is shown in Fig. 2.

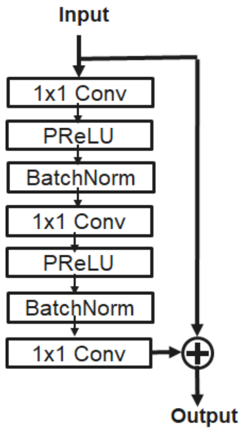

Fig. 2. ConvBlock architecture of Conv-TasNet model

The idea of using the convolutional operators instead of actual STFT is related to the fact that processing of the exact type of signal (speech signal) is more efficient, as neural network can find the best suitable filters and ignore some parts of the spectrum, where speech is not presented. Processing a single slice of the FFT with a convolution neural network is quite efficient, as convolution operators demonstrate high efficiency for processing patterns. However, using the same regular convolution layers for processing the STFT is not as efficient, and using a filter with fixed height and width does not include the relations between the farthest points of the spectrogram, which reduces efficient of processing time signals. For this reason, the original TasNet model includes LSTM layers in the separator part of the model, as the recurrent neurons demonstrate high efficiency for processing the time-series data.

Replace the LSTM or other recurrent neurons with convolutional operators to obtain the same performance in processing time-sequence data was done with TCN (temporal convolutional network), which contains nested convolution layers with different dilation factors. In the Conv-TasNet, the separator contains N number of convolutions with increasing value of the dilator factor that is linked to the index of the convolution layer. The more layers the model contains, the higher efficiency of the processed signal separations is. The proposed implementation of the Conv-TasNet demonstrates the overperforming of the original TasNet structure for 5 db in SI-SDR metric for WSJ0-2mix dataset with 10.8 db SI-SDR for TasNet and 15.3 db SI-SDR for Conv-TasNet.

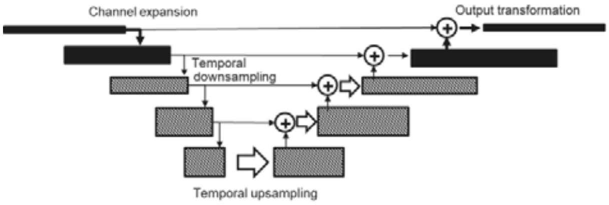

Fig. 3. U-ConvBlock architecture of SuDoRM-RF model

The fully-convolutional implementation of the speech separation model also has another benefit as a fully convolutional structure, which makes it possible to use accelerators to boost calculations and reduce latency. It is important for ASR systems, as they are usually built as low-latency applications. Using models with LSTM or other recurrent layers requires more time for calculations, as processing recurrent flow.

Next fully-convolutional model that uses the encoder-decoder structure is SuDoRM-RF (Successive Down-sampling and Resampling of Multi-Resolution Features), which is a different convolutional approach [16]. This model proposes another separation design. It contains a group of sequential U-ConvBlock of U-Net structures with skip connections (Fig. 3). This approach demonstrates higher separation rate than Conv-TasNet even though requires 2 time less parameters. The SuDoRM-RF model has 17.02 db SI-SNR versus 15.3 db of Conv-TasNet. Original Conv-TasNet model has 5.05 million of trainable parameters, while SuDoRM-RF uses only 2.66 million that is almost twise less. As architecture of this model is also fully-convolutional, these models demonstrate highest opt

Another deep learning model is the Deep Casa model (deep learning and computational auditory scene analysis), which is designed as a two-step approach based on the divide-and-conquer idea [17]. The Deep Cases model uses two types of grouping: simultaneous and sequential. Firstly, the simultaneous grouping stage is done by a neural network that separates the spectral components of each speaker at the level of audio frames. While TasNet and Conv-TasNet models process time-domain signal, Deep Casa processes spectrogram calculated by STFT. The simultaneous grouping step is done by the Dense-UNet structure, which contains a fully-connected layer of neurons, a convolution layer, pooling, and up-sample layers. It calculates signal characteristics for mask estimations that apply to the original spectrogram of the input signal. It generates two spectrograms of speakers that transform back into the time domain by inverse STFT. Obtained signals are later used as an input to the second neural network, that performs grouping of the signal based on the clusterization process. Additional input used is the time domain representation of the signal, which was processed by the TCN neural network, which helps to classify and group parts of the signal of each speaker. Such skip connections and the structure of the overall system make it more complex compared to TasNet and Conv-TasNet models. Implementation of such complex systems and skip connections usually requires more RAM memory and is hard to accelerate with GPUs. The proposed DeepCasa model demonstrates a 3.1 db improvement relative to the TasNet model but also requires 4M parameters more, which makes the model bigger by 45%.

Another approach, which is quite similar to Deep Casa, is Wavesplit [18]. It also contains several models inside, that focus on different parts of the overall system (Fig. 4). The first model is used to split the input mixture into a part of the separate speaker by frames, while the second model processes these speaker vectors, customizes them, and combines them into grouped channels for each speaker. The second neural network processes the speaker vectors and pre-calculated clusters to group vectors by the speaker. Both neural networks for the estimation of speaker vectors and group vectors by clusters use ResNet architecture. The Wavesplit split model demonstrates improvement in speech separation compared to a model with a similar approach Deep Casa a 3.3 db with 21 db of SI-SNR versus 17.7 db SI-SNR for Wavesplit and Deep Casa, respectively. The authors do not provide the number of parameters of the model. Possibly, the size of this approach is even bigger than the Deep Casa model, as ResNet models usually use skip connection between layers, which leads to increasing the number of trainable parameters. The authors of the next SepFormer model mention the size of the Wavesplit model in their comparison and set it to be 29 million, which is more than 2 times bigger than the Deep Casa model.

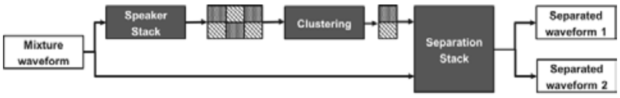

Fig. 4. Wavesplit model architecture

The next model is a speech separation model of transformer architecture called SepFormer [19]. Transformer demonstrated boost in processing sequential data for other tasks. Proposed model also reuses the encoder-decoder structure proposed for TasNet by using convolutional and transposed convolutional layers for the encoder and decoder respectively to transform the input signal into a pseudo-spectrogram that is used as an input for the separator part of the model (Fig. 5). This separator part is a SepFormer block that consists of two sequential layers of the transformer. The first layer is named IntraTransformer (IntraT) and designed to process short-term dependencies, while the second layer entitled InterTransformer (InterT) processes the longer-term dependencies of the audio characteristics. As in the TasNet, the SepFormer generates two masks that are applied to the encoder output and transferred back into the time domain with a transposed convolution operator. The proposed SepFormer model demonstrates 2 times higher SI-SNR metric than TasNet, but demonstrates a slight decrease in the SI-SNR metric compared to Wavesplit, where SepFormer demonstrates 20.4 db SI-SNR versus 21 db of Wavesplit. It is important to note that while SepFormer is slightly below the Wavesplit in the performance comparison, it requires 3M parameters less.

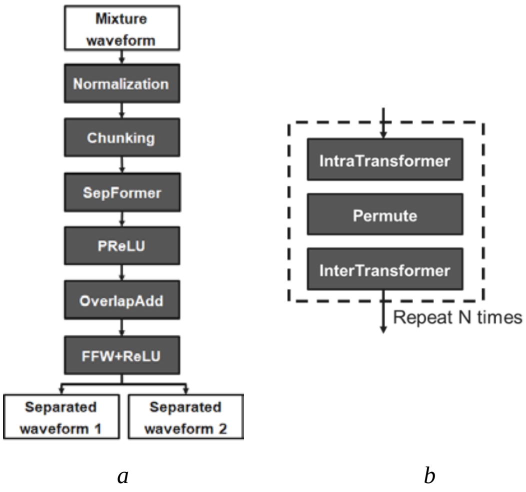

Fig. 5. SepFormer model architecture (a); SepFormer block structure (b)

Monaural Speech Separation TransFormer (Moss Former) is another deep neural network approach that uses transformer blocks [20]. As with most of the described approaches, the MossFormer model is also designed based on the encoder-decoder structure to process the pseudo-spectrogram calculated by the convolution operator. The separation part of the model also generates the masks of speakers to apply them onto the encoder output. This part is designed in the same manner as Conv-TasNet or SuDoRM-RF, as it uses a sequence of blocks. The MossFormer model uses MossFormer type of blocks of Gated Single-head Transformer (GSHT) architecture with convolution self-attention connections. The MossFormer outperformed SepFormer with 20.9 db SI-SNR versus 20.4 db. It is important to note that the proposed MossFormer architecture uses 2 times fewer parameters than SepFormer 10.8 million versus 25.7 million and still gets the improvement in the separation task.

Authors of the Separate and Diffuse approach proposed the combination of the pre-trained model with the diffusion model DiffWave to improve the performance of existing models [21]. The DiffWave model is the generative adversarial network (GAN) for audio generation. The proposed architecture processes input signal by a pre-trained SepFormer model to obtain primary separation signals. Obtained results later are converted into mel-spectrum and processes each signal by DiffWave model. In this pipeline, the DiffWave model is used as a vocoder that suppresses the non-speech audio components that help to remove noises. However, signal processes do not have phases as mel-spectrum signal representation and require phase correction. This phase correction is fixed by another neural network model that processes input mel-spectrum signal and primary separated signal by SepFormer. This approach produces improvement of the original SepFormer model 1.6 db in SI-SNR. However, then number of parameters of the overall approach is much bigger as contains three different neural network models.

The authors of the MossFormer model later proposed an extended version named MossFormer2 that uses an additional recurrent module combined with the original MossFormer [22] They have a hypothesis that signal processes by transformer block with self-attention connections can have recurrent patterns in it that additionally can be processes that possibly will lead to increasing the performance of the speech separation. As a result, the proposed second version has an increasing 0.2 db in SI-SNR than the Separate and Diffuse approach. However, the size of the MossFormer2 model has increased by 13.6 million parameters than the first version. This increase in parameters is 25% of the original MossFormer model, while the SI-SNR increase is 5%.

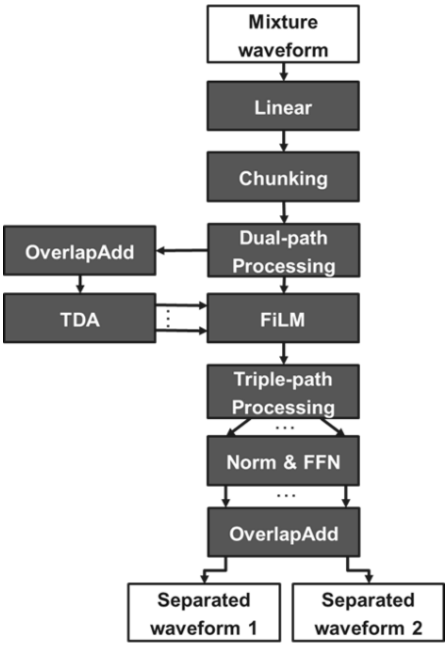

Fig. 6. SepTDA model architecture

Another deep neural network model called SepTDA (Speaker Separation with transformer decoder-based attractor) also inherited the encoder-decoder structure with a separator part [23]. The separator part of the model uses a dual-path processing block, a transformer decoder-based attractor, and a third-path processing block (Fig. 6).

The dual-path block is implemented with a self-attention LSTM block for intra-dependencies and another self-attention LSTM block for inter-dependencies of the audio signal. These inter- and intra-dependencies in the dual-path block are quite similar to the SepFormer approach. Third-path blocks have a similar structure, but as we generate several masks for each speaker, another type of dependency between the speaker's channel. The SepTDA model demonstrates the highest separation level with SI-SNR 24 db with a relatively small number of parameters of 21.2 million. This number of parameters is smaller than MossFormer, SepFormer and Wavesplit.

VI.CONCLUSION

The most suitable architecture for designing a neural network for speech separation is an encoder-decoder-based structure, with a separator block. A neural network with a separator part that aimed to estimate the mask of the separated speaker demonstrated the highest performance. Such masks should be applied to the encoder output to obtain the pseudo-spectrogram of each speaker separately. Generally, approaches with transformer blocks in the architecture produced the highest performance rate, but usually require significantly more memory, than fully-convolutional neural networks. The SepTDA transformer-based neural networks produced the highest separation rate with 24 db in SI-SNR with the relatively small size of the network of 21.2 million of parameters.

On the other hand, the SuDoRM-RF neural network model was the optimal choice in case of a number of parameters, optimization factors, and separation performance. The SuDoRM-RF model was a fully convolutional model that possibly to be boosted by the usage of GPU or NPU. This model had the highest SI-SNR value of 17.02 db for the WSJ0-2mix dataset compared to other fully convolutional neural networks and requires only 2.66 million trainable parameters.

Designing of complex speech separation system that contained several neural networks provided a small improvement in SI-SNR compared to systems with a single neural network, but number of trainable parameters and calculation time made them inconvenient to use in real-time ASR applications.

References

[1] M. Lichouri, K. Lounnas, R. Djeradi & A. Djeradi. (2022). Performance of End-to-End vs Pipeline Spoken Language Understanding Models on Multilingual Synthetic Voice. In *2022 International Conference on Advanced Aspects of Software Engineering* (pp. 1-6). ICAASE. DOI: https://doi.org/10.1109/icaase56196.2022.9931594

[2] Z. -Q. Wang, J. L. Roux & J. R. Hershey. (2018). Alternative Objective Functions for Deep Clustering. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 686-690). ICASSP. DOI: https://doi.org/10.1109/icassp.2018.8462507

[3] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan & P. Smaragdis. (2020). Two-Step Sound Source Separation: Training On Learned Latent Targets. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 31-35). ICASSP. DOI: https://doi.org/10.1109/icassp40776.2020.9054172

[4] Y. Luo, Z. Chen & T. Yoshioka. (2020). 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* (pp. 46-50). ICASSP. DOI: https://doi.org/10.1109/icassp40776.2020.9054266

[5] J. Q. Yip *et al.* (2024). SPGM: Prioritizing Local Features for Enhanced Speech Separation Performance. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 326-330). ICASSP. DOI: https://doi.org/10.1109/icassp48485.2024.10447030

[6] Pu H, Cai C, Hu M, Deng T, Zheng R, Luo J. (2021). Towards Robust Multiple Blind Source Localization Using Source Separation and Beamforming. In *Sensors*. DOI: https://doi.org/10.3390/s21020532

[7] D. Wang & J. Chen. (2018, October). Supervised Speech Separation Based on Deep Learning: An Overview. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (vol. 26, no. 10, pp. 1702-1726). IEEE. DOI: https://doi.org/10.1109/taslp.2018.2842159

[8] Y. Luo & N. Mesgarani. (2018). TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation. *In IEEE International Conference on Acoustics, Speech and Signal Processing (*pp. 696-700). ICASSP. DOI: 10.1109/ICASSP.2018.8462116

[9] Tzinis, E., Wang, Z., Jiang, X. *et al.* (2021). Compute and Memory Efficient Universal Sound Source Separation. In *Journal of Signal Processing Systems 94* (pp. 245–259). DOI: https://doi.org/10.1007/s11265-021-01683-x

[10] J. R. Hershey, Z. Chen, J. Le Roux & S. Watanabe. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (*pp. 31-35). ICASSP. DOI: https://doi.org/10.1109/icassp.2016.7471631

[11] J. L. Roux, S. Wisdom, H. Erdogan & J. R. Hershey. (2019). SDR – Half-baked or Well Done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (*pp. 626-630). ICASSP. DOI: https://doi.org/10.1109/icassp.2019.8683855

[12] M. Kolbæk, D. Yu, Z.-H. Tan & J. Jensen. (2017). Multitalker Speech Separation with Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (vol. 25, no. 10, pp. 1901-1913, Oct. 2017). DOI: https://doi.org/10.1109/taslp.2017.2726762

[13] Cosentino, Joris et al. (2020). LibriMix: An Open-Source Dataset for Generalizable Speech Separation. In *arXiv: Audio and Speech Processing*. DOI: https://doi.org/10.48550/arxiv.2005.11262

[14] Dauphin, Yann et al. (2016). Language Modeling with Gated Convolutional Networks. In *International Conference on Machine Learning*. DOI: https://doi.org/10.48550/arxiv.1612.08083

[15] Y. Luo & N. Mesgarani. (2019, August). Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (vol. 27, no. 8, pp. 1256-1266). DOI: https://doi.org/10.1109/taslp.2019.2915167

[16] E. Tzinis, Z. Wang & P. Smaragdis. (2020). Sudo RM -RF: Efficient Networks for Universal Audio Source Separation. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (*pp. 1-6). MLSP. DOI: https://doi.org/10.1109/mlsp49062.2020.9231900

[17] Y. Liu & D. Wang. (2019, December). Divide and Conquer: A Deep CASA Approach to Talker-Independent Monaural Speaker Separation. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (vol. 27, no. 12, pp. 2092-2102). DOI: https://doi.org/10.1109/taslp.2019.2941148

[18] N. Zeghidour & D. Grangier. (2021). Wavesplit: End-to-End Speech Separation by Speaker Clustering. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (vol. 29, pp. 2840-2849). DOI: https://doi.org/10.1109/taslp.2021.3099291

[19] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi & J. Zhong. (2021). Attention Is All You Need In Speech Separation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 21-25). ICASSP. DOI: https://doi.org/10.1109/icassp39728.2021.9413901

[20] S. Zhao & B. Ma. (2023). MossFormer: Pushing the Performance Limit of Monaural Speech Separation Using Gated Single-Head Transformer with Convolution-Augmented Joint Self-Attentions. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1-5). ICASSP. DOI: https://doi.org/10.1109/icassp49357.2023.10096646

[21] Lutati, Shahar et al. (2023). Separate and Diffuse: Using a Pretrained Diffusion Model for Improving Source Separation. In *ArXiv* abs/2301.10752. DOI: https://doi.org/10.48550/arxiv.2301.10752

[22] S. Zhao et al. (2024). MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 10356-10360). ICASSP. DOI: https://doi.org/10.1109/ICASSP48485.2024.10445985

[23] Y. Lee, S. Choi, B. -Y. Kim, Z. -Q. Wang & S. Watanabe. (2024). Boosting Unknown-Number Speaker Separation with Transformer Decoder-Based Attractor. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 446-450). ICASSP. DOI: https://doi.org/10.1109/icassp48485.2024.10446032

Andrii Tsemko was born in Melitopol, Ukraine, in 2000. Starting from 2023, he has been studying for a PhD in Computer Science at the Faculty of Electronics and Computer Technologies of Ivan Franko National University of Lviv. His research interests encompass machine learning, signal processing, embedded systems, and wireless communication technologies such as Wi-Fi and Bluetooth Low Energy (BLE).

Ivan Karbovnyk, PhD, Dr. Sci., was born in Lviv, Ukraine, in 1978, is the Chair of Radiophysics and Computer Technologies Department at Ivan Franko National University of Lviv. With over 20 years of research experience, he specializes in automation, embedded systems, Internet of Things, computer modeling and electronics.