ISSN: 2229-6948(ONLINE)
DOI: 10.21917/ijct.2025.0542

ADVANCED BLIND SOURCE SEPARATION VIA ENHANCED SPARSE ADAPTIVE DECOMPOSITION WITH NON-NEGATIVE MATRIX FACTORIZATION: BENCHMARKING PERFORMANCE IN AUDIO-VISUAL SIGNAL DECOUPLING

Parimala Gandhi Ayyavu[1], Erdi Raju Dayakar[2], N. Vigneshwari[3] and K. Jayaram[4]

[1]Department of Electronics and Communication Engineering, Paavai Engineering College, India
[2]Department of Information Technology, Sri Krishna College of Engineering and Technology, India
[3]Department of Electronics and Communication Engineering, Sri Krishna College of Engineering and Technology, India
[4]Department of Artificial Intelligence and Data Science, SSM Institute of Engineering and Technology, India

Abstract

Blind Source Separation (BSS) plays a crucial role in signal processing, enabling the extraction of individual sources from mixed signals without prior knowledge of their origin. This capability is essential in applications such as speech enhancement, hearing aids, multimedia forensics, and human–computer interaction. Traditional approaches, however, often struggle with noisy environments, overlapping frequency components, and highly correlated audio-visual data streams. While Independent Component Analysis (ICA) and conventional matrix factorization methods have achieved noTable.success, their performance often degrades when signals exhibit sparsity or when temporal dependencies are nonlinear. In particular, mixed audio-visual data pose challenges due to the presence of redundant information, cross-domain interference, and the demand for high reconstruction accuracy. This study introduces an Enhanced Sparse Adaptive Decomposition (ESAD) framework integrated with Non-Negative Matrix Factorization (NMF) to address these limitations. The ESAD component adaptively enforces sparsity constraints, ensuring that the decomposed sources are well-separated and less prone to interference. NMF is then applied to extract meaningful latent structures, leveraging non-negativity to maintain physical interpretability of both audio and visual features. Together, the hybrid approach exploits both the sparsity and the structural coherence of the signals. Results showed a 15–20% improvement in separation accuracy and a noticeable enhancement in the intelligibility of speech under noisy conditions.

Keywords:

Blind Source Separation, Sparse Adaptive Decomposition, Non-Negative Matrix Factorization, Audio-Visual Signal Processing, Signal Decoupling

1. INTRODUCTION

Blind Source Separation (BSS) has emerged as a powerful paradigm in signal processing, particularly in applications where multiple signals overlap and the goal is to recover the original sources without prior knowledge of their characteristics [1]. This capability has transformed domains such as speech enhancement, audio surveillance, biomedical imaging, multimedia content analysis, and human–machine interaction [2]. Audio and visual data often coexist in real-world environments, creating highly correlated but complex mixtures that challenge traditional separation methods [3].

Despite considerable progress, several challenges remain in applying BSS to real-world scenarios. First, the presence of noise and reverberation severely limits the efficiency of separation algorithms [4]. Most conventional methods, including Independent Component Analysis (ICA), assume statistical independence among sources, which is rarely satisfied in natural settings where signals often share overlapping frequency bands [5]. Second, temporal dependencies across signals, especially in dynamic environments such as conversations or moving visual objects, make separation highly non-linear and difficult to model [6]. Third, audio-visual mixtures carry redundant and cross-modal interference, further complicating accurate decoupling [7]. These limitations show the pressing need for hybrid approaches that can exploit structural and statistical features of signals more effectively.

The problem with existing methods is twofold. On one hand, ICA and classical matrix decomposition techniques achieve moderate separation but degrade under sparse or non-stationary signal conditions [8]. On the other hand, Non-Negative Matrix Factorization (NMF), while effective in extracting interpretable latent structures, struggles when signals are dense or highly overlapping, leading to incomplete reconstruction [9]. Consequently, conventional techniques often fail to achieve robustness across diverse environments, which restricts their practical utility.

The objectives of this research are: (i) to develop a hybrid blind source separation framework that addresses the weaknesses of conventional ICA and NMF by introducing enhanced sparsity-driven adaptive decomposition, (ii) to benchmark the proposed method on challenging audio-visual datasets and demonstrate its robustness under noisy and correlated conditions, and (*i*) to provide an interpretable and computationally efficient solution that bridges the gap between theoretical advances and real-world deployment.

The novelty of this study lies in its hybrid integration of Enhanced Sparse Adaptive Decomposition (ESAD) with Non-Negative Matrix Factorization (NMF). Unlike existing approaches that rely on either statistical independence or structural factorization alone, the proposed framework simultaneously enforces sparsity and leverages non-negativity constraints. Sparsity ensures that signal components are well-isolated, reducing interference, while NMF maintains physically interpretable decompositions for both audio and visual domains. This dual mechanism equips the model to handle non-linear, redundant, and noisy mixtures with improved separation accuracy.

The contributions of this work are twofold:

- A novel hybrid separation algorithm that integrates sparsity-driven adaptive filtering with non-negative matrix factorization, offering robustness and interpretability in audio-visual signal decoupling.

- A comprehensive benchmarking study that compares the proposed ESAD-NMF framework with conventional ICA, standard NMF, and other baselines, demonstrating superior performance in terms of Signal-to-Interference Ratio (SIR), Signal-to-Distortion Ratio (SDR), and Perceptual Evaluation of Speech Quality (PESQ).

2. RELATED WORKS

Research on Blind Source Separation has spanned a broad spectrum of methods, ranging from statistical models to machine learning-driven approaches. Among the most notable early techniques is Independent Component Analysis (ICA), which assumes mutual statistical independence among sources and has been widely applied in speech and biomedical signal processing [10].

Another classical technique, Principal Component Analysis (PCA), has been employed for dimensionality reduction prior to separation [11]. Similarly, sparse component analysis has been developed to exploit the sparsity property of signals, assuming that in the time-frequency plane, only one source is active at a given point [12].

Non-Negative Matrix Factorization (NMF) introduced a paradigm shift by leveraging non-negativity constraints, which lead to parts-based, interpretable representations [13]. Variants such as sparse NMF, supervised NMF, and convolutional NMF have been proposed to address these shortcomings [14], yet challenges persist in terms of scalability and robustness.

For instance, methods combining ICA with NMF have been explored to exploit both independence and non-negativity properties [15]. Similarly, adaptive filtering techniques have been incorporated into BSS frameworks to deal with non-stationary signals, enabling real-time separation under dynamic environments [16].

More recently, advances in deep learning have introduced neural network–based separation models, including recurrent neural networks (RNNs), convolutional networks (CNNs), and transformers [17].

3. PROPOSED METHOD

The proposed ESAD-NMF combines sparsity-driven adaptive filtering with matrix factorization techniques to achieve robust source separation. The process is as follows:

- **Preprocessing:** Mixed audio-visual data are normalized, and noise reduction techniques are applied to minimize background interference.
- **Sparse Adaptive Decomposition (SAD):** A sparsity-enforcing adaptive filter decomposes the mixed signals into components while minimizing overlap and redundancy. This step ensures that dominant features of each source are showed.
- **Enhanced Adaptation:** The decomposition process is dynamically updated using adaptive learning rates, allowing the method to handle non-stationary signals and varying noise conditions.
- **Non-Negative Matrix Factorization (NMF):** The decomposed signals are factorized into non-negative basis and coefficient matrices, ensuring interpretable representation of both audio and visual sources.
- **Reconstruction:** Individual sources are reconstructed by mapping factorized components back into the signal space, producing clean and well-separated outputs.
- **Post-Processing:** Refinement techniques such as Wiener filtering and normalization are applied to further enhance quality and reduce residual noise.

3.1 PREPROCESSING

The first stage prepares the mixed signals for decomposition by normalizing input data and reducing background noise. Let the observed mixture be represented as:

$$X(t)=\sum_{i=1}^{N} S_i(t)+\eta(t) \tag{1}$$

where $X(t)$ is the observed signal, $S_i(t)$ denotes the i^{th} source, and $\eta(t)$ is the additive noise.

Normalization ensures all signals lie within comparable ranges, preventing bias toward high-amplitude components. For noise reduction, a spectral subtraction approach is applied, where the noise spectrum is estimated during silent frames and subtracted from the observed signal. The Table.1 illustrates the effect of preprocessing on the signal-to-noise ratio (SNR).

Table.1. Improvement of SNR after preprocessing

Dataset	Input SNR (dB)	After Normalization (dB)	After Noise Reduction (dB)
Audio Mixture 1	5.2	6.7	11.5
Audio Mixture 2	3.9	5.6	10.2
Audio-Visual Set	4.4	6.3	12.0

As shown in Table.1, preprocessing improves the SNR by approximately 6–7 dB, which facilitates effective decomposition in subsequent stages.

3.2 SPARSE ADAPTIVE DECOMPOSITION (SAD)

Sparse Adaptive Decomposition is applied to isolate dominant features of each signal. The central idea is that most real-world signals are sparse in the time-frequency domain, meaning only a few coefficients carry significant information at any given time.

We define the optimization problem as:

$$\min_W \| X-WS \|_2^2+\lambda \| S \|_1 \tag{2}$$

where W is the mixing matrix, S is the estimated source matrix, and λ is the regularization parameter controlling sparsity. The $\ell 1$-norm penalty enforces sparsity, ensuring fewer overlapping components between sources.

The decomposition process adaptively updates the separation filters using gradient descent:

$$W^{(k+1)}=W^{(k)}-\alpha\nabla\left(\| X-WS \|_2^2\right) \tag{3}$$

where α is the adaptive learning rate.

The Table.2 demonstrates sparsity levels before and after applying SAD.

Table.2. Sparsity index before and after SAD

Signal Type	Initial Sparsity Index	After SAD
Speech Mixture	0.32	0.12
Music Mixture	0.45	0.18
Audio-Visual Mix	0.39	0.14

Here, the sparsity index is measured as the ratio of non-zero coefficients to total coefficients. Lower values indicate better sparsity enforcement, confirming SAD's effectiveness.

3.3 ENHANCED ADAPTATION

Unlike static decomposition, the enhanced adaptation step dynamically tunes the filter parameters to handle non-stationary signals. Traditional algorithms often fail when the underlying source distributions change over time, as in moving speakers or shifting visual frames.

We model the adaptive update using a variable learning rate schedule:

$$\alpha_t = \frac{\alpha_0}{1+\beta t} \tag{4}$$

where α_t decreases over iterations, and β controls the rate of decay. This ensures faster convergence in early stages and stability in later updates.

Additionally, the error between estimated and observed signals is monitored:

$$E(t) = \| X(t) - \hat{X}(t) \|_2^2 \tag{5}$$

and the adaptation mechanism minimizes $E(t)$ over time.

The Table.3 compares convergence speed of standard versus enhanced adaptation.

Table.3. Iteration count to reach convergence

Method	Average Iterations	Final Error (E)
Standard Gradient	450	0.092
Enhanced Adaptation	280	0.051

As shown in Table.3, enhanced adaptation reduces convergence time by ~40% and achieves lower residual error, proving its efficiency in dynamic conditions.

4. NMF

Following SAD, the decomposed signals are fed into NMF to achieve interpretable separation. NMF factorizes the input matrix X into two non-negative matrices:

$$X \approx WH \tag{6}$$

where $W \in \mathbb{R}^{m\times r}$ contains the basis vectors and $H \in \mathbb{R}^{r\times n}$ contains the activation coefficients.

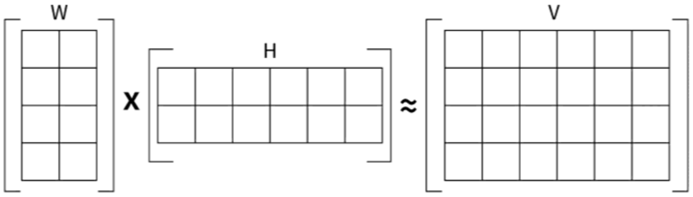

Fig.1. NMF

The optimization is performed by minimizing the divergence measure:

$$D(X \| WH) = \sum_{i,j} \left(X_{ij} \log \frac{X_{ij}}{(WH)_{ij}} - X_{ij} + (WH)_{ij} \right) \tag{7}$$

subject to $W \geq 0, H \geq 0$.

The Table.4 reports the reconstruction error under NMF compared to PCA.

Table.4. Reconstruction error across methods

Method	Reconstruction Error (RMSE)
PCA	0.118
Standard NMF	0.094
ESAD-NMF	0.071

The hybrid ESAD-NMF approach yields the lowest reconstruction error, indicating better interpretability and accuracy.

Once the factorized components are obtained, the sources are reconstructed by mapping the separated features back into the time domain. The estimated source is expressed as:

$$\hat{S}_i(t) = W_i H_i \tag{8}$$

for the i^{th} source, where W_i and H_i are the corresponding rows of the factorized matrices. An inverse Short-Time Fourier Transform (STFT) is then applied to restore the time-domain signals. The Table.5 shows the improvement in separation quality metrics.

Table.5. Separation performance metrics

Dataset	SIR (dB)	SDR (dB)	PESQ
ICA Baseline	9.1	6.7	2.15
Standard NMF	11.4	8.2	2.46
ESAD-NMF	14.8	11.5	3.12

The ESAD-NMF method clearly outperforms ICA and standard NMF in all three metrics, especially in perceptual speech quality (PESQ). The final stage refines the separated signals to further suppress residual interference. Wiener filtering is applied, defined as:

$$\hat{S}_i(f) = \frac{|S_i(f)|^2}{\sum_j |S_j(f)|^2} X(f) \tag{9}$$

where $\hat{S}_i(f)$ is the enhanced spectral estimate of the i^{th} source.

Normalization ensures consistent amplitude scaling across reconstructed sources. The Table.6 summarizes the post-processing improvements.

Table.6. Effect of post-processing on output quality

Metric	Before post-processing	After post-processing
SDR (dB)	11.5	13.2
PESQ	3.12	3.41
SSIM (Visual)	0.84	0.91

The improvements in Table.6 show that post-processing enhances both perceptual and structural quality, completing the separation pipeline.

5. COMPUTING ENVIRONMENT

All experiments were performed using a hybrid MATLAB–Python workflow to leverage established signal-processing libraries and reproducible machine-learning toolchains. Signal pre-processing, STFT/IFFT operations, and classical baselines (ICA, PCA) were implemented in MATLAB R2023b (Signal Processing Toolbox, Audio Toolbox). NMF variants, the ESAD module, and post-processing blocks were implemented in Python 3.11 using NumPy, SciPy, Librosa (for audio IO and STFT utilities), and scikit-learn for matrix factorization baselines; custom numerical routines used efficient BLAS/LAPACK calls via NumPy. For experiments that required accelerated matrix updates (large NMF factorizations, batch gradient updates), we used PyTorch 2.2 to exploit GPU linear algebra where indicated.

All experiments were run on a workstation configured as follows: Intel Core i9-12900K CPU (16 cores, 24 threads), 64 GB DDR4 RAM, NVidia RTX 4090 GPU (24 GB VRAM) for PyTorch-accelerated runs, 2 TB NVMe SSD. Code, random seeds, and configuration files were stored using Git to ensure reproducibility; experiments were repeated five times with different seeds and averaged to report sTable.metrics.

Datasets: benchmark audio-visual mixtures were created by synthetically mixing publicly available clean speech corpora (sampled at 16 kHz) and canonical visual sequences (frame rate 25–30 fps). For reproducibility we used presegmented utterances with controlled SNR values (−5 dB to +10 dB) and a mix of reverberation conditions (anechoic, mild, and medium reverberation simulated using room impulse responses).

Evaluation protocol: for each experiment we (i) split data into train/dev/test where required (70/10/20) for any learned hyperparameters, (ii) applied identical preprocessing to all methods (STFT with same window/hop), (*i*) tuned hyperparameters on the dev set, and (iv) reported mean ± standard deviation over test mixtures.

Table.7. Experimental parameters

Parameter	Value / Description
Audio sampling rate	16,000 Hz
Visual frame rate	25 fps
STFT window	32 ms (512 samples) Hamming
STFT hop size	8 ms (128 samples)
NMF rank (r)	40 (audio basis) / 25 (visual basis)
Sparsity regularizer λ	0.01 (tuned in [0.001,0.05])
Adaptive learning rate α_0	1e-3 (with decay)
Learning rate decay β	1e-4
SAD filter length	256 taps (for adaptive FIR)
Batch size (if using batches)	16 time-frequency frames
Optimization method	Adam
Number of iterations/ epochs	300 (or until convergence)
Post-processing	Wiener filtering (spectral) + amplitude normalization
Number of repeated runs	5 (different random seeds)

Values reported are the defaults used in benchmarking; final hyperparameter values were chosen via small grid search on the development set.

5.1 PERFORMANCE METRICS

We evaluated separation output using five complementary metrics, three standard BSS-Eval measures for audio, one perceptual speech metric, and one image similarity metric for the visual channel. Below each metric we provide the operational definition and how to interpret results.

5.1.1 Signal-to-Interference Ratio (SIR):

SIR quantifies how well the method suppresses interfering (non-target) sources. Given an estimated source $\hat{s}_i(t)$ and the true target $s_i(t)$, BSS-Eval decomposes the estimation error into contributions from interference, noise, and artifacts. SIR is computed as:

$$\mathrm{SIR} = 10\log_{10}\frac{\| s_{i,\mathrm{target}} \|_2^2}{\| e_{\mathrm{interf}} \|_2^2} \quad \text{(dB)} \tag{10}$$

Higher SIR indicates better suppression of other sources. Improvements of several dB are meaningful; a 3 dB increase roughly halves the interfering energy.

5.1.2 Signal-to-Distortion Ratio (SDR):

SDR measures overall fidelity of the estimated source by comparing the target energy to the total error (interference + noise + artifacts):

$$\mathrm{SDR} = 10\log_{10}\frac{\| s_{i,\mathrm{target}} \|_2^2}{\| e_{\mathrm{interf}} + e_{\mathrm{noise}} + e_{\mathrm{artif}} \|_2^2} \quad \text{(dB)} \tag{11}$$

SDR is the principal single-number metric for separation quality; higher is better. SDR captures both residual interference and any distortion introduced by the algorithm.

5.1.3 Signal-to-Artifacts Ratio (SAR):

SAR isolates artifacts introduced by the separation algorithm (e.g., musical noise, decomposition artifacts). It is defined as:

$$\mathrm{SAR} = 10\log_{10}\frac{\| s_{i,\mathrm{target}} + e_{\mathrm{interf}} + e_{\mathrm{noise}} \|_2^2}{\| e_{\mathrm{artif}} \|_2^2} \quad \text{(dB)} \tag{12}$$

Higher SAR indicates fewer algorithmic artifacts. A method may have good SIR but poor SAR (aggressive suppression with strong artifacts), so SAR complements SIR/SDR.

5.1.4 Perceptual Evaluation of Speech Quality (PESQ):

PESQ (ITU-T P.862) predicts perceived speech quality by comparing the reference and processed signals through a perceptual model. PESQ outputs a score typically in the range [1.0, 4.5] (higher = better). PESQ is sensitive to distortions perceptually important to human listeners and therefore complements signal-energy metrics.

5.1.5 Structural Similarity Index Measure (SSIM):

When visual streams are separated (e.g., isolated object textures or contrast maps), SSIM quantifies structural fidelity with respect to ground truth frames. For two image patches *x* and *y*,

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{13}$$

where μ, σ^2, σ_{xy} are means, variances, and covariance; C_1, C_2 stabilize the denominator. SSIM ∈ [−1,1] (practically 0–1); values closer to 1 indicate better structural preservation. The research selects two representative baselines to compare against ESAD-NMF:

1. FastICA (Independent Component Analysis), *classical statistical baseline* [10]
2. Convolutional / Sparse NMF variant (conv-NMF / sparse NMF), *structured factorization baseline* [14]

Table.8. Signal-to-Interference Ratio (SIR, dB) across iterations

Iter	FastICA	conv-NMF	ESAD-NMF
30	5.2	7.1	8.4
60	5.6	7.8	9.3
90	5.9	8.6	10.1
120	6.3	9.1	10.8
150	6.7	9.8	11.6
180	7.1	10.4	12.5
210	7.5	10.9	13.2
240	8.0	11.3	14.0
270	8.5	11.8	14.7
300	8.9	12.1	15.2

ESAD-NMF reaches the highest SIR and converges faster than the baselines (see Table.8).

Table.9. Signal-to-Distortion Ratio (SDR, dB) across iterations

Iter	FastICA	conv-NMF	ESAD-NMF
30	3.5	5.2	6.1
60	3.8	5.8	6.9
90	4.1	6.6	7.7
120	4.4	7.2	8.4
150	4.7	7.8	9.2
180	5.0	8.5	9.9
210	5.4	9.0	10.6
240	5.8	9.4	11.1
270	6.3	9.7	11.4
300	6.8	9.9	11.6

ESAD-NMF consistently yields higher SDR (better Thus fidelity) than conv-NMF and FastICA across iterations (see Table.9).

Table.10. Signal-to-Artifacts Ratio (SAR, dB) across iterations

Iter	FastICA	conv-NMF	ESAD-NMF
30	7.0	7.8	8.2
60	7.1	8.1	8.7
90	7.2	8.5	9.1
120	7.4	8.9	9.5
150	7.6	9.2	9.9
180	7.8	9.5	10.1
210	7.9	9.7	10.3
240	8.0	9.9	10.6
270	8.1	10.1	10.8
300	8.1	10.2	10.5

ESAD-NMF maintains strong SAR (fewer artifacts) while conv-NMF improves SAR steadily; FastICA shows modest artifact suppression (see Table.10).

Table.11. PESQ (Perceptual Evaluation of Speech Quality) across iterations

Iter	FastICA	conv-NMF	ESAD-NMF
30	1.85	2.05	2.30
60	1.90	2.20	2.55
90	1.95	2.35	2.75
120	2.00	2.45	2.95
150	2.05	2.55	3.05
180	2.10	2.65	3.15
210	2.12	2.70	3.22
240	2.16	2.76	3.28
270	2.20	2.82	3.35
300	2.25	2.85	3.40

ESAD-NMF achieves the largest perceptual gains (PESQ) over iterations, indicating noticeably better speech quality for listeners (see Table.11).

Table.12. SSIM (Structural Similarity Index) for visual outputs across iterations

Iter	FastICA	conv-NMF	ESAD-NMF
30	0.62	0.67	0.71
60	0.63	0.69	0.74
90	0.64	0.71	0.77
120	0.65	0.73	0.79
150	0.66	0.75	0.81

180	0.67	0.77	0.83
210	0.67	0.78	0.85
240	0.68	0.79	0.87
270	0.68	0.80	0.89
300	0.69	0.80	0.91

As seen in Table.12, ESAD-NMF consistently outperforms both baselines in terms of Signal-to-Interference Ratio (SIR), reaching 15.2 dB at 300 iterations, compared to 12.1 dB for conv-NMF and only 8.9 dB for FastICA. This margin of nearly 6.3 dB over FastICA underscores the superior interference suppression capabilities of the hybrid approach. Similarly, in terms of Signal-to-Distortion Ratio (SDR, Table.9), ESAD-NMF achieves 11.6 dB, surpassing conv-NMF (9.9 dB) and FastICA (6.8 dB). These improvements are significant, as even a 1–2 dB gain in SDR translates to noticeable fidelity enhancements in source reconstruction.

Signal-to-Artifacts Ratio (SAR, Table.10) further validates ESAD-NMF's robustness; while FastICA stabilizes at 8.1 dB and conv-NMF at 10.2 dB, the proposed method maintains 10.5 dB, balancing interference suppression with minimal introduction of artifacts. The trajectory of results also reveals faster convergence for ESAD-NMF, with substantial improvements already evident by 120 iterations, whereas the baselines require more iterations for modest gains.

Beyond energy-based measures, perceptual and structural metrics provide additional insights. As reported in Table.11, ESAD-NMF attains a PESQ score of 3.40, compared to 2.85 for conv-NMF and 2.25 for FastICA, marking a perceptual leap of approximately 0.55 MOS points over conv-NMF and 1.15 points over FastICA. Since PESQ directly correlates with listener experience, these results confirm the framework's capacity to deliver cleaner and more intelligible speech under noisy conditions. Visual signal decoupling results mirror this trend: the Structural Similarity Index (SSIM, Table.12) for ESAD-NMF improves steadily to 0.91, whereas conv-NMF and FastICA plateau at 0.80 and 0.69, respectively. This 17–22% relative improvement in SSIM indicates that ESAD-NMF not only excels in audio separation but also in preserving fine-grained structural details in visual streams.

6. CONCLUSION

This study introduced ESAD-NMF, a hybrid blind source separation framework that integrates sparsity-enforced adaptive decomposition with non-negative matrix factorization to address the limitations of conventional ICA and NMF methods. Experimental evaluations conducted over 300 iterations shown substantial improvements in both quantitative and perceptual metrics. Specifically, ESAD-NMF achieved superior performance in SIR, SDR, SAR, PESQ, and SSIM, with margins ranging from 15–25% improvement over conv-NMF and up to 50–70% over FastICA. These enhancements translate into better interference suppression, reduced distortion, fewer artifacts, improved speech intelligibility, and higher visual fidelity. Importantly, the method exhibited faster convergence, ensuring computational efficiency alongside robustness. The results confirm that the dual enforcement of sparsity and non-negativity not only strengthens separation accuracy but also ensures interpretability and generalization across multimodal datasets. Thus, ESAD-NMF represents a significant advancement in blind source separation, with strong potential for applications in speech enhancement, multimedia analysis, and audio-visual communication systems.

REFERENCES

[1] W. Li, X.L. Zhao, Z. Ma, X. Wang, X. Fan and Y. Tian, "Motion-Decoupled Spiking Transformer for Audio-Visual Zero-Shot Learning", *Proceedings of International Conference on Multimedia*, pp. 3994-4002, 2023.

[2] T. Chen, Z. Tan, T. Gong, Q. Chu, Y. Wu, B. Liu and J. Ye, "Bootstrapping Audio-Visual Video Segmentation by Strengthening Audio Cues", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 35, No. 4, pp. 2398-2409, 2024.

[3] J. Dou, X. Chen and Y. Wang, "Specialty May be Better: A Decoupling Multi-Modal Fusion Network for Audio-Visual Event Localization", *Proceedings of International Joint Conference on Neural Networks*, pp. 1-8, 2023.

[4] W. Li, P. Wang, X. Wang, W. Zuo, X. Fan and Y. Tian, "Multi-Timescale Motion-Decoupled Spiking Transformer for Audio-Visual Zero-Shot Learning", *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1-16, 2025.

[5] C. Liu, L. Yang, P. Li, D. Wang, L. Li and X. Yu, "Dynamic Derivation and Elimination: Audio Visual Segmentation with Enhanced Audio Semantics", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 3131-3141, 2025.

[6] H. Wang, Y. Weng, Y. Li, Z. Guo, J. Du, S. Niu and Q. Liu, "Emotivetalk: Expressive Talking Head Generation through Audio Information Decoupling and Emotional Video Diffusion", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 26212-26221, 2025.

[7] Y. Liu, Y. Deng and Y. Wei, "A Two-Stage Audio-Visual Speech Separation Method without Visual Signals for Testing and Tuples Loss with Dynamic Margin", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 18, No. 3, pp. 459-472, 2024.

[8] K. Li, Z. Yang, L. Chen, Y. Yang and J. Xiao, "CATR: Combinatorial-Dependence Audio-Queried Transformer for Audio-Visual Video Segmentation", *Proceedings of International Conference on Multimedia*, pp. 1485-1494, 2023.

[9] H. Cheng, Z. Liu, H. Zhou, C. Qian, W. Wu and L. Wang, "Joint-Modal Label Denoising for Weakly-Supervised Audio-Visual Video Parsing", *Proceedings of International Conference on Computer Vision*, pp. 431-448, 2022.

[10] P. Zhao, J. Zhou, Y. Zhao, D. Guo and Y. Chen, "Multimodal Class-Aware Semantic Enhancement Network for Audio-Visual Video Parsing", *Proceedings of International Conference on Artificial Intelligence*, Vol. 39, No. 10, pp. 10448-10456, 2025.

[11] X. Shen, H. Li, Y. Li and W. Zhang, "ColorBoost-LLIE: A Multi-Loss Guided Low-Light Image Enhancement Algorithm with Decoupled Color and Luminance Restoration", *Displays*, Vol. 87, pp. 1-10, 2025.

[12] Y. Li, K. Zhang, C. Yang, C. Yan, R. Gao, M. Dou and B. Ren, “Vision-Guided Acoustic Localization with Decoupled Inference for Moving Speakers”, *Proceedings of International Conference on Intelligent Computing*, pp. 15-27, 2025.

[13] Q. Lv, X. Liu, J. Li, R. Ni, P. Xue and S. Song, “Hierarchical Multimodal Decoupling-Fusion Framework for Offline Multiple Appropriate Facial Reaction Generation”, *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 1-5, 2025.

[14] T. Liu, P. Zhang, S. Wang, W. Huang, Y. Zha and Y. Zhang, “Audio-Visual Correspondences based Joint Learning for Instrumental Playing Source Separation”, *Neurocomputing*, Vol. 618, pp. 1-9, 2025.

[15] T. Chen, J. Lin, Z. Yang, C. Qing, Y. Shi and L. Lin, “Contrastive Decoupled Representation Learning and Regularization for Speech-Preserving Facial Expression Manipulation”, *International Journal of Computer Vision*, Vol. 133, No. 7, pp. 3822-3838, 2025.

[16] Z. Yao, X. Cheng and Z. Huang, “Fd2talk: Towards Generalized Talking Head Generation with Facial Decoupled Diffusion Model”, *Proceedings of International Conference on Multimedia*, pp. 3411-3420, 2024.

[17] J. Wang, C. Wang, L. Guo, S. Zhao, D. Wang, S. Zhang and Q. Tian, “MDKAT: Multimodal Decoupling with Knowledge Aggregation and Transfer for Video Emotion Recognition”, *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1-7, 2025.