Electronics and information technologies, 2026, 34, 217–224
http://publications.lnu.edu.ua/collections/index.php/electronics/index

UDC: 004.8

SPEECH SEPARATION BY MODIFIED DEEP NON-LINEAR FILTERING MODEL

Andrii Tsemko**, Ivan Karbovnyk
Faculty of Electronics and Computer Technologies
Department of Radiophysics and Computer Technologies
Ivan Franko National University of Lviv,
107 Gen. Tarnavskoho St., Lviv, 79017, Ukraine
** Corresponding author e-mail: andrii.tsemko@lnu.edu.ua*

Tsemko A., & Karbovnyk I. (2026). Speech Separation by Modified Deep Non-Linear Filtering Model. *Electronics and Information Technologies*, 34, 217–224. https://doi.org/10.30970/eli.34.14

ABSTRACT

Background. Automatic Speech Recognition (ASR) systems in single-channel experimental setups perform poorly in real-world scenarios when several talkers are placed near a microphone. While for multi-channel setups, speech (or source) separation can be solved by algorithmic approaches such as beamforming or machine learning approaches using deep non-linear filtering models, there is no such monaural algorithm that can efficiently separate speech.

Materials and Methods. We focus on a successive model designed for multi-channel speech extraction with a steering mechanism called deep non-linear filtering and modify it for single-channel speech separation. This LSTM-based (LSTM - Long short-term memory) model without a steering mechanism demonstrates high-quality speech extraction from a fixed angle, which makes the model promising for single-channel separation. We modified such a model by appending a parallel output head that generates a second gain mask in addition to the existing one. We train the model on the Libri2Mix dataset in noisy conditions to separate speeches and extract them without background noise. Additionally, we investigate the influence of bidirectional LSTM layers versus the unidirectional version. As the DNLF (Deep non-linear filtering) model is built with bidirectional LSTM layers, it cannot be used for real-time separation processing.

Results and Discussion. Our trained DNLF-SS-UNI (Deep non-linear filtering speech separation unidirectional) model with unidirectional LSTM layers demonstrates 3.41 dB of SI-SDR (Scale-invariant signal-to-distortion ratio) and 5.34 dB of SI-SDRi (Scale-invariant signal-to-distortion ratio improvement) on test subset of Libri2Mix dataset with background noise, while its bidirectional version called DNLF-SS-BI (Deep non-linear filtering speech separation bidirectional) achieves 5.82 dB and 7.76 dB for SI-SDR and SI-SDRi, respectively. However, the bidirectional version of the model requires approximately 2.3 times more parameters than its unidirectional counterpart.

Conclusion. The use of bidirectional LSTM layers in speech separation models of architectures such as deep non-linear filtering is crucial, as a model with unidirectional LSTM layers instead shows a performance drop of approximately 2.4 dB for SI-SDR and SI-SDRi metrics.

Keywords: recurrent neural networks, speech separation, machine learning

BACKGROUND

The main aim of this research is to investigate the deep non-linear filtering model [1] for a single-channel speech separation task. Originally, this model was built for speech extraction and beamforming tasks in a multi-channel microphone setup. The model demonstrated significant results for speech extraction from a specified relative angle by additional input processed by a steering mechanism in a condition with 4 speakers talking simultaneously. It is designed to process an input of stacked matrices of calculated spectrograms from microphones and generate complex gain masks applied to a reference channel. The training framework processes the input complex stacked spectrograms of multi-talker audio scenarios to extract only a single specified person. In one version of the model, the desired speaker to extract is placed at a specified relative or fixed angle to the microphone setup. The trained model demonstrates high-quality separation/isolation of the target speaker, making it highly useful for real-world scenarios.

We gained insight from this relatively small model and decided to experiment with its architecture for a speech separation task focusing on the single-channel setup. While it is possible to add output head to the existing DNLF model and train it for speech separation, a multi-channel setup usually does not require machine learning approaches; since talkers are mostly placed at different angles to the microphone array, they can be efficiently separated by beamforming algorithms or even by the existing DNLF model applied iteratively. It is possible to run the DNLF model several times (many times as needed to extract all speakers) by specifying each time required angle of each speaker time, thereby allowing the existing model to work for a speech separation task. Therefore, we decided to focus on updating the model structure for a single-channel setup, modifying it to process only one complex spectrogram (real and imaginary parts) while generating two complex gain masks for each speaker.

Designing efficient neural network models for speech separation in a single-channel setup requires capturing and processing both long- and short-term audio dependencies. While classical convolutional neural networks (CNNs) demonstrate high performance in various pattern recognition tasks like image processing, their fundamental characteristic of focusing on local features makes them less effective for tasks such as speech separation.

As an efficient speech separation convolution-based model, Conv-TasNet [2] demonstrates a high-quality separation level. Instead of classical convolutional neural network layers, this model relies on temporal convolutional network (TCN) layers [3], which are designed to process input data with an increasing dilation factor. This allows the convolutional model to process both long- and short-term audio dependencies. While our investigation focuses on complex data representation and relies on the Short-Time Fourier Transform (STFT) for the feature preparation phase—and the inverse STFT (iSTFT) to convert the signal from the frequency domain back into the time domain—Conv-TasNet implements a fully trainable time-domain model using trainable encoder-decoder layers. These encoder-decoder layers use single Convolutional and Transposed Convolutional layers to convert the time-domain signal into a pseudo-spectrogram format and vice versa. These pseudo-spectrograms can be described as magnitude spectrograms of the input signal, as they consist of only a single matrix with a fixed number of frequency bins. The primary advantage of the Conv-TasNet model is that it is a fully convolutional model designed for time-domain processing which, when using causal convolutions, can be deployed for real-time speech separation.

Another fully-convolutional model that demonstrates even higher efficiency than Conv-TasNet is a model called SuDoRM-RF [4]. This model is designed in a similar way, using trainable encoder-decoder layers, but relies on U-Net-style convolutional blocks. These U-Net blocks utilize downsampling and upsampling operations with skip connections, which help the model capture both long- and short-term dependencies. This architecture results in a smaller model size with fewer floating-point operations while still yielding highly

competitive separation results in both clean and noisy conditions. Specifically, the SuDoRM-RF model uses only 2.6M parameters—which is roughly half that of Conv-TasNet (5.1M parameters)—and requires only 7.6 GFLOPs compared to Conv-TasNet's 10 GFLOPs. Despite this smaller footprint, it still demonstrates a high separation level, achieving 9.24 dB SI-SDRi compared to the 9.52 dB SI-SDRi of Conv-TasNet [5] for a clean speech mixture without background noise. Ultimately, our investigation focuses on real-world scenarios for speech separation in these types of noisy conditions.

The Deep Non-Linear Filtering model relies on recurrent neural network layers—specifically Long Short-Term Memory (LSTM) layers, which were originally designed to capture both long- and short-term dependencies. Another core difference is its use of an STFT signal representation instead of a trainable encoder layer. While a trainable encoder-decoder layer can potentially be trained to denoise a signal from non-speech components, the STFT provides a complex representation that supplies the model with both magnitude and phase components. This representation helps the model more clearly and precisely process the input information to reconstruct cleanly separated audio signals. Another important architectural feature is the processing of input matrices in both the wideband and narrowband dimensions [6]. This technique helps the model capture feature dependencies across both frequencies and time frames. Combined, these features allow the model to distinguish one speaker from another by focusing not just on isolated characteristics like pitch or talking speed, but on a combination of both.

MATERIALS AND METHODS

A real-world scenario for a speech separation task can be described by next equations:

$$Y(t) = \sum_{i}^{N} x_i(t) + n(t), \tag{1}$$

where $Y(t)$ is a time-domain representation of the mixture of N talkers defined as $x_i(t)$ where i is the index of a speaker and $n(t)$ is an additional background noise. In our particular case, we consider only a two-speaker scenario where N equals 2. The main task for our model is to process the input mixture and reconstruct $x_i(t)$ for each speaker. Our investigated model is designed to process input complex spectrograms as a tensor Y of shape [T, F, 2C], where T is a number of time frames, F number of frequency bins of STFT, and thirds dimension is equals to C by 2, where C is number of channels/microphones and 2 for both real and imaginary parts of spectrogram. In our particular case, C is equal to 1, so the input matrixes shape is [T, F, 2]. However, as original model was built for a single complex gain mask estimation of shape [T, F, 2], our output shape is [T, F, 2*N], so for our case we have an output shape of [T, F, 4], where first estimated gain mask Y_E_1 contains real and imaginary parts at third-dimension's index 0 and 1, and second estimated gain mask Y_E_2 contains real and imaginary parts at index 2 and 3, respectively.

$$Y_{re} = \mathrm{Re}[STFT(y)], \tag{2}$$

$$Y_{im} = \mathrm{Im}[STFT(y)], \tag{3}$$

The model architecture of the deep non-linear filtering model consists of 2 stacked LSTM layers, followed by a fully-connected layer with a tanh activation function. The overall model architecture is illustrated in **Figure 1**. The first LSTM layer, referred to as the

frequency-based LSTM (F-LSTM), processes the input complex spectrogram of a shape [T, F, 2] across the first dimension, so it obtains dependencies across frequencies (inter-frequency dependencies). To capture information across time frames, the resulting output is then permuted into a [F, T, N_F1] format, where N_F1 is the number of hidden features from the first LSTM layer. Due to this permutation, the second LSTM layer features dependencies (temporal dependencies) across time frames rather than frequencies. Through this dual-path approach, the model processes features along both the frequency and time dimensions. The output is then permuted back to a shape of [T, F, N_F2], where N_F2 is the number of features from the second LSTM layer. Next, the resulting matrices must be mapped back into real and imaginary components for our two target masks. To achieve this, we employ a fully-connected layer that processes each [T, F] slice from the [T, F, N_F2] tensor independently, mapping the N_F2 values to 4 output channels to form a vector of values: [$mask1_{real}$, $mask1_{imag}$, $mask2_{real}$, $mask2_{imag}$]. The obtained tensor of shape [T, F, 4] is then processed by a tanh activation function. Finally, the resulting complex masks are separated and applied to the input [T, F, 2] spectrogram independently to yield the individual spectrograms for each speaker, which are subsequently converted back into the time domain via the inverse STFT (iSTFT).

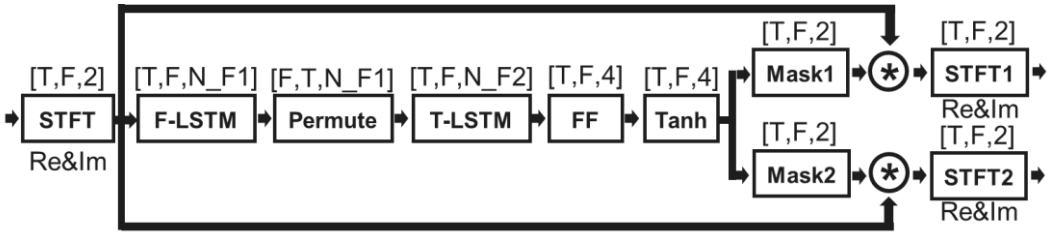

Fig. 1. DNLF-SS model architecture.

Since the original deep non-linear filtering model uses only bidirectional LSTM layers, it cannot be deployed in real-time applications. To address this, we decided to train two versions of the DNLF-SS model: one utilizing the original bidirectional LSTM layers and a second using unidirectional LSTM layers. A bidirectional LSTM layer use pair of weights and processes the input signal from the past-to-future (forward) and from the future-to-past (backward). While this allows the model to leverage future context, it prevents real-time execution because the model must wait for future frames that are unavailable in real-world streaming scenarios. In contrast, a unidirectional LSTM layer processes the sequence only in the forward direction. This makes real-time deployment possible, as the model can process the current time frame using only the current input and the hidden states from past frames [7]. Furthermore, switching to unidirectional LSTM layers reduces the parameter size of the recurrent block by half.

We utilized the open-source Libri2Mix dataset [8] with additional noise for training and evaluation. The dataset consists of mixtures of two speakers with background noise, where the Signal-to-Noise Ratio (SNR) ranges from -6 dB to +3 dB between the loudest speaker and the background noise. The audio data was segmented into 3-second frames at a 16 kHz sampling rate. We use the train-360 dataset subset for training and the test subset for evaluation. The network was trained for 100 epochs with 2 samples per batch. The Adam optimizer was used for training with an initial learning rate of 10^{-3}, which is reduced by a factor of 1.1 every 30 epochs.

As the loss function for training, we use the negative scale-invariant signal-to-distortion ratio (SI-SDR) [9]. The SI-SDR is calculated directly in the time domain rather than the frequency domain by comparing the target time-domain speech signals with the estimated ones. Because speech separation training focuses on obtaining outputs without a

predefined speaker order, we apply an utterance-level permutation-invariant training (uPIT) technique [10].

$$\alpha = \frac{\hat{x}^T s}{\|x\|^2}, \tag{4}$$

$$\text{SI-SDR}(x, \hat{x}) = 10 \log_{10} \left(\frac{\|\alpha x\|^2}{\|\alpha x - \hat{x}\|^2} \right), \tag{5}$$

where x is a target signal, and $\hat{x}$ is estimated speech.

$$L(X, \hat{X}) = -\max_{\pi \epsilon \Pi_N} \left[\frac{1}{N} \sum_{i=1}^{N} \text{SI-SDR}(x_i, \hat{x}_{\pi(i)}) \right], \tag{6}$$

where X is the list of N target sources, $\hat{X}$ is the list of N estimated sources, and the $\max[\,]$ operation ensures the optimal assignment.

In addition to SI-SDR, we also employ the SI-SDR improvement (SI-SDRi) metric for evaluation. This metric estimates the improvement in the separation level achieved by the model compared to a baseline. As a baseline, we calculate the initial SI-SDR metric for the original mixed signal

$$\text{SI-SDRi}(x, \hat{x}, mix) = \text{SI-SDR}(x, \hat{x}) - \text{SI-SDR}(mix, \hat{x}). \tag{7}$$

RESULTS AND DISCUSSION

We trained two LSTM-based models for speech separation in noisy conditions: DNLF-SS-UNI, which utilizes unidirectional LSTM layers, and DNLF-SS-BI, which utilizes bidirectional LSTM layers. The comparative results of our proposed models against the Conv-TasNet baseline are presented in Table 1.

Table 1. **Comparison for Libri2Mix test**

Models	SI-SDR (dB)	SI-SDRi (dB)	Params. (M)	FLOPs (G)
Conv-TasNet	4.48	6.41	5.1	9.992
DNLF-SS-UNI	3.41	5.34	1.4	0.013
DNLF-SS-BI	5.82	7.76	3.3	0.026

Our DNLF-SS-BI model demonstrates an improvement of 2.4 dB over the DNLF-SS-UNI variant across both the SI-SDR and SI-SDRi metrics. All results listed in the table were evaluated on the test subset of the Libri2Mix dataset under noisy background conditions. Although the bidirectional version of the model requires roughly 2.3 times more parameters than its unidirectional counterpart, it still maintains 1.8M fewer parameters than Conv-TasNet. Furthermore, both of our proposed models require fewer than 250M FLOPs to process 1 second of audio, whereas the fully-convolutional Conv-TasNet model requires approximately 10 GFLOPs.

CONCLUSION

In this study, we investigated a deep non-linear filtering model modified for single-channel speech separation. Our best model achieves a performance of 5.82 dB in the SI-SDR metric and 7.76 dB in SI-SDRi, outperforming the Conv-TasNet model by 1.34 dB across both metrics. Additionally, we demonstrate that the backward temporal direction is crucial for the separation performance of LSTM-based models. Utilizing unidirectional layers in the deep non-linear filtering model resulted in a performance drop of over 2.4 dB for each metric, while reducing the model size by only 1.9M parameters. Although bidirectional LSTM layers yield higher separation results and better quality in the reconstructed audio, their architecture prevents them from being used in online separation tasks because they require buffering at least 1 second of audio. In contrast, both Conv-TasNet and our proposed unidirectional model, DNLF-SS-UNI, support efficient real-time online processing.

ACKNOWLEDGMENTS AND FUNDING SOURCES

The authors received no financial support for the research, writing, and/or publication of this article.

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that the research was conducted in the absence of any conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, [A.T., I.K.]; methodology, [A.T., I.K.]; validation, [I.K.]; formal analysis, [A.T., I.K.]; investigation, [A.T., I.K.]; writing – original draft preparation, [A.T.]; writing – review and editing, [I.K.]; visualization, [A.T.]; supervision, [I.K.].

All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Tesch, K., & Gerkmann, T. (2023). Insights into Deep Non-linear Filters for Improved Multi-channel Speech Enhancement. *IEEE/ACM Transactions of Audio, Speech and Language Processing, vol 31.,* 563-575. https://doi.org/10.1109/TASLP.2022.3221046

[2] Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio, Speech, Language Process, 27(8),* 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167.

[3] Lea, C., Vidal, R., Reiter, A., & Hager, G. D. (2016). Temporal Convolutional Networks: A Unified Approach to Action Segmentation. https://doi.org/10.48550/arXiv.1608.08242

[4] Efthymios., T., Zhepei., W., Paris., S. (2020). Sudo rm -rf: Efficient Networks for Universal Audio Source Separation. *IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. https://doi.org/10.1109/MLSP49062.2020.9231900

[5] Tsemko, A., Karbovnyk, I. (2025). Edge-ready speech separation with Sudo-TasNet. *Advances in cyber-physical systems,* Vol. 10, No. 2. https://doi.org/10.23939/acps2025.02.202

[6] Chenda, L., Zhuo, C., Yi, C., Cong, H., Tianya, Z., Keisuke, K., Marc, D., Shinji, W., Yanmin, Q. (2021). Dual-Path Modeling for Long Recording Speech Separation in Meetings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30,* 1508-1520. https://doi.org/10.1109/TASLP.2022.3165712

[7] Tsemko, A., Karbovnyk, I. (2024). Models and Methods for speech separation in digital systems. *Advances in cyber-physical systems*, Vol. 9, No. 2. https://doi.org/10.23939/acps2024.02.121
[8] Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., & Vincent, E. (2020). LibriMix: An open-source dataset for generalizable speech separation. *ArXiv preprint.* arXiv:2005.11262. https://doi.org/10.48550/arXiv.2005.11262.
[9] Le Roux, J., Wisdom, S., Erdoğan, H., & Hershey, J. R. (2018). SDR – half-baked or well done? *ArXiv preprint*. arXiv:1811.02508. https://doi.org/10.48550/arXiv.1811.02508.
[10] Morten, K., Dong, Y., Zheng-Hua, T., & Jesper, J. (2017). Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks. *IEEE/ACM Trans. Audio, Speech and Lang. Proc. 25(10),* 1901-1913. https://doi.org/10.1109/TASLP.2017.2726762.

РОЗДІЛЕННЯ МОВЛЕННЯ З ВИКОРИСТАННЯМ МОДИФІКОВАНОЇ ГЛИБИННОЇ МОДЕЛІ НЕЛІНІЙНОЇ ФІЛЬТРАЦІЇ

Андрій Цемко*, Іван Карбовник
Кафедра радіофізики та комп'ютерних технологій,
Львівський національний університет імені Івана Франка
вул. Ген. Тарнавського, 107, 79017 Львів, Україна

АНОТАЦІЯ

Вступ. Ефективність автоматизованих систем розпізнавання мовлення для одного мікрофона суттєво знижується в реальних умовах, коли одночасно говорять декілька людей поруч із пристроєм. Хоча для систем із багатьма мікрофонами задача розділення мовлення або розділення джерел сигналу загалом вирішується за допомогою таких алгоритмів, як адаптивне формування діаграми спрямованості (англ. beamforming), або з використанням нейронних мереж як глибокі моделі нелінійної фільтрації, ці підходи є неефективними для розділення мовлення в системах з одним мікрофоном.

Матеріали та методи. Ми обрали ефективну глибоку модель нелінійної фільтрації, розроблену для багатоканальних систем виділення мовлення зі стеринг-механізмом (керуванням діаграмою спрямованості), і модифікували її для одноканальної системи розділення мовлення. Ця модель в основі побудована з використанням так званих ДКЧП-шарів (ДКЧП – довга короткочасна пам'ять, англ. Long short-term memory, LSTM), однак без застосування поворотного механізму і показує високу ефективність для виділення одного мовця на фіксованому куті відносно лінійки мікрофонів, що наштовхнуло нас на ідею, що дана модель, перебудована для одноканальної системи, може бути також ефективно. Ми додали до цієї моделі додатковий вихід для генерації другої маски для другого мовця на додачу до вже існуючої. Використовуючи набір даних Libri2Mix з додатковим шумом ми навчили дану модель розділенню мовлення без фонових шумів. Також ми дослідили вплив використання саме дво-направлених ДКЧП шарів та односпрямованих ДКЧП шарів, оскільки оригінальна глибинна модель нелінійної фільтрації через використання дво-направлених шарів не може бути застосована для розділення мовлення у реальному часі.

Результати й обговорення. Модель ОН-ГНФ-РМ (ОН-ГНФ-РМ – одно-напрямлена глибока нелінійна фільтрація для розділення мовлення, англ. DNLF-SS-UNI – Deep non-linear filtering speech separation unidirectional) з одно-напрямленими ДКЧП шарами демонструє рівень розділення в 3,41 дБ для метрики MI-BCC

(масштабно-інваріантного відношення сигналу до спотворень, англ. SI-SDR – Scale-invariant signal-to-distortion ratio) та 5,34 дБ для пMI-BCC (покращення масштабно-інваріантне відношення сигналу до спотворень, англ. SI-SDRi – Scale-invariant signal-to-distortion ration improvement) на тестовому наборі даних з Libri2Mix з фоновими шумами, в той час як модель ДН-ГНФ-РМ (ДН-ГНФ-РМ – дно-напрямлена глибока нелінійна фільтрація для розділення мовлення, англ. DNLF-SS-BI - Deep non-linear filtering speech separation bidirectional) з дво-напрямленими ДКЧП шарами показує вище результати в 5,82 дБ та 7,76 дБ для метрик SI-SDR та SI-SDRi відповідно. Однак, дво-направлена версія моделі потребує майже в 2,3 рази більше навчальних параметрів, порівняно з одно-направленою версією.

Висновки. Використання дво-напрямлених ДКЧП шарів для задачі розділення мовлення в архітектурах по типу глибинної моделі нелінійної фільтрації є вирішальним, оскільки модель з однонапрямленими ДКЧП шарами показує зниження ефективності розділення на 2,4 дБ для обох метрик, – як MI-BCC, так і пMI-BCC.

Ключові слова: рекурентні нейронні мережі, розділення мовлення, машинне навчання

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
27 May, 2026	22 June, 2026	26 June, 2026	29 June, 2026