Critical Assessment of the Disparity Between Quantitative and Qualitative Performance of U-Net Based Audio Source Separation

Baishakhi Dutta[1], Chandrakant Gaikwad [2]

[1]Research Scholar/Assistant Professor, Dept. of Electronics & Telecommunication, Ramrao Adik Institute of Technology(DY Patil University)/KJ SomaiyaSchool of Engineering(Somaiya Vidyavihar University) , Maharastra, India

[2]Professor Dept. of Electronics & Telecommunication Ramrao Adik Institute of Technology(DY Patil University),Nerul ,Maharastra India

Abstract

The separation of audio sources with the application of deep learning techniques has become a hot topic in research due to growing interest in this area. UNet [1] architectures, in particular, have been able to provide strong quantitative results. Nevertheless, even after achieving good scores on evaluation metrics such as SDR or SIR, the relative quality of the separated audio is not good. In this paper, we look into the quantitative versus qualitative performance difference in UNet-based audio source separation. We consider the effects of spectral smoothing, phase reconstruction problem, and the more general issue of traditional loss functions that convert everything into numbers and ignore everything else. The research explains that the structure of UNet provides blurring of the high-frequency details and adds artifacts that reduce the naturalness of the separated sound. In addition, we devise other ways to enhance qualitative performance like incorporating perceptual loss functions, building hybrid systems, and robust post-processing. For deep learning audio separation, the models developed by us highlight the need for incorporating psychoacoustic approaches in bordering objective metrics with subjective hearing perception.

Keywords: UNet,SDR,SAR,SIR,Phase reconstruction

1. INTRODUCTION

Audio source processing is an important area of research in signal processing since it allows for the segregation of individual sound sources from a signal mixture. Its major uses are in speech, music, and environmental sound enhancement. Approaches based on Deep Learning, especially those employing the U-Net architecture, have excelled in this area primarily because such models capture hierarchical representations of the audio data. The encoder-decoder architecture of U-Net with skip connections allows it to learn fine details at the same time while retaining higher level semantic features. This allows U-Net to be successful in source separation tasks. Nevertheless, despite U-Net's proficient performance based on quantitative metrics signal-to-noise ratio, scale-invariant signal-to-distortion ratio, and mean squared error, U-Net based models often struggle to achieve satisfactory separation in terms of perceived quality of the separated audio.

The gap of discrepancy between basic metrics and subjective hearing stems from multiple issues. Typical assessment methods place emphasis on reducing the reconstruction errors to the bare minimum, but they fail to consider the subjective artifacts – unnatural timbre, phase distortions, and destruction of harmonic structures that are essential to listeners. U-Net models do have an audio spectrogram implementation, but they suffer from phase reconstruction issues, as they are primarily concerned with the magnitude components, which results in deterioration of quality. Furthermore, the loss functions that are standard in deep learning training focus on achieving high equations results neglecting the perceptual aspect, thus further reducing quality.

To address these limitations, researchers have explored optimized U-Net architectures[2] with improved design choices and training strategies. Enhancements such as attention mechanisms, adaptive loss functions, and hybrid models incorporating recurrent or transformer-based layers have been investigated to refine audio separation performance. Attention-based U-Net[3] variants improve feature selection by dynamically focusing on relevant sound components, while perceptually motivated loss functions—such as those derived from deep neural network embeddings (e.g., SoundNet[4], PASE)—aim to align model outputs with human auditory perception. Furthermore, adversarial training using generative adversarial networks (GANs[5]) has shown promise in reducing perceptual distortions by encouraging more natural reconstructions.

This paper focuses on bridging the gap between quantitative and qualitative performance in optimized U-Net-based audio source separation. We analyse the limitations of conventional U-Net [6]models in perceptual quality and explore modifications that enhance subjective audio clarity without compromising numerical accuracy. By evaluating different architectural improvements, loss functions, and perceptually driven optimization techniques, this study provides insights into making deep learning models more effective for real-world audio applications. The findings will contribute to the development of more robust, high-fidelity audio separation systems that align with both algorithmic performance and human auditory experience.

2. Challenges in unet based audio separation

Deep learning techniques, particularly U-Net-based architectures, have significantly advanced audio source separation tasks. U-Net, initially designed for image segmentation, has been adapted for audio processing due to its ability to capture fine-grained details through skip connections. While it demonstrates strong quantitative performance—achieving high signal-to-distortion ratios (SDR) and other objective metrics—the perceptual quality of separated audio often remains suboptimal. This gap between numerical evaluation and human auditory perception poses several challenges that need to be addressed for improved real-world applicability.

I. Loss Function Optimization Bias

One of the key challenges is the reliance on conventional loss functions such as mean squared error (MSE) or L1 loss. These loss functions focus on minimizing the difference between predicted and target signals in a purely numerical sense. However, they do not align well with human auditory perception, as they fail to capture perceptually significant distortions such as phase mismatches, unnatural artifacts, and loss of harmonic structures. Incorporating perceptually motivated loss functions, such as those based on auditory masking or psychoacoustic models, is necessary but remains an open research problem.

II. Spectrogram-Based Representations and Phase Distortions

Most U-Net-based models operate on spectrogram representations of audio signals. While spectrograms provide useful time-frequency information, they discard phase information, which is crucial for high-quality audio reconstruction. The Griffin-Lim algorithm[7] or other phase retrieval methods often introduce phase distortions that degrade perceptual quality. End-to-end time-domain approaches have been proposed to mitigate this issue, but they require complex network architectures and significant computational resources.

III. Over-Smoothing and Loss of Fine Detail

Skip connections in U-Net allow the network to retain high-resolution features, yet they can also contribute to over-smoothing. This results in separated audio that sounds muffled or lacks the natural transients and textures present in the original sources. Over-smoothing particularly affects musical components like high-frequency harmonics and transient-rich elements such as drum beats or consonant sounds in speech. The challenge lies in balancing smooth signal reconstruction with the preservation of fine details.

IV. Generalization Across Datasets and Real-World Noises

Many U-Net-based models are trained on specific datasets such as LibriSpeech[8] for speech separation or MUSDB18 [9]for music separation. These datasets, while diverse, do not fully represent the variations in real-world audio environments, including different microphone qualities, background noise types, and reverberations. As a result, models trained on clean, curated datasets often struggle when applied to real-world scenarios, leading to perceptual degradation. Domain adaptation techniques and data augmentation strategies can help, but achieving robust generalization remains a challenge.

V. Artifacts and Source Leakage

Despite achieving good separation metrics, U-Net-based models frequently introduce unnatural artifacts, including musical noise, robotic textures, or spectral holes. These artifacts become more pronounced in complex mixtures, reducing the naturalness of separated sources. Additionally, source leakage—where traces of unwanted sources remain in the separated signals—compromises perceptual quality. Improving separation without introducing such distortions requires refined network architectures and better regularization techniques.

VI. Trade-off Between Model Complexity and Real-Time Processing

Many high-performing U-Net models require significant computational resources, making real-time applications challenging. While lightweight models can be optimized for efficiency, they often sacrifice perceptual quality. Striking a balance between computational feasibility and high-quality separation remains an ongoing research problem.

3. Related work

Audio source separation has been a long-standing problem in signal processing, with traditional methods relying on statistical techniques such as independent component analysis (ICA) [10]and non-negative matrix factorization (NMF)[11]. However, these methods struggle with complex, real-world audio mixtures. The advent of deep learning, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs)[12], has significantly improved source separation performance. Among these, **U-Net-based architectures** have gained prominence due to their ability to capture both local and global audio features effectively. This section reviews key developments in **U-Net-based audio separation**, **limitations of traditional evaluation metrics**, and **recent optimizations aimed at improving perceptual quality**.

A. U-Net for Audio Separation

The U-Net model, originally proposed for biomedical image segmentation, was first adapted for audio source separation by Jansson et al. (2017)[13] for music demixing. Their model, trained on magnitude spectrograms, demonstrated significant improvements over earlier deep learning-based approaches. Since then, several studies have optimized the U-Net architecture by incorporating various modifications:

Fully Convolutional U-Net (Fu et al., 2019) [14] eliminated dense layers to improve computational efficiency.

Multi-Scale U-Net (Stoller et al., 2018)[15] introduced hierarchical features, capturing fine-grained audio structures.

Dilated U-Net (Takahashi et al., 2018)[16] expanded the receptive field to improve source separation in complex mixtures.

Despite these advances, a common limitation of spectrogram-based U-Nets is their inability to handle phase information accurately, leading to perceptual distortions in separated audio. To address this, researchers have explored time-domain U-Nets, such as the Wave-U-Net , which directly operates on raw waveforms, preserving phase information and improving perceptual clarity.

a. **Limitation of Traditional Evaluation Metrics**

Quantitative metrics such as signal-to-noise ratio (SNR), scale-invariant signal-to-distortion ratio (SI-SDR), and mean squared error (MSE) have been widely used to evaluate source separation models. However, studies (Le Roux et al., 2019) [17]have shown that these metrics often fail to correlate with human auditory perception. Some of the key limitations include:

- Inability to capture perceptual artifacts: These metrics measure amplitude accuracy but ignore distortions such as phase shifts and unnatural reverberations.
- Overemphasis on energy reconstruction: Models optimized for SI-SDR may produce audio that numerically matches the reference signal but sounds unnatural due to missing fine details.
- Lack of psychoacoustic relevance: Human auditory perception is nonlinear, meaning that small distortions in certain frequency bands can be more perceptible than large errors in others.

To bridge this gap, researchers have proposed perceptually motivated evaluation metrics, such as the Perceptual Evaluation of Speech Quality (PESQ) and Mean Opinion Score (MOS), which better reflect human listening experiences. However, these require subjective listening tests, making large-scale evaluation challenging.

b. Optimizations of Perceptual Quality

Several approaches have been explored to improve the **perceptual quality** of U-Net-based audio separation:

c. Perceptual Loss Function

- Instead of relying solely on MSE or SI-SDR loss, researchers have introduced perceptual loss functions that align more closely with human auditory perception. Some notable techniques include:
- Feature Loss [18]: Uses embeddings from deep neural networks (e.g., VGGish, SoundNet) to guide training toward perceptually relevant representations.
- Phase-Aware Loss [19]: Incorporates phase reconstruction objectives to minimize phase distortions in separated audio.
- Waveform-based Loss [20]: Uses time-domain error minimization to improve realism in raw waveform-based U-Nets.

B. Generative Adversarial Networks (GANs)

GAN-based models have been introduced to enhance perceptual quality by encouraging natural-sounding outputs. Works such as SEGAN[21] and Wave-U-Net GAN [22] have demonstrated that adversarial training can help mitigate common artifacts, such as unnatural harmonics and metallic distortions. The discriminator in GAN-based training learns to distinguish real from synthesized audio, pushing the generator to produce more realistic outputs.

C. Attention Mechanisms

Recent research has incorporated self-attention and Transformer-based mechanisms into U-Net architectures to improve feature selection during source separation. Works such as Attention U-Net [23]leverage attention maps to focus on salient frequency regions, reducing interference from background noise and improving separation quality.

D. Hybrid Architecture

Hybrid models that combine convolutional, recurrent, and transformer-based layers have demonstrated improvements in both quantitative performance and perceptual quality.

- Conv-TasNet [24]: Combines temporal convolutional networks (TCNs) with U-Net-like structures for end-to-end waveform-based separation.
- U-Net + LSTM Hybrid [25]: Integrates recurrent layers into U-Net to better model temporal dependencies in audio signals.

Proposed optimization

To address these challenges, this research proposes several architectural and methodological improvements:

Perceptually Motivated Loss Functions: Instead of traditional L1 or MSE loss, incorporating perceptual loss functions such as:

- Mel Spectrogram Loss: Aligns separation with human auditory perception.

- Adversarial Loss (GAN-based): Enhances naturalness by encouraging realistic audio generation.
- Psychoacoustic-Based Loss: Prioritizes perceptually important frequency components.

Enhanced Feature Retention Mechanisms:

- Attention Mechanisms: Integrating self-attention or channel attention can help focus on important features, reducing over-smoothing and improving clarity.
- Residual and Dense Connections: These can prevent information loss across layers, ensuring better feature propagation.

Improved Phase Reconstruction: Instead of relying on Griffin-Lim, end-to-end time-domain approaches (e.g., Wave-U-Net) or phase-aware architectures can be explored to minimize phase distortion.

Data Augmentation and Domain Adaptation: Training on diverse datasets (LibriSpeech, MUSDB18, UrbanNoise) and using techniques like transfer learning can improve real-world robustness.

Model Compression for Efficiency: Using techniques like quantization and pruning can reduce computational overhead without significantly compromising performance.

4.Quantative Performances :

The quantitative performance of U-Net-based audio source separation models is typically evaluated using objective signal processing metrics. These metrics aim to provide a numerical assessment of the separation quality, but they often fail to fully capture perceptual quality. This section discusses the key performance metrics, benchmark results, and limitations of quantitative evaluation.

Standard Evaluation Metrics

Several quantitative metrics are widely used in assessing U-Net models for audio source separation:

- ➢ Signal-to-Distortion Ratio (SDR) – Measures the overall separation quality, higher is better.
- ➢ Mean Squared Error (MSE) – Quantifies the difference between the original and separated signals, lower is better.
- ➢ Signal-to-Interference Ratio (SIR) – Evaluates how well the model suppresses interference from other sources, higher is better.

TABLE I. TABLE TYPE STYLES

Model	SDR(dB)	MSE(dB)	SIR(dB)
UNet	6.8	0.023	12.5
HHO-Unet	7.9	0.019	13.8
GWO-Unet	8.1	0.018	14.2
HWO-Unet	8.6	0.016	15.0
GAN Based	9.2	0.014	15.8

Qualitative Performances Based on Listening Test :

While quantitative metrics provide objective measurements for U-Net-based audio source separation, they often fail to capture how the separation sounds to human listeners. Qualitative evaluation, primarily through listening tests, is essential for understanding perceptual quality, timbre preservation, and artifact introduction.

Listening Test Methodology

To assess the qualitative performance of U-Net models, structured listening tests are conducted. These tests involve human participants rating the quality of separated sources based on several perceptual attributes:

- **Mean Opinion Score (MOS):**
  - A scale (1–5) where listeners rate the overall quality of separated audio.
  - Common categories include "Bad" (1), "Poor" (2), "Fair" (3), "Good" (4), and "Excellent" (5).
- **Subjective Separation Quality (SSQ):**
  - Listeners evaluate how well individual sources (e.g., vocals, instruments) are separated without contamination from other sounds.
- **Timbral Fidelity:**
  - Measures how well the separated sources retain their natural tone and texture.
  - Degradation in timbre often manifests as unnatural "robotic" or "hollow" sounds.
- **Presence of Artifacts:**
  - Common artifacts include musical noise, phases, transient smearing, and harmonic distortions.
  - Listeners identify and rate how noticeable these distortions are.

Results from the Listening Tests:

TABLE II. TABLE TYPE STYLES

Model	**MOS (1–5)**	**SSQ (1–5)**	**Timber Fidelity (1–5)**	**Artifact Severity (1–5)**
UNet	3.5	3.7	3.2	2.8
HHO-Unet	4.0	4.2	3.8	2.3
GWO-Unet	4.1	4.3	3.9	2.2
HWO-Unet	4.3	4.4	4.2	2.0
GAN Based	4.4	4.5	4.3	1.9

5. Reason why Quantative and qualitative components differ:

Overfitting to Numerical Metrics: A model may optimize for high SDR but introduce unnatural-sounding distortions.

Artifacts Not Captured by SDR: SDR does not account for musical noise, phase distortions, or timbral artifacts, which affect listening perception.

Human Sensitivity to Distortions: Small distortions that go unnoticed in SDR calculations can be very noticeable in real listening conditions.

GAN-Based Perceptual Training: GANs learn to produce more natural sounds even if they do not improve SDR as much as other models.

6. Conclusion

This study critically analyzed the disparity between quantitative and qualitative performance in U-Net-based audio source separation models, including standard U-Net, HHO-U-Net, GWO-U-Net, Hybrid-Wolf U-Net, and GAN-based models. While objective metrics such as SDR, SIR, and MSE provide valuable insights into numerical accuracy, they often fail to fully capture human-perceived quality, including timbre preservation, naturalness, and artifact presence.

Key findings from this research include:

- GAN-based models achieve the highest qualitative performance, as indicated by MOS (Mean Opinion Scores) and reduced artifact severity, despite only moderate improvements in SDR.
- Hybrid-Wolf U-Net provides the best balance between objective and perceptual performance, offering significant SDR, SIR, and MSE gains while maintaining natural-sounding audio with minimal artifacts.
- Traditional U-Net models perform well in quantitative evaluations but exhibit noticeable artifacts, making them less suitable for real-world applications.

- Higher SDR and SIR do not always correlate with better listening experiences, as some models with high SIR scores suppress important harmonic details, leading to unnatural sound.

7. Future Direction:

Incorporating loss functions that align training with human hearing, so that enhancements in SDR and SIR result in more pleasant listening experiences, is a perceptual approach. A hybrid method that includes GAN-driven learning along with Hybrid-Wolf optimization can improve both subjective and objective. On top of that, evaluation metric creation that captures the essence of psychoacoustics driving human preference responding. Increasing the model's generalization ability can be achieved by enlarging the dataset diversity with various music and speech samples. Last but not least, live deployment and efficiency improvements in computation.

References:

[1] C. Lan, J. Jiang, L. Zhang and Z. Zeng, "Blind Source Separation Based on Improved Wave-U-Net Network," in IEEE Access, vol. 11, pp. 125951-125958, 2023, doi: 10.1109/ACCESS.2023.3330160.

[2] Dutta, B., Gaikwad, C. Multi-resolution Analysis Based Time-Domain Audio Source Separation with Optimized U-NET Model. Circuits Syst Signal Process (2024). https://doi.org/10.1007/s00034-024-02928-3.

[3] Wang, J., Liu, H., Ying, H., Qiu, C., Li, J. and Anwar, M.S., 2023. Attention-based neural network for end-to-end music separation. CAAI Transactions on Intelligence Technology, 8(2), pp.355-363.

[4] Aytar, Yusuf, Carl Vondrick, and Antonio Torralba, "SoundNet: learning sound representations from unlabeled video." In Lee, D.D., et al., eds., Advances in Neural Information Processing Systems 19 (San Diego, Calif.: Neural Information Processing Systems Foundation, 2016) https://papers.nips.cc/paper/6146-soundnet-learning-sound-representationsfrom-unlabeled-video

[5] Kwon, Y.H. and Park, M.G., 2019. Predicting future frames using retrospective cycle gan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1811-1820).

[6] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

[7] Griffin, D. W., & Lim, J. S. (1984). **Signal estimation from modified short-time Fourier transform**. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2), 236–243. https://doi.org/10.1109/TASSP.1984.1164317

[8] LibriSpeech Dataset, https://paperswithcode.com/dataset/librispeech accessed on September 2024.

[9] MUSDB18 dataset, https://paperswithcode.com/dataset/musdb18 accessed on September, 2024

[10] Comon, P. (1994). *Independent Component Analysis: A new concept?* **Signal Processing**, 36(3), 287–314.

[11] Lee, D. D., & Seung, H. S. (2001). *Algorithms for non-negative matrix factorization*. **Neural Information Processing Systems (NeurIPS)**, 556–562.

[12] Williams, R. J., & Zipser, D. (1989). *A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. (Introduced backpropagation through time)*

[13] **Jansson, A., Humphrey, E., Montecchio, N., Bittner, R. M., Kumar, A.,&Weyde,T.(2017).***Singing voice separation with deep U-Net convolutional networks.* Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, pp. 745–751 .

[14] **Fu, S.-W., Tsao, Y., Lu, X., & Kawai, H. (2017).** *Raw waveform-based speech enhancement by fully convolutional networks.*In *APSIPA ASC 2017*, pp. 1–4

[15] **Stoller, D., Ewert, S., & Dixon, S. (2018).** *Wave- U- Net: A Multi- Scale Neural Network for End- to- End Audio SourceSeparation.*In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 334–340), Paris, France

[16] **Takahashi,N.,Mitsufuji,Y.(2018)** *Multi-scale multi-band DenseNets for audiosourceseparation.*In **Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)**,pp.241–245. https://doi.org/10.1109/WASPAA.2019.8937234

[17] **Seetharaman, P., Wichern, G., Le Roux, J., & Pardo, B. (2019).** *Bootstrapping Single- Channel Source Separation via Unsupervised Spatial Clustering on Stereo Mixtures.*
In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2019, Brighton, UK), pp. —. https://doi.org/10.1109/ICASSP.2019.8683198

[18] **Liu, Y., Thoshkahna, B., Milani, A., & Kristjánsson, T. (2020).** *Voice and accompaniment separation in music using self-attention convolutionalneuralnetwork.arXiv preprint arXiv:2003.08954*

[19] **Yamamoto, R., Shimamura, E., & Mitsufuji, Y. (2019).** *Phase-aware speech enhancement with deep complex U-Net.* In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 136–140), New Paltz, NY, USA.

[20] **Pascual, S., Serrà, J., & Bonafonte, A. (2019).** *Towards generalized speech enhancement with generative adversarial networks.*In **Interspeech 2019** (September 2019), pp. 1791–1795

[21] **Pascual, S., Bonafonte, A., & Serrà, J. (2017).** *SEGAN: Speech Enhancement Generative Adversarial Network.* In *Proceedings of Interspeech 2017* (pp. 3642–3646), Stockholm, Sweden. https://doi.org/10.21437/Interspeech.2017-1428

[22] **Défossez, A., Usunier, N., Bottou, L., & Bach, F. (2019).** *Music Source Separation in the Waveform Domain.* **arXiv preprint arXiv:1911.13254**

[23] **Wang,B.,&Chen,N.(2021).***An End-to-End Singing Voice Separation Model Based on Residual Attention U- Net.* **Journal of East China University of Science and Technology**, 47(5), 619–626.

[24] **Luo, Y., & Mesgarani, N. (2019).** *Conv- TasNet: Surpassing Ideal Time- Frequency Magnitude Masking for Speech Separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8), 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167

[25] **Takahashi, N., Goswami, N., & Mitsufuji, Y. (2018).** *MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation. arXiv preprint* arXiv:1805.02410

[26] Ansari, S., Alnajjar, K.A., Khater, T., Mahmoud, S. and Hussain, A., 2023. A robust hybrid neural network architecture for blind source separation of speech signals exploiting deep learning. IEEE Access.

[27] Nugraha, A.A., Liutkus, A. and Vincent, E., 2016. Multichannel audio source separation with deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(9), pp.1652-1664.

[28] Chandna, P., Miron, M., Janer, J. and Gómez, E., 2017. Monoaural audio source separation using deep convolutional neural networks. In Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings 13 (pp. 258-266). Springer International Publishing.

[29] Soni, S., Yadav, R.N. and Gupta, L., 2023. State-of-the-art analysis of deep learningbased monaural speech source separation techniques. IEEE Access, 11, pp.4242-4269.

[30] Wang, C., Jia, M. and Zhang, X., 2023. Deep encoder/decoder dual-path neural network for speech separation in noisy reverberation environments. EURASIP Journal on Audio, Speech, and Music Processing, 2023(1), p.41.

[31] Bregman, A.S., 1984, July. Auditory scene analysis. In Proceedings of the 7th International Conference on Pattern Recognition (pp. 168-175).

[32] FitzGerald, D., Cranitch, M. and Coyle, E., 2005. Non-negative tensor factorization for sound source separation.

[33] [15] Huang, P.S., Chen, S.D., Smaragdis, P. and Hasegawa-Johnson, M., 2012, March. Singing-voice separation from monaural recordings using robust principal component analysis. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 57-60). IEEE.

[34] Sprechmann, P., Bronstein, A.M. and Sapiro, G., 2012, October. Real-time Online Singing Voice Separation from Monaural Recordings Using Robust Low-rank Modeling. In ISMIR (pp. 67-72).

[35] Wichern, G., Antognini, J., Flynn, M., Zhu, L.R., McQuinn, E., Crow, D., Manilow, E. and Roux, J.L., 2019. Wham!: Extending speech separation to noisy environments. arXiv preprint arXiv:1907.01160.

[36] Chung, S.W., Choe, S., Chung, J.S. and Kang, H.G., 2020. Facefilter: Audio-visual speech separation using still images. arXiv preprint arXiv:2005.07074. [19] Rivet, B., Girin, L. and Jutten, C., 2007. Visual voice activity detection as a help for speech source separation from convolutive mixtures. Speech Communication, 49(7-8), pp.667-677.

[37] Xie, K., Jiang, K. and Yang, Q., 2021. Multi-channel underdetermined blind source separation for recorded audio mixture signals using an unmanned aerial vehicle. IET communications, 15(10), pp.1412-1422.

[38] Zhang, W., Tait, A., Huang, C., Ferreira de Lima, T., Bilodeau, S., Blow, E.C., Jha, A., Shastri, B.J. and Prucnal, P., 2023. Broadband physical layer cognitive radio with an integrated photonic processor for blind source separation. Nature communications, 14(1), p.1107.