



Deep-Simplex Multichannel Speech Separation

Tzlil Avidan, Bracha Laufer-Goldshtein

School of Electrical and Computer Engineering, Tel Aviv University, Israel

tzlilavidan@mail.tau.ac.il, blaufer@tauex.tau.ac.il

Abstract

Numerous methods exist for sound source separation, leveraging either classical signal processing or deep learning approaches. While deep-learning-based models often outperform conventional methods, they require large training datasets and struggle to generalize to new settings. To address this, we propose *Deep-Simplex*, a deep prior-based method that reconstructs the probability simplex of speaker activity over time. This global activity probability guides the estimation of a local mask per frequency, identifying the dominant speaker in each time-frequency (TF) bin. We then use this mask for both spatial and spectral separation. Experimental results demonstrate that Deep-Simplex outperforms competing baselines in different reverberation conditions.

Index Terms: source separation, simplex, deep prior

1. Introduction

Sound source separation is widely used in audio processing, music production, and speech enhancement [1]. Over the years, various approaches have been developed to tackle this problem, ranging from classical signal processing methods to modern deep learning techniques. Classical sound source separation methods exploit properties of audio signals such as sparsity, continuity, and harmonic structure. Notable approaches include nonnegative matrix factorization (NMF) [2, 3] and independent component analysis (ICA) [4, 5], which do not require training data are typically based on well-understood mathematical models. However, they often require careful parameter tuning or initialization and may exhibit limited performance on complex, real-world audio mixtures.

In recent years, modern deep learning techniques have revolutionized the field of sound source separation, achieving state-of-the-art performance [6], with models based on CNNs [7], RNNs [8], and Transformers [9]. These models learn complex patterns from data but require large datasets of paired mixtures and their corresponding sources under various acoustic conditions. To reduce data requirements, unsupervised methods have been developed to learn directly from mixtures without isolated sources, including single channel approaches [10, 11] approaches, as well as recent multichannel extensions [12, 13, 14, 15]. However, both supervised and unsupervised approaches may lack interpretability compared to classical methods and often cannot generalize well to unseen settings, different from those encountered during training.

In this paper, we introduce Deep-Simplex, a novel approach for multi-microphone sound source separation that integrates a deep prior-based method [16] with simplex-based speech source separation [17, 18]. The key idea is to decompose a temporal correlation matrix, representing relationships between

time frames, into a low-dimensional simplex that captures the global probability of speaker activity over time. A TF mask is then constructed by identifying the dominant speaker in each TF bin, leveraging local relations within each frequency bin while using the global probabilities as soft labels. However, the original simplex separation method assumes the existence of frames dominated by a single speaker, corresponding to the simplex vertices. This may not hold under heavy overlap or reverberation, leading to distorted simplex geometry and poor vertex recovery. To overcome this challenge, we propose a deep neural network (DNN) to learn the global probabilities directly from the correlation matrix, inspired by the deep image prior framework [16], which requires no pretraining. Likewise, our method operates without external training data, optimizing the model solely on the test mixture. The full pipeline is illustrated in Fig. 1. Our approach leverages the flexibility of neural networks while retaining the structure and interpretability of the simplex-based formulations. Our results demonstrate that Deep-Simplex enhances global probability estimation and leads to superior source separation performance compared to baselines.

2. Method

2.1. Preliminaries

Consider a mixture of J sound sources, recorded by M microphones. Assuming a short-time Fourier transform (STFT) representation over $t \in \{1, \dots, T\}$ time-frames and $f \in \{1, \dots, F\}$ frequency bins, the received signal in the m th microphone, is given by

$$X^m(t, f) = \sum_{j=1}^J A_j^m(f) S_j(t, f) = \sum_{j=1}^J H_j^m(f) X_j^1(t, f) \quad (1)$$

where $S_j(t, f)$ is the j th source signal, $A_j^m(f)$ is the acoustic transfer function (ATF) relating the j th source and the m th microphone, $H_j^m(f) = \frac{A_j^m(f)}{A_j^1(f)}$ is the relative transfer function (RTF) between the m th microphone and the first microphone, and $X_j^m(t, f) = A_j^m(f) S_j(t, f)$ is the j -th source signal measured by the m th microphone.

Let $R^m(t, f) = \frac{X^m(t, f)}{X^1(t, f)}$ denote the ratio between the measured signal in the m th microphone w.r.t. the first microphone. Assuming speech sparsity in the STFT domain [17, 18], this ratio corresponds to the RTF of the dominant speaker, i.e., $R^m(t, f) \approx H_{G(t, f)}^m(f)$, where $G(t, f)$ denotes the dominant speaker at the (t, f) -th bin. We assume that $G(t, f)$ has a categorical distribution $G(t, f) \sim \text{Categorical}(p_1(t), p_2(t), \dots, p_J(t))$, with $p_j(t)$ denoting the probability of activity of the j -th speaker at frame t . Identifying $G(t, f)$ is of key importance in solving the source separation

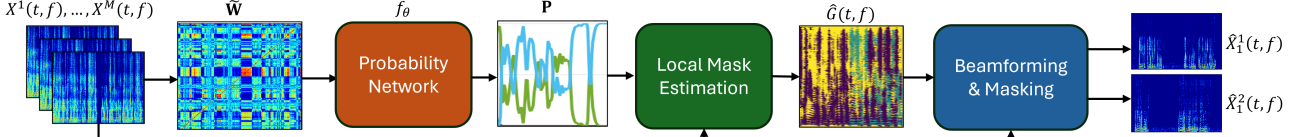


Figure 1: Illustration of the proposed Deep-Simplex method. We first compute microphone ratio values and construct the correlation matrix $\widetilde{\mathbf{W}}$ across time frames. A model is then trained to decompose $\widetilde{\mathbf{W}}$ into $\widehat{\mathbf{P}}$, yielding the global speaker activity probabilities over time. These global probabilities are then used to estimate a local speaker assignment map, which is exploited to separate the mixture.

problem, as it enables both the estimation of a beamformer for spatial separation and the construction of a mask for spectral separation.

The simplex separation method [17, 18] addresses this problem through a two-stage process. First, it estimates the global probabilities of speaker activity over time, $p_j(t)$. Then, it determines the local speaker dominance in individual TF bins, $G(t, f)$. The following sections provide a detailed explanation of both stages.

2.2. Global probabilities

The goal is to estimate the global frame-wise activity probabilities $\mathbf{p}(t) = [p_1(t), \dots, p_J(t)]^T$, which lie in the standard J -dimensional probability simplex. This is achieved through the following key steps: (1) compute per-frame features; (2) construct a correlation matrix across frames; (3) recover a simplex from the eigenvectors of the correlation matrix; (4) identify the simplex vertices corresponding to frames dominated by a single speaker; and (5) transform the eigenvector simplex into a probability simplex using the identified vertices. In our proposed Deep-Simplex method we replace steps (3)-(5) with a DNN that directly learns the global probabilities from the correlation matrix. We now elaborate on each step and our modifications.

We first compute a real-valued feature vector per frame $\mathbf{r}_{\text{glob}}(t) \in \mathbb{R}^{2(M-1)F}$ by aggregating the real and imaginary parts of the microphone ratios $\{R_m(t, f)\}_{m,t,f}$. Based on these features we compute the temporal correlation matrix $\mathbf{W} \in \mathbb{R}^{T \times T}$ with elements given by $W_{tt'} = \frac{1}{F} \mathbb{E}\{\mathbf{r}_{\text{glob}}^\top(t) \mathbf{r}_{\text{glob}}(t')\}$. The core idea of the simplex method lies in the following decomposition of \mathbf{W} in terms of the global probabilities:

$$\mathbf{W} \approx \mathbf{P}\mathbf{P}^\top \quad (2)$$

where \mathbf{P} is a $T \times J$ probability matrix with $P_{tj} = p_j(t)$. Note that there is a slight discrepancy between the diagonal elements of \mathbf{W} which equal 1 (assuming that the feature vectors are normalized to unit-norm) and the diagonal elements of $\mathbf{P}\mathbf{P}^\top$ that equal $\sum_{j=1}^J p_j^2(t)$. As follows by this decomposition, the rank of \mathbf{W} equals that of \mathbf{P} , which corresponds to the number of speakers J . In practice, \mathbf{W} is estimated as $\widetilde{W}_{tt'} \approx \frac{1}{F} \mathbf{r}_{\text{glob}}^\top(t) \mathbf{r}_{\text{glob}}(t')$, using the same statistical assumptions described in [17, 18].

The eigenvalue decomposition of the estimated correlation matrix $\widetilde{\mathbf{W}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ yields an orthonormal matrix \mathbf{U} containing its eigenvectors $\{\mathbf{u}_j\}_{j=1}^T$ and a diagonal matrix $\mathbf{\Lambda}$ containing its eigenvalues $\{\lambda_j\}_{j=1}^T$. We extract and sort the first J eigenvectors, and for each time-frame t define the vectors $\boldsymbol{\nu}(t) = [u_1(t), \dots, u_J(t)]^\top$. These vectors $\boldsymbol{\nu}(t)$ span a rotated and scaled simplex referred to as U-simplex, whose J vertices correspond to single-speaker-dominant frames. To recover the source probabilities per time-frame $\mathbf{p}(t)$, we first identify J vertices $\{\boldsymbol{\nu}(t_j)\}_{j=1}^J$ using the successive projection algorithm (SPA) [19]. Then we construct the back-transformation matrix from the identified vertices $\widehat{\mathbf{Q}} = [\boldsymbol{\nu}(t_1), \dots, \boldsymbol{\nu}(t_J)]^\top$,

and transform $\boldsymbol{\nu}(t)$ back to the standard probability by $\widehat{\mathbf{p}}(t) = \widehat{\mathbf{Q}}^{-1} \boldsymbol{\nu}(t)$.

If the speakers exhibit significant overlap, the extracted simplex may lack well-defined vertices, as illustrated in Fig. 2. SPA relies on the assumption that certain frames are dominated by a single speaker, identifying these frames by detecting the simplex vertices. Specifically, the first vertex is selected as the frame with the maximum norm, while the second is chosen as the one farthest from the first. The remaining vertices are iteratively identified by maximizing their projection onto the orthogonal complement of the subspace spanned by the previously selected vertices. However, if no single-speaker frames exist for each source, SPA fails to accurately recover the true probabilities.

Instead, we propose using a DNN to directly map the input correlation matrix $\widetilde{\mathbf{W}}$ to the corresponding activity probabilities for the given test mixture, without requiring prior training. Specifically, we learn a model f_θ , parameterized by θ , which takes the matrix $\widetilde{\mathbf{W}}$ as input and outputs the global probabilities, i.e. $\widehat{\mathbf{P}} = f_\theta(\widetilde{\mathbf{W}})$. The model parameters θ are optimized in an unsupervised manner by minimizing the discrepancy between $\widetilde{\mathbf{W}}$ and $\widehat{\mathbf{P}}\widehat{\mathbf{P}}^\top$, as implied by Eq. (2). Since \mathbf{W} and $\mathbf{P}\mathbf{P}^\top$ differ in their diagonal elements, we enforce equality by setting these elements to 1:

$$\widehat{W}_{ij} = \begin{cases} [\widehat{\mathbf{P}}\widehat{\mathbf{P}}^\top]_{ij}, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases} \quad (3)$$

We then define the loss function, inspired by [20] and structurally related to deep clustering [21], which matches an embedding similarity matrix to supervised labels per frequency. In contrast, we align our model output to a correlation matrix derived directly from the input, without supervision:

$$\mathcal{L} = \lambda_1 \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_F^2 + \lambda_2 \sum_{t=1}^T \|\widetilde{\mathbf{W}}_t\|_2 \arccos \left(\frac{\widetilde{\mathbf{W}}_t^\top \widehat{\mathbf{W}}_t}{\|\widetilde{\mathbf{W}}_t\|_2 \|\widehat{\mathbf{W}}_t\|_2} \right) \quad (4)$$

where $\widetilde{\mathbf{W}}_t$ denotes the t -th column of $\widetilde{\mathbf{W}}$, $\|\cdot\|_F$ is matrix Frobenius norm, and λ_1, λ_2 are loss weighting coefficients. The first term in Eq. (4) minimizes reconstruction error using element-wise squared loss, while the second term aligns the matrix columns. This promotes a structured probability matrix that preserves the mixture's underlying correlations. Notably, this loss function requires no supervised information and is optimized independently for each test example.

2.3. Model Architecture

Our network for estimating the global speaker probabilities draws an inspiration from architectures used in hyperspectral

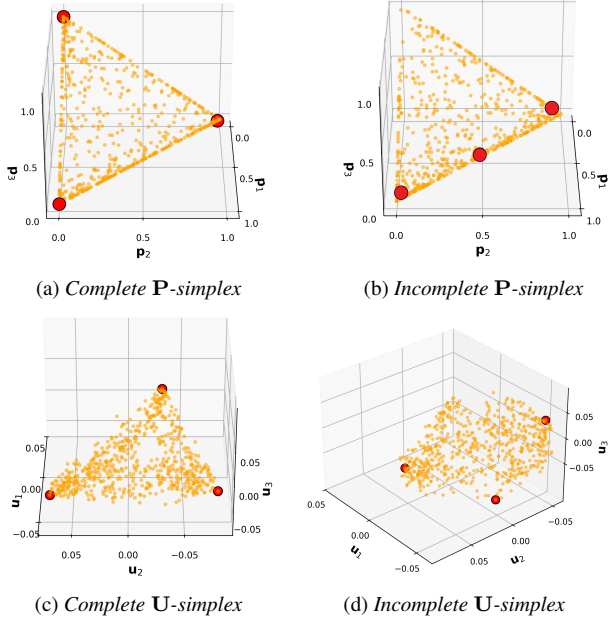


Figure 2: Two cases of the simplex mapping ($J = 3$): (a)+(c) complete and (b)+(d) incomplete. First row shows the true probabilities, and the second row shows the first J eigenvectors. Red points correspond to the vertices detected by SPA.

unmixing [20, 22] and deep clustering [23]. The input correlation matrix is treated as a sequence of T time-frames, each represented by a T -dimensional vector. These are transformed into a sequence of T probability vectors over J speakers, capturing speakers’ activity probability over time. The architecture, shown in Fig. 3, consists of the following components:

1. **Multi-Head Attention:** Captures long-range temporal dependencies via self-attention applied to the $T \times T$ input, refining the representation of each frame by incorporating global temporal context.
2. **Two Bidirectional LSTM (BLSTM) Layers:** Bidirectional LSTMs encode temporal context in both forward and backward directions, producing an output of size $T \times H$.
3. **Conv1D Encoder Block:** Applies 1D convolutions along the time axis, treating T as the sequence dimension and H as channels. Each layer uses LayerNorm and LeakyReLU for stability and nonlinearity. The feature dimension is gradually reduced, and two skip connections are included to preserve information flow and mitigate vanishing gradients.
4. **Fully Connected Layer:** A dense layer enables interaction across all channel dimensions, transforming the encoded features into a $T \times J$ matrix of frame-wise speaker probabilities.
5. **Softmax Layer:** Enforces valid probabilities by normalizing each row (frame) to sum to one.

2.4. Local Mask

We exploit the estimated global probabilities to recover the dominant source in each frequency bin. We define the local per-frequency ratio vector $\mathbf{r}_{\text{loc}}(t, f) \in \mathbb{R}^{2M}$ aggregating real and imaginary ratio values $\{R_m(t, f)\}_m$ across microphones.

The index of the dominant component in each TF bin is chosen by combining the relations between the local mappings with the global probabilities $\mathbf{p}(t)$. For each frame, the assignment is determined based on the following weighted nearest-

neighbor rule:

$$\hat{G}(t, f) = \arg \max_{j \in 1, \dots, J} \frac{1}{\pi_j} \sum_{t'=1}^T \omega_{tt'}(f) \cdot p_j(t') \quad (5)$$

where the weight $\omega_{tt'}(f) = \exp\{-\|\mathbf{r}_{\text{loc}}(t, f) - \mathbf{r}_{\text{loc}}(t', f)\|_2^2\}$ measures the similarity between frames t and t' , with closer embeddings $\mathbf{r}_{\text{loc}}(t, f)$ and $\mathbf{r}_{\text{loc}}(t', f)$ receiving higher influence. The normalization term $\pi_j = \sum_{t=1}^T p_j(t)$ ensures proper class weighting. Notably, aligning local decisions with the same global probability $\mathbf{p}(t)$ mitigates permutation ambiguity across frequencies—a common issue in separation methods that perform clustering per frequency [24].

2.5. Separation

We perform separation by combining spatial filtering with spectral masking. For each speaker, we compute the spatial covariance matrix at each frequency using TF bins where $\hat{G}(t, f) = j$, and estimate the corresponding RTF as the principal eigenvector of this matrix, normalized with respect to the first microphone [1]. The RTFs are then used to construct a linearly constrained minimum variance (LCMV) beamformer per speaker (frequency index omitted): $\mathbf{B}_j = \mathbf{H}(\mathbf{H}^H \mathbf{H})^{-1} \mathbf{q}_j$, where the j th column of \mathbf{H} is the estimated RTF of the j th speaker and \mathbf{q}_j is a selection vector with 1 at the j th entry. As a post-processing step, we apply the estimated mask:

$$\hat{X}_j(t, f) = \mathbf{1}_{\hat{G}(t, f)=j} \cdot \hat{X}_j^{\text{B}}(t, f) + \gamma \cdot \mathbf{1}_{\hat{G}(t, f) \neq j} \cdot \hat{X}_j^{\text{B}}(t, f), \quad (6)$$

where $\hat{X}_j^{\text{B}}(t, f) = \mathbf{B}_j^H(f) \mathbf{X}(t, f)$ is the beamformer output, $\mathbf{X}(t, f) = [X^1(t, f), \dots, X^M(t, f)]^T$ is the multichannel mixture, and γ is an attenuation factor.

3. Experiments

3.1. Dataset

We simulated $6 \times 6 \times 2.4 \text{ m}^3$ rooms with reverberation times of 300 ms or 600 ms¹. We used a uniform linear array of $M = 4$ microphones, centered at $[3, 3, 1.5] \text{ m}$, with 30 cm distance between adjacent microphones. We generated mixtures of $J = 3$ speakers, with source positions randomly sampled on a half-circle at a 2 m distance from the array center, ensuring a minimum angular separation of 30° between speakers. Speech utterances were drawn from the Librispeech dev-set [25]. For each speaker, 4 – 5 recordings were concatenated to form a continuous 20 s long signals, thus speakers have close to 100% overlap. The signals were processed using an STFT with 1024 frequency bins and 75% frame overlap. The global probability estimation network requires approximately 29 s per 20 s mixture on a single NVIDIA RTX 6000 Ada GPU. Our code is available at².

3.2. Model and Training Settings

The correlation matrix is computed based on ratio vectors in the frequency range of 1000–2000 Hz. The model hyperparameters are detailed in Table 1. We train the model for 200 epochs (one-sample iterations), using the Adam optimizer with a learning rate of $1e-5$ and $\beta_1 = 0.5$, $\beta_2 = 0.99$. The loss function hyperparameters used in Eq. (4) are $\lambda_1 = 1000$ and $\lambda_2 = 1$ to align scales. We use $\gamma = 0.3$ for spectral mask attenuation.

¹<https://pypi.org/project/rir-generator/>

²<https://github.com/tzlilavi/deep-simplex>

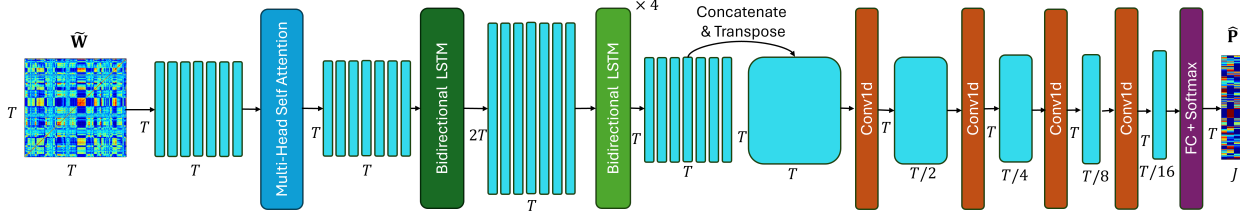
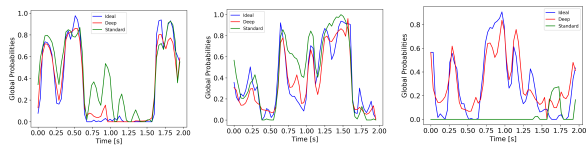


Figure 3: Architecture of the proposed probability network for predicting activity probabilities based on an input correlation matrix.

Table 1: Layerwise Summary of Network Blocks

Layer	Composition	In	Out	Hyperparams
1	Multi-Head Attention	$T \times T$	$T \times T$	$n_{heads} = 8$
2	BiLSTM Layer 1	$T \times T$	$T \times (H_1 \times 2)$	$H_1 = L$
3	BiLSTM Layer 2 ($\times 4$)	$T \times (H_1 \times 2)$	$T \times (H_2 \times 2)$	$H_2 = T/2$
4	Conv1D, LN, LeakyReLU	$T \times T$	$T \times T/2$	$kernel = 3, pad = 1$
5	Conv1D, LN, LeakyReLU	$T \times T/2$	$T \times T/4$	$kernel = 3, pad = 1$
6	Conv1D, LN, LeakyReLU	$T \times T/4$	$T \times T/8$	$kernel = 3, pad = 1$
7	Conv1D, LN, LeakyReLU	$T \times T/8$	$T \times T/16$	$kernel = 3, pad = 1$
8	Skip Connection Conv1D	$T \times T$	$T \times T/4$	$kernel = 3, pad = 1$
9	Skip Connection Conv1D	$T \times T/4$	$T \times T/16$	$kernel = 3, pad = 1$
10	Fully Connected	$T \times T/16$	$T \times J$	-
11	Softmax	$T \times J$	$T \times J$	-



(a) 1st Speaker (b) 2nd Speaker (c) 3rd Speaker

Figure 4: Comparison of global probabilities computed for the example in Fig. 2 using the Ideal baseline, Standard Simplex and Deep-Simplex. The average SI-SDR scores for this example are - Ideal: 8.24 dB, Deep: 6.15 dB, and Standard: -5.42 dB.

3.3. Baselines and Evaluation

We compare Deep-Simplex to the Standard Simplex [18] and a conventional approach based on independent vector analysis (IVA), using the TorchIVA toolkit³. We also evaluate an ideal baseline using the same procedure described in Sec. 2.5, but with an ideal mask computed from the individual source signals.

We first compare Deep-Simplex and the Standard Simplex in terms of global probability prediction and mask estimation. For global probability prediction, we use the mean squared error (MSE) between the true and estimated probabilities. For mask estimation, error is defined as the percentage of TF bins assigned to the wrong speaker. The true probabilities are derived from the ideal mask, based on the proportion of TF bins dominated by each speaker. Additionally, we evaluate the separation performance across all baselines using SI-SDR, PESQ, and STOI metrics. Results are averaged over 30 random mixtures, with both mean scores and standard deviations reported.

3.4. Results

Probability prediction and mask estimation. Table 2 summarizes the errors in global probability prediction and local mask estimation. Our results show that Deep-Simplex outperforms the Standard Simplex in global probability estimation, which in turn enhances local mask estimation. This improvement is evident in both lower mean error values and reduced standard deviation, and it remains consistent for moderate and high reverberation levels. As illustrated in Figs. 2 and 4, when the simplex structure is incomplete, the SPA algorithm fails to recover the vertex of at least one speaker. This failure significantly

³<https://github.com/fakufaku/torchiva>

Table 2: Performance comparison in terms of global probability and local mask estimation.

Method	Global MSE	Mask Err
Reverberation Time - 300 ms		
Standard Simplex	0.026 ± 0.026	0.330 ± 0.103
Deep-Simplex	0.020 ± 0.008	0.317 ± 0.081
Reverberation Time - 600 ms		
Standard Simplex	0.025 ± 0.027	0.339 ± 0.101
Deep-Simplex	0.018 ± 0.007	0.319 ± 0.079

Table 3: Separation Performance in terms of SI-SDR(dB), PESQ, and STOI(%).

Method	SI-SDR (dB)	STOI	PESQ
Reverberation Time - 300ms			
IVA	1.4 ± 4.1	0.622 ± 0.085	1.75 ± 0.297
Standard Simplex	5.3 ± 4.8	0.795 ± 0.095	2.45 ± 0.311
Deep-Simplex	6.4 ± 2.4	0.821 ± 0.035	2.53 ± 0.158
Ideal Mask	8.6 ± 2.0	0.851 ± 0.030	2.68 ± 0.177
Reverberation Time - 600 ms			
IVA	-0.1 ± 3.8	0.592 ± 0.076	1.67 ± 0.224
Standard Simplex	4.8 ± 4.5	0.783 ± 0.089	2.38 ± 0.273
Deep-Simplex	5.9 ± 2.4	0.806 ± 0.038	2.45 ± 0.165
Ideal Mask	7.6 ± 2.1	0.836 ± 0.032	2.58 ± 0.191

degrades the performance leading to inaccurate probability estimations and poor mask reconstructions. In contrast, Deep-Simplex presents robust performance across these challenging cases, offering a more reliable probability prediction.

Separation Performance. Table 3 compares the separation performance across all methods. It can be seen that Deep-Simplex outperforms the baselines at both reverberation levels, achieving higher scores. Moreover, these results confirm that the improvement in global probability estimation contributes to enhanced separation performance with lower variance compared to the Standard Simplex approach.

4. Conclusions

In this paper, we propose a deep learning model for estimating the global probability of speaker activity over time from a single multi-microphone recording. The model decomposes the correlation matrix between time frames into a low-rank approximation that captures speaker activity probabilities within a simplex. These estimated probabilities are then used to derive a local mask for the dominant speaker in each TF bin, enabling source separation. Our results show that the proposed method improves global probability estimation over the standard simplex approach and achieves superior performance compared to competing baselines across reverberation levels.

5. Acknowledgments

This work was supported by the Israeli Ministry of Innovation, Science and Technology (Grant No.1001818518).

6. References

- [1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on Multimicrophone speech enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.
- [2] T. Virtanen, "Monaural sound source separation by Nonnegative Matrix Factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.
- [3] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2009.
- [4] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *2004 IEEE international conference on acoustics, speech, and signal processing*, vol. 3. IEEE, 2004, pp. iii–889.
- [5] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e12, 2019.
- [6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [7] Y. Luo and N. Mesgarani, "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.
- [8] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.
- [9] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 21–25.
- [10] S. Wisdom, H. Erdogan, J. R. Hershey *et al.*, "Unsupervised sound separation using mixture invariant training," in *Proc. NeurIPS*, 2020.
- [11] K. Schulze-Forster, G. Richard, L. Kelley, C. S. Doire, and R. Badeau, "Unsupervised music source separation using differentiable parametric source models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1276–1289, 2023.
- [12] Z.-Q. Wang and S. Watanabe, "UNSSOR: Unsupervised neural speech separation by leveraging over-determined training mixtures," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [13] K. Saijo and R. Scheibler, "Spatial loss for unsupervised multi-channel source separation," in *Proc. Interspeech*, 2022, pp. 166–170.
- [14] K. Han, F. Zhang, and P. Smaragdis, "Unsupervised multi-channel separation and adaptation," in *Proc. ICASSP*, 2024, pp. 1–5.
- [15] Y. Bando, Y. Koizumi, and T. Nakatani, "Neural fast full-rank spatial covariance analysis for blind source separation," in *Proc. EUSIPCO*, 2023, pp. 1–5.
- [16] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9446–9454.
- [17] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source counting and separation based on simplex analysis," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6458–6473, 2018.
- [18] —, "Global and local simplex representations for multichannel source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 914–928, 2020.
- [19] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvao, T. Yoneyama, H. C. Chame, and V. Visani, "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis," *Chemometrics and intelligent laboratory systems*, vol. 57, no. 2, pp. 65–73, 2001.
- [20] P. Ghosh, S. K. Roy, B. Koirala, B. Rasti, and P. Scheunders, "Hyperspectral unmixing using transformer network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 516–527, 2022.
- [21] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.
- [22] B. Rasti, B. Koirala, P. Scheunders, and J. Chanussot, "MiSiCNet: Minimum simplex convolutional network for deep hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022.
- [23] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.
- [24] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2010.
- [25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.