Audio-only and Audio-Visual Source Separation: A Comparative Study

Nikolay Uvarov
FES, HSE, Moscow
nruvarov@hse.ru

Kirill Pribora
FES, HSE, Moscow
krpribora@hse.ru

Last Edit Date: November 24, 2025

***Abstract*—Recent progress in the field of source separation has shown that incorporating information from multiple modalities can significantly improve overall separation quality. In this paper, we present an empirical study of modern state-of-the-art audio-only and audio-visual speech separation models. We evaluate several architectures, i.e. ConvTasNet and DPTNet, with an emphasis on various ways of using visual modality. Our experiments demonstrate that applying different fusion techniques yields a significant boost in separation quality. Among the tested models, our audio-visual DPTNet consistently outperforms other solutions while preserving competitive computational efficiency.**

I. Introduction

Audio source separation (SS) aims to recover individual signals - known as sources - from their observed mixtures. Its applications vary from denoising to meeting transcription, hearing aids, music remixing and general speech enhancement. In multi-speaker scenarios, SS is particularly important, as overlapping sources can severely degrade performance of Automatic Speech Recognition (ASR) and speaker diarization systems. Existing methods are often broadly divided into blind source separation (BSS) and target source separation (TSS) approaches. BSS aims to recover all underlying sources from a mixture without any prior knowledge about the speakers, which provides flexibility but often is rather challenging in noisy conditions. In contrast, main focus of TSS models lies in extracting a single target speaker given some additional information (e.g., short speech example or visual cues such as lip-motion). This work primarily focuses on the BSS setting while leveraging ideas from TSS approaches through the use of visual modality, lip-motion specifically, and fusion techniques.

Training a BSS model is a non-trivial task, as one has to deal with different technical challenges. Since multiple sources are separated simultaneously, the model has no notion of which output corresponds to which target source. This permutation problem alone makes supervised training unstable because the loss depends on an inconsistent ordering of model outputs. To address this issue, permutation invariant training (PIT) [1], which evaluates the separation loss over all possible target-source permutations and chooses the ordering providing the minimum loss is implemented. By applying this method, the model optimization is independent from output ordering, thus providing stable training of BSS systems.

A large part of recent speech separation research has been focused on time-domain models, which directly work with the waveform without time-frequency representations. In contrast, spectrogram-based systems estimate masks over a spectrogram and reconstruct the waveform via inverse transforms. This solution introduces additional challenges such as sensitivity to empirical parameters (e.g., window size and hop length) and increasing computational requirements. As a result, frequency information is typically incorporated implicitly, either through learned time-domain filters [2] or via hybrid architectures that combine waveform processing with spectrogram cues [3].

First mentions of audio-visual learning date back to late 1990s, when researchers first observed that adding visual information, such as lip motion, could improve source separation. [4] These early studies showed that combining audio and visual information enhances separation quality in noisy conditions. In recent years, this topic has been gaining increasing attention, especially in context of speech separation. A common practice is to treat visual modality as an auxiliary cue and fuse it with audio either early, by combining features in the first layers, or late, by merging them in the prediction stage.

In this work, we primarily focus on early fusion techniques. We investigate several state-of-the-art architectures for two-speaker speech separation with the goal of maximizing the separation quality under limited computational resources. The remainder of this paper is organized as follows. Section II reviews existing SS models. Section III describes our fusion and training methodology. Section IV shows our experimental setup. Section V presents and discusses a series of experiments. Finally, Section VI concludes the paper and briefly overviews our work.

II. Related Work

Many advances in speech separation have been driven by time-domain architectures based on temporal convolutional networks (TCNs). These models work directly on the waveform and follow an encoder-separator-decoder pipeline.

ConvTasNet [5] is considered to be one of the most influential representatives of this family: it uses a learnable convolutional encoder to map the input mixture into a latent representation, apply a stack of TCN blocks to estimate source specific masks and then reconstructs the separated signals with a decoder. By working purely in the time domain, this design avoids explicit phase reconstruction and offers low latency, making ConvTasNet a strong baseline for speech separation.

Architectures based on recurrent neural networks (RNNs), specifically long short-term memory (LSTM) or gated recurrent units (GRU) have also played an important role in source separation. Early models mostly operated in the time-frequency domain by using bidirectional LSTMs to estimate spectral masks [6]. However such systems were limited by computational costs and sequence length, therefore they were not able to capture very long context. Dual-path recurrent neural networks (DPRNN) [7] adress this issue by splitting the encoded mixture representation into small overlapping chunks. Two recurrent modules, one by one, are applied: an intra-chunk RNN models local dependencies within each chunk whereas an inter-chunk RNN captures global context across all chunks. This dual-path architecture significantly improves model's receptive field without substantial increases in the amount of parameters or runtime.

Dual-path transformer networks (DPTNet) [8] further develop the dual-path design by replacing recurrent blocks with self-attention allowing the model to capture long-range dependencies more effectively. DPTNet, similar to DPRNN, splits the encoded mixture into overlapping chunks and uses intra-inter modules, but each stage is implemented with transformer layers instead of RNNs. This architecture improves the modeling of global context and leads to consistently better separation quality making DPTNet a stronger dual-path baseline for modern speech separation systems.

Audio-visual source separation builds on these audio-only architectures by incorporating synchronized video streams, mainly focusing on the speaker's face or lip region. This provides additional cues that are difficult to infer from audio alone. Early deep learning based audio-visual separation systems operated in the time–frequency domain and used CNNs to extract visual embeddings from the face or lip region. They were then aligned with the audio, broadcasted over the spectrogram and fused with audio features before being passed to a U-Net separator [9].

More recent approaches extend ConvTasNet and dual-path based designs to the audio-visual setting by injecting lip-based visual embeddings into the encoder–separator pipeline via early feature fusion or attention mechanisms [10]. These models consistently report improvements in separation quality, therefore they provide the main reference point for our experiments with early fusion in audio-visual BSS.

III. Methodology

A. Dataset All models were trained and validated on a provided two-speaker speech separation dataset consisting of 2-second-long mixture signals and the corresponding clean sources for each target speaker, all sampled at 16 kHz. Visual modality is represented by 50 frame videos of lip movement for each speaker in 96×96 resolution.

B. Training objective Main training objective lies in maximizing the scale-invariant signal-to-noise ratio (SI-SNR), which is calculated as:

$$\text{SI-SNR} = 10 \cdot \log_{10} \frac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{noise}}\|^2}, \tag{1}$$

where

$$\mathbf{s}_{\text{target}} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle}{\|\mathbf{s}\|^2} \mathbf{s}, \tag{2}$$

$$\mathbf{e}_{\text{noise}} = \hat{\mathbf{s}} - \mathbf{s}_{\text{target}}. \tag{3}$$

where $\hat{s}$ is the prediction and s is the target source.

C. Encoder As the video encoder, we use the pre-trained Multi-Scale TCN lip-reading network [11]. This encoder takes a sequence of 50 grayscale frames of size 96 ×96, i.e., a video clip. For each frame it produces a 512-dimensional embedding, and during training we keep this encoder frozen.

D. Fusion techniques Our main goal is to explore different ways of fusing audio and video embeddings, with linear fusion serving as our baseline fusion method.

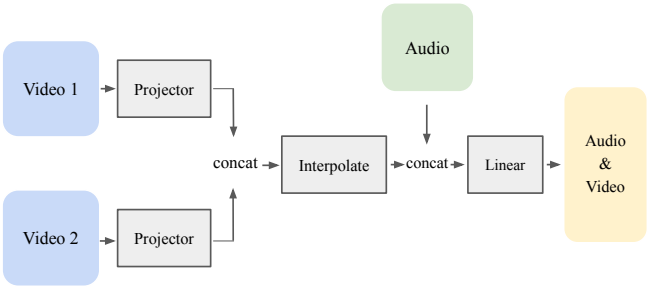


Figure 1. Linear Embedding Fusion

Our linear fusion module, illustrated in Fig. 1, takes the mixture audio embedding and two video embeddings as input. The video embeddings are passed through separate linear projectors that map them into a shared embedding space. The projected video embeddings are then concatenated, followed by a non-linear GELU activation function, and linearly interpolated to match the temporal length of the audio embeddings. At last, this interpolated video representation is concatenated with the audio embeddings and passed through a linear layer, producing a joint audio–visual embedding that is fed into the separation network.

Our second approach is similar to linear fusion but augments it by adding a gating mechanism with a residual connection.

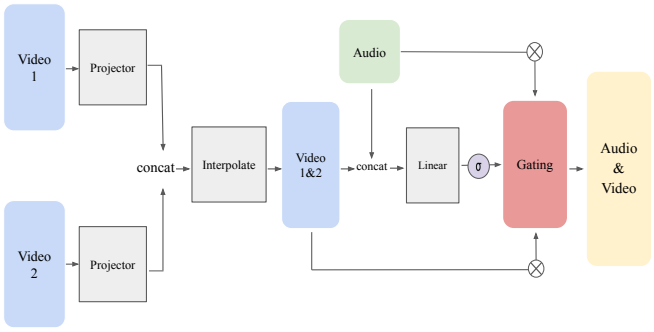


Figure 2. Gated Embedding Fusion

Figure 2 illustrates our gated fusion mechanism. It starts in the same way as linear fusion: audio and video embeddings are obtained, and the projected video representations are concatenated and aligned with the audio in time. The key difference is that after concatenating audio and video embeddings, we pass this representation through a sigmoid activation function to produce a gate, which we apply to audio and video embeddings to obtain a gated joint representation, which is followed by a residual connection to the original audio embedding. This method is expected to perform at least as well as the linear version.

Our final and most advanced approach is utilizing cross-modal attention mechanism to obtain embeddings enriched with information from another modality.

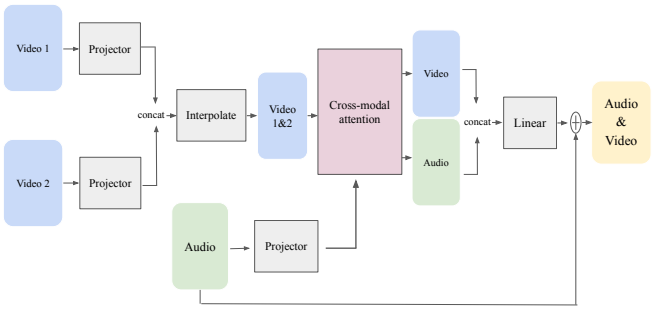


Figure 3. Attention Embedding Fusion

The fusion method shown in Fig. 3 uses the projected audio and video embeddings as queries and keys in an attention module, allowing each modality to be enriched with information. Resulting cross-modal embeddings are concatenated and are combined with the original audio embeddings via a residual connection to obtain highly informative joint representation. We expect this attention-based method to outperform other strategies.

D. Evaluation metrics Our main evaluation metric is the scale-invariant signal-to-noise ratio improvement (SI-SNRi), defined as:

$$\text{SI-SNRi} = \text{SI-SNR}(\hat{s}, s) - \text{SI-SNR}(\text{mix}, s) \quad (4)$$

This metric measures how much the model improves SI-SNR relative to simply using the mixture as the prediction. In addition, we report two widely used speech separation metrics: perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI).

IV. Experimental Setup

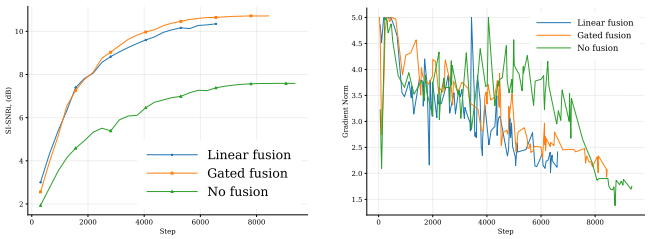
1. ConvTasNet ConvTasNet, implemented with the hyper-parameters from the original paper, serves as our baseline solution. We use the AdamW optimizer together with a OneCycleLR scheduler based on a cosine annealing strategy where max learning rate equals $1e-3$. All models were trained for 30 epochs with a batch size of 64 and a maximum gradient norm of 5, using early stopping when the validation loss reached a plateau. Since ConvTasNet is not a competitive state-of-the-art model, we chose not to use attention-based fusion in order to preserve computational resources and instead applied only linear and gated fusion methods.

2. DPTNet One of the main challenges we encountered when training DPTNet was that only specific combinations of hyperparameters (chunk size, hop size, encoder kernel size) would allow us to recover the same temporal length as the input audio. Therefore, we explicitly list the hyperparameters used in our experiments in our code implementation[1]. All DPTNet variants except the transformer-fusion model were trained for 50 epochs with a batch size of 16. We used the AdamW optimizer with a OneCycleLR scheduler and applied early stopping based on the validation loss, with a maximum gradient norm of 10. For all models that use fusion techniques, the encoder was configured with N = 128 filters, whereas the audio-only version used only 64 filters. Our largest model, DPTNet with cross-modal transformer fusion, was trained with a batch size of 8 due to its substantially higher computational requirements.

V. Results

The first outcome of our experiments is that incorporating visual information yields a substantial improvement in source separation performance compared to purely audio-only systems. This holds consistently across all architectures we tested. Moreover, all fusion-based models - except for the attention-based variant - not only achieved higher metrics, but also converged faster than the audio-only baseline, requiring fewer training steps to reach a comparable validation loss. We suggest that visual cues provide a strong inductive bias which stabilizes optimization.

[1]https://github.com/kre1ses/AVSS



(a) SI-SNRi (dB) (b) L2 Gradient Norm

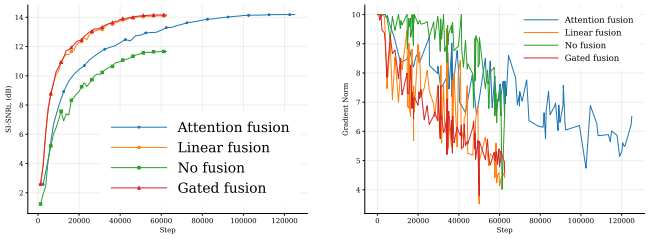
Figure 4. Training dynamics of ConvTasNet with different fusion strategies

Table I
CONVTASNET PERFORMANCE WITH DIFFERENT FUSION STRATEGIES

Fusion	Model size (M)	GMACS	SI-SNRi	PESQ	STOI
no	5.00	20.46	7.59	-	-
Linear	5.79	22.57	10.35	-	-
Gated	5.79	22.57	10.72	1.94	0.86

In addition to the observations above, we confirm that our initial hypothesis that gated fusion provides better results than linear fusion was correct, although the gap between them is smaller than we originally expected.

Our previous points were further confirmed when we evaluated the DPTNet models, which not only supported our conclusions but also generally outperformed ConvTasNet.



(a) SI-SNRi (dB) (b) L2 Gradient Norm

Figure 5. Training dynamics of DPTNet with different fusion strategies

Table II
DPTNET PERFORMANCE WITH DIFFERENT FUSION STRATEGIES

Fusion	Model size (M)	GMACS	SI-SNRi	PESQ	STOI
no	2.60	43.61	11.66	2.12	0.88
Linear	3.72	60.95	14.1	2.32	0.91
Gated	3.72	60.95	14.16	2.43	0.94
Attention	3.85	61.48	14.19	2.40	0.93

As shown in Table 2, all fusion strategies provide a substantial improvement over the audio-only DPTNet baseline. Incorporating visual information increases SI-SNRi from 11.66dB to around 14dB, and also leads to consistent gains in PESQ (from 2.12 to 2.32–2.43) and STOI (from 0.88 to 0.91–0.94). The differences between fusion methods are relatively small, but gated fusion achieves the best overall scores, with the highest PESQ (2.43) and STOI (0.94), while the attention-based variant attains the highest SI-SNRi (14.19dB). Gated fusion has the same computational cost as the linear variant but achieves higher scores, effectively providing an improvement in performance at no additional cost. In contrast, training the attention-based fusion was considerably more demanding in terms of memory, making the gated version the best trade-off between quality and efficiency.

As part of the efficiency contest, we also provide a lightweight version of our DPTNet model. By reducing the hidden dimensions and disabling bidirectional LSTMs, we were able to substantially decrease the model size while maintaining competitive separation quality.

- SI-SNRi - 11.05
- Model size (M) - 0.73
- GMACS - 11.73
- Model size (in MB) - 149.90
- Time per step - 0.036
- Throughput (in seconds) - 27.54
- Peak GPU memory usage (in GB) - 0.18

VI. CONCLUSION

In this work, we conducted an empirical comparison of audio-only and audio-visual speech separation models based on ConvTasNet and DPTNet, focusing on early fusion strategies. It was concluded that incorporating visual information consistently improved separation performance in terms of SI-SNRi and perceptual metrics such as PESQ and STOI. Among the fusion strategies, gated fusion was found to be the best trade-off between quality and efficiency, matching attention-based fusion at a lower computational cost. These results suggest that lightweight early fusion architectures may already utilize most of the benefits of audio-visual information, making them efficient candidates for practical deployment.

VII. ACKNOWLEDGEMENTS

This work was carried out as part of the Deep Learning in Audio course at HSE. We thank our instructors for their guidance and for providing the computational resources that made these experiments possible.

REFERENCES

[1] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," 2017. [Online]. Available: https://arxiv.org/abs/1607.00325

[2] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," 2018. [Online]. Available: https://arxiv.org/abs/1711.00541

[3] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," 2022. [Online]. Available: https://arxiv.org/abs/2211.08553

[4] Y. Nakagawa, H. G. Okuno, and H. Kitano, "Using vision to improve sound source separation," in *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, ser. AAAI '99/IAAI '99. USA: American Association for Artificial Intelligence, 1999, p. 768–775.

[5] 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, p. 1256–1266, Aug. 2019. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2019.2915167

[6] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[7] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," 2020. [Online]. Available: https://arxiv.org/abs/1910.06379

[8] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," 2020. [Online]. Available: https://arxiv.org/abs/2007.13975

[9] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, p. 1–11, Jul. 2018. [Online]. Available: http://dx.doi.org/10.1145/3197517.3201357

[10] D. Liu, T. Zhang, M. Christensen, Y. Wei, and Z. An, "Audio-visual fusion using multiscale temporal convolutional attention for time-domain speech separation," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2023-August. International Speech Communication Association, 2023, pp. 3694–3698, publisher Copyright: © 2023 International Speech Communication Association. All rights reserved.; 24th International Speech Communication Association, Interspeech 2023 ; Conference date: 20-08-2023 Through 24-08-2023.

[11] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," 2020. [Online]. Available: https://arxiv.org/abs/2001.08702

VIII. Contributions

Kolya: models, schedulers, metrics, losses, report, debugging, experiments brainstorming

Kirill: datasets, demo, readme, trainer/inferencer, project infrastructure, GPU resources, configs, debugging, experiments brainstorming