Deep-Simplex Multichannel Speech Separation

Anonymous submission to Interspeech 2025

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, aiming to extract the individual sounds from a recorded mixture of multiple sources, has numerous applications 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 approaches to sound source separation typically exploit the inherent properties of audio signals, such as sparseness, continuity, and harmonic structure. One notable classical approach is nonnegative matrix factorization (NMF), which decomposes the magnitude spectrogram of an audio mixture into a set of basis functions and their corresponding activations [2, 3]. Another important line of research is based on independent component analysis (ICA), which assumes statistical independence between the source signals [4, 5]. The advantage of these methods is that they do not require large datasets for training. In addition, they are often based on well-understood mathematical principles, and thus can be more interpretable than deep learning models. However, they often require careful parameter tuning or initialization and may struggle with complex, real-world audio mixtures.

In recent years, deep learning has revolutionized the field of sound source separation, achieving state-of-the-art performance on various settings [6]. Recent advancements in deep learning for source separation include the use of convolutional neural networks (CNNs) [7], recurrent neural networks (RNNs) [8], and transformer architectures [9]. Their strong performance can be attributed to their ability to learn complex patterns and representations from data. Nevertheless, they require large datasets of paired individual speech signals and their corresponding mixture under various acoustic conditions. To reduce data demands, unsupervised methods have been proposed, relying on mixtures (no isolated sources) for learning [10, 11]. 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 with simplex-based speech source separation [12, 13]. The key idea is to decompose a temporal affinity matrix, representing relationships between time frames, into a low-dimensional simplex that captures the 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, when speaker overlap is high, some frames may lack a dominant speaker, leading to a simplex without distinct corners and impairing the accurate recovery of global probabilities. To overcome this challenge, we propose leveraging a deep neural network to learn the global probabilities directly from the affinity matrix, using an unsupervised loss function for training on a single sample. The overall framework is illustrated in Fig. 1. Our results demonstrate that this approach 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. The short-time Fourier transform (STFT) representation of the received signal in the mth 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 jth source signal, $A_j^m(f)$ is the acoustic transfer function (ATF) relating the jth source and th mth microphone, $H_j^m(f) = \frac{A_j^m(f)}{A_1^m(f)}$ 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 [12, 13], this ratio corresponds to an RTF of one of the speakers $R^m(t,f) \approx H^m_{M(t,f)}(f)$, where $M(t,f)$ is the index of the dominant speaker in the (t,f)-th bin.

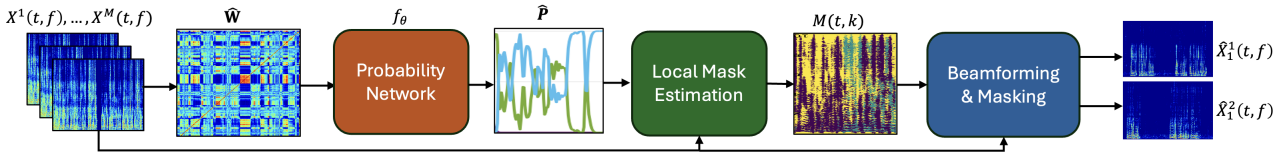


Figure 1: ***Illustration of the proposed Deep-Simplex method.*** *We first compute the ratio values between microphones and construct the correlation matrix* $\mathbf{W}$ *across time frames. A model is then trained to decompose* $\mathbf{W}$ *into* $\mathbf{PP}^T$*, yielding the global speaker activity probabilities over time. Using these global probabilities, we estimate a local mask, which is then exploited to separate the mixture.*

We assume that $M(t,f)$ has a categorical distribution $M(t,f) \sim \text{Categorical}(p_1(t), p_2(t), \ldots, p_J(t))$, with $p_j(t)$ denoting the probability of activity of the j-th speaker at time t. Identifying $M(t,f)$ is tantamount to solving the source separation problem, as each source signal can be extracted by masking, i.e. $\hat{X}_j^1(t,f) = I_j(t,f) \cdot X^1(t,f)$ where $I_j(t,f) = 1$ if $M(t,f) = j$ and 0 otherwise.

The simplex separation method [12, 13] addresses this problem through a two-stage process. First, it estimates the global probabilities of speaker activity over time, $p_j(t)$, for all $t \in \{1,\ldots,T\}$ and $j \in \{1,\ldots,J\}$. Then, in the second stage, it determines the local dominance of each speaker within individual TF bins, $M(t,f)$, for all $t \in \{1,\ldots,T\}$ and $f \in \{1,\ldots,F\}$. The following sections provide a detailed explanation of both steps.

2.2. Global probabilities

The goal is to estimate the frame-wise activity probabilities $\mathbf{p}(t) = [p_1(t),\ldots,p_J(t)]^T$, which lie in the standard J-dimensional probability simplex. We construct a real-valued feature vector $\mathbf{r}_g(f) \in \mathbb{R}^{2(M-1)F}$, aggregating real and imaginary ratio values $\{R_m(t,f)\}_{m,t,f}$ across all frequency bins and microphones. It has been shown that the $T \times T$ correlation matrix $\mathbf{W}$, with elements $W_{tt'} = \frac{1}{F}\mathbb{E}\{\mathbf{r}_g^\top(t)\mathbf{r}_g(t')\}$, can be approximated as:

$$\mathbf{W} = \mathbf{PP}^\top + \Delta\mathbf{W}, \tag{2}$$

where $\mathbf{P}$ is a $T \times J$ probability matrix containing source activity probabilities, i.e., $P_{lj} = p_j(f)$, and $\Delta\mathbf{W}$ is a diagonal matrix with $\Delta W_{tt} = \sum_{j=1}^{J}(1 - p_j(t))p_j(t) \leq 1$. Consequently, 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 $\mathbf{W}_{tt'} \approx \frac{1}{F}\mathbf{r}_g^\top(t)\mathbf{r}_g(t')$, a statistically justified approximation discussed in [12, 13].

The eigenvalue decomposition of the estimated correlation matrix $\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ yields an orthonormal matrix $\mathbf{U}$ containing its eigenvectors $\mathbf{u}_j$ and a diagonal matrix $\mathbf{\Lambda}$ containing its eigenvalues λ_j. Each time frame t can be represented by a slice of the extracted and sorted eigenvectors as $\boldsymbol{\nu}(t) = [u_1(t),\ldots,u_J(t)]^\top$. These vectors $\boldsymbol{\nu}(t)$ span a rotated and scaled simplex, whose J vertices $\{\boldsymbol{\nu}(t_J)\}_{j=1}^J$ can be computed using the successive projection algorithm (SPA) [14]. To finally obtain the source probabilities per time frame $\mathbf{p}(t)$, we construct the back-transformation matrix from the identified vertices $\hat{\mathbf{Q}} = [\boldsymbol{\nu}(t_1),\ldots,\boldsymbol{\nu}(t_J)]^\top$, and transform $\boldsymbol{\nu}(t)\,\forall t$ back to the standard simplex to obtain the source probabilities, i.e. $\hat{\mathbf{p}}(t) = \hat{\mathbf{Q}}^{-1}\boldsymbol{\nu}(t)$.

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 Deep Neural Network (DNN) to directly map the input correlation matrix $\mathbf{W}$ 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 $\mathbf{W}$ as input and outputs the global probabilities, i.e. $\widehat{\mathbf{P}} = f_\theta(\mathbf{W})$. The model parameters θ are optimized in an unsupervised manner by minimizing the discrepancy between $\mathbf{W}$ and $\widehat{\mathbf{P}}\widehat{\mathbf{P}}^\top$, as implied by (2). Since $\mathbf{W}$ and $\mathbf{PP}^\top$ differ in their diagonal elements, we enforce equality by setting these diagonal elements to 1:

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

We then use the following loss function, which is similar to the loss defined in [15], used in the context of hyperspectral unmixinig:

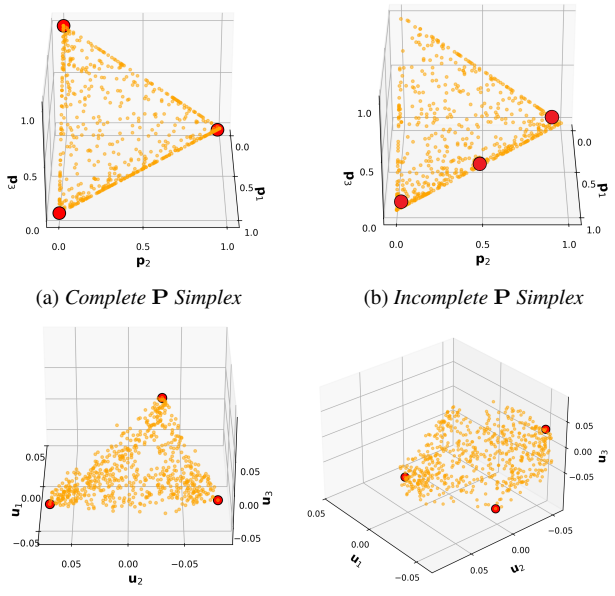
$$\mathcal{L} = \lambda_1 ||\mathbf{W} - \widehat{\mathbf{W}}||_F^2 + \lambda_2 \sum_{t=1}^{T} ||\mathbf{W}_t||_2 \arccos\left(\frac{\mathbf{W}_t^\top \widehat{\mathbf{W}}_t}{||\mathbf{W}_t||_2 ||\widehat{\mathbf{W}}_t||_2}\right) \tag{4}$$

where $\mathbf{W}_t$ denotes the t-th column of $\mathbf{W}$, $\|\cdot\|_F$ is matrix Frobenius norm, and λ_1, λ_2 are loss weighting coefficients. The first term in Eq. (4) is an element-wise squared loss, which minimizes the reconstruction error between $\widehat{\mathbf{P}}$ and $\mathbf{W}$, while the second term aligns the matrix columns. By enforcing this alignment in both angle and scale, the model derives a structured probability matrix that preserves the essential correlations within the mixture. Notably, this loss function requires no supervised information and is optimized independently for each test example.

2.3. Model Architecture

Our global probability network consists of multiple units designed for both local and global processing, drawing inspiration from architectures used in hyperspectral unmixing. The input correlation matrix is interpreted as a sequence of T time frames, each embedded as a T-dimensional vector. Through the network, these vectors are transformed into a sequence of T probabilities of dimension J, capturing the speakers' probabilities over time. The model architecture is illustrated in Figure. 3, and consists of the following elements.

1. **Multi-Head Attention Mechanism:** This block captures long-range dependencies across time frames, refining



(a) *Complete* $\mathbf{P}$ *Simplex* (b) *Incomplete* $\mathbf{P}$ *Simplex*

(c) *Complete* $\mathbf{U}$ *Simplex* (d) *Incomplete* $\mathbf{U}$ *Simplex*

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

the representation of each frame by incorporating global temporal context. It applies self-attention to the given T correlation vectors of size T, generating an output of the same dimensions ($T \times T$).

2. **Two Bidirectional LSTM (BLSTM) Layers [16]:** The bidirectional LSTM layers process the sequence in both forward and backward directions, capturing past and future temporal dependencies. Given an input of shape $T \times T$, each BLSTM layer preserves the sequence length T while transforming the feature dimension, producing an output tensor of shape $T \times H$.
3. **Conv1D Encoder Block [17]:** We treat T as the sequence dimension and H as the channel dimension, applying one-dimensional convolutions along the sequence axis. By using padding, we preserve the sequence length while progressively reducing the channel size. Since the model operates on single samples without batch processing, LayerNorm is applied after each encoding layer instead of BatchNorm. LeakyReLU is used as the activation function. We incorporate two skip connections across layers to improve gradient flow and mitigate the vanishing gradient problem.
4. **Final Fully Connected Layer:** Finally, we have a fully connected (dense) layer, enabling interaction of all channel dimensions. This layer transforms the encoded representation into the required probability simplex matrix of size $T \times J$, where each row represents a probability distribution over J speakers.
5. **Softmax Layer:** The output is passed through a Softmax function applied to each row, ensuring that the final probabilities for each frame sum to one, thus forming a valid probability simplex.

2.4. Local Mask

We exploit the estimated global probabilities to recover the dominant source in each frequency bin. We define the per-frequency ratio vector $\mathbf{r}_l(t,f) \in \mathbb{R}^{2M}$ 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:

$$\hat{M}(t,f) = \underset{j \in 1,\ldots,J+1}{\arg\max} \frac{1}{\pi_j} \sum_{t=1}^{T} \omega_{tt'}(f) \cdot p_j(t') \qquad (5)$$

where the weight $\omega_{tt'}(f) = \exp\{-\|\mathbf{r}_l(t,f) - \mathbf{r}_l(t',f)\|_2^2\}$ measures the similarity between frames t and t', with closer embeddings $\mathbf{r}_l(t,f)$ and $\mathbf{r}_l(t',f)$ 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 [18].

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	$T \times J$	$T \times J$	-

2.5. Separation

We perform separation using the estimated mask by integrating spatial filtering with spectral post-processing. Specifically, we estimate each speaker's RTF based on the TF bins where they are dominant. These estimates are then used to construct a linearly constrained minimum variance (LCMV) beamformer. For post-processing, we refine the beamformer output by applying the estimated mask, attenuating TF bins dominated by other speakers by a factor of β.

3. Experiments

3.1. Dataset

We simulated a $6 \times 6 \times 2.4$ m^3 room with reverberation times of 300ms or 600ms [19]. We used a uniform linear array of $M = 4$ microphones, centered at $[3, 3, 1.5]$ m, with 30cm distance between adjacent microphones. We generated mixtures of $J = 3$ speakers, with source positions randomly sampled on a half-circle with a 2m distance from the array center, ensuring a minimum angular separation of 30° between speakers. Speech utterances were drawn from the Librispeech dev-set [20]. 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.

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 100 epochs (one-sample iterations), using the Adam optimizer with a learning rate of $1e-5$ and $\beta = (0.5, 0.99)$. We use $\beta = 0.3$ for mask

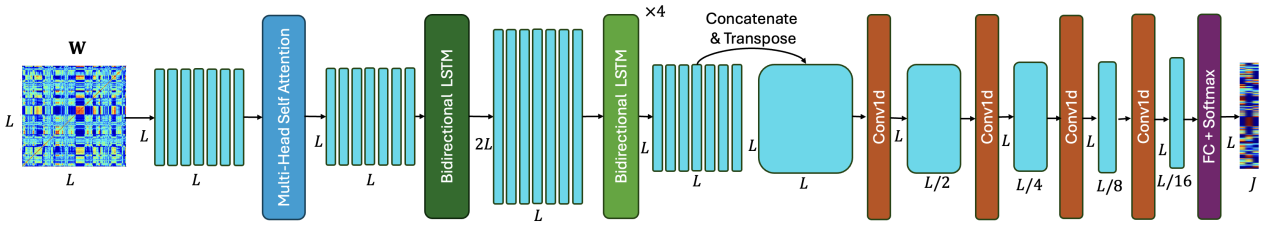
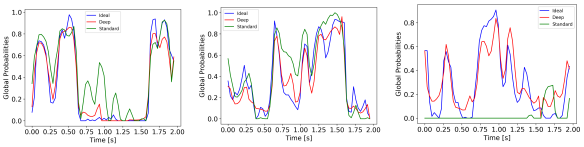


Figure 3: *Architecture of the proposed probability network for predicting activity probabilities based on an input correlation matrix.*



(a) 1^{st} *Speaker* (b) 2^{nd} *Speaker* (c) 3^{rd} *Speaker*

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

attenuation. The loss function hyperparameters used in Eq. (4) are $\lambda_1 = 1000$ and $\lambda_2 = 1$ to obtain align scales.

3.3. Baselines and Evaluation

We compare Deep-Simplex to the learning-free simplex method and a conventional approach based on independent vector analysis (IVA), using the implementation provided by the TorchIVA toolkit[1]. Additionally, we evaluate an ideal baseline, where the same separation procedure described in §2.5 is applied using an ideal mask computed from the individual speaker signals.

We first compare Deep-Simplex and the Standard Simplex in terms of global probability prediction and mask estimation. For global probability prediction, we measure the mean squared error (MSE) between the true and estimated probabilities. For mask estimation, the error is defined as the percentage of TF bins assigned to the wrong speaker. The true probability is derived from the ideal mask, computed as the relative proportion of TF bins dominated by each speaker. Additionally, we evaluate 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 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.

[1]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

Separation Performance. Table 3 compares the separation performance across all methods. Deep-Simplex consistently 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 demonstrate that the proposed method improves global probability estimation over the standard simplex approach and achieves superior separation performance compared to competing baselines for different reverberation levels.

5. 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] 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.

[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] 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.

[13] ——, "Global and local simplex representations for multichannel source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 914–928, 2020.

[14] 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.

[15] B. K. B. R. P. Ghosh, S. K. Roy and P. Scheunders, "Hyperspectral unmixing using transformer network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 516–527, 2022.

[16] J. L. R. Z.-Q. Wang and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022.

[17] P. S. B. Rasti, B. Koirala and J. Chanussot, "Misicnet: Minimum simplex convolutional network for deep hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022.

[18] 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.

[19] E. A. P. Habets, "Room impulse response (rir) generator," Online, July 2006, available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.