RecognAVSE-V2: An Improved Cross-Attention-Based Audio-Visual Speech Enhancement Approach

Leandro A. Passos, João R. Manesco, and João P. Papa
Dept. of Computing
São Paulo State University
Bauru, Brazil
leandro.passos@unesp.br

Rahma Fourati
REGIM-Lab.: REsearch Groups in Intelligent Machines, University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia.
Université de Jendouba, Faculté des Sciences Juridiques, Economiques et de Gestion de Jendouba, 8189, Jendouba, Tunisie.

Amir Hussain
School of Computing
Edinburgh Napier University
Edinburgh, U.K.

***Abstract*—Audio-visual speech enhancement (AVSE) is a subfield of machine learning that aims to improve speech quality from noisy audio signals, leveraging visual cues to guide information flow and enhance learning. In this context, the COG-MHEAR program supports an early competition, namely the AVSE Challenge (AVSEC), to advance the development of new solutions and AVSE methods while fostering the community. This paper introduces the RecognAVSE-V2, an improved and more efficient version of RecognAVSE proposed at AVSEC 2024, that implements a Time-Synced Cross-Attention Mechanism to exploit temporal correlations between audio and video features. The method's architecture comprises a video encoder that extracts spatiotemporal features from raw video frames, an STFT-based audio encoder that captures spectral features of the audio signal, and a time-synchronized cross-attention module that aligns audio and video features. Experimental results conducted over AVSE Challenge and CHiME3 datasets show RecognAVSE-V2 is capable of outperforming the baseline and its prior version in some cases while consisting of a model with reduced complexity, 20 times smaller, whose inference is 26% faster, and training demands a quarter of the total time.**

***Index Terms*—Audio-Visual Speech Enhancement, Cross-Attention Mechanism, Feature Fusion, Lip-Speech Synchronization, Audio-Video Temporal Alignment**

I. Introduction

THE last decades have witnessed exponential growth in computer-based and machine learning (ML) methods developed to improve the quality of life. In the context of speech enhancement (SE), such advancements may directly impact the lives of more than 430 million people who have hearing impairment to some degree [1], which also affects their social interactions [2] and psychological health [3].

Recent studies on SE have combined speech audio data with visual information in the so-called audio-visual speech enhancement (AVSE) [4]–[6]. Such an approach allows inferring knowledge extracted from the environment, e.g., reading lips or analyzing body movement, thus providing contextual visual information that contributes to a more substantial meaning to the spoken sentences.

Many efforts have been made to foster the AVSE community. In this context, the COG-MHEAR programme[1] promotes an annual challenge, namely COG-MHEAR Audio-Visual Speech Enhancement Challenge (AVSE Challenge) [7], which stimulates the development of novel, energy-efficient, multimodal approaches that fuse audio-visual information into robust mechanisms capable of extracting clean speech from boisterous environments, even considering using small processing devices.

In prior work, Manesco et al. [8], [9] proposed the RecognAVSE, a method presented in the AVSE Challenge 2024 that comprises an audio-visual speech enhancement architecture that combines the discriminative power of a Separable 3D Convolutional Neural Network (S3DNN) [10] for visual feature extraction with Deep Complex U-Net (DCU-Net) [11] for audio information encoding and decoding, such that the information from both modalities is fused using a cross-attention mechanism.

Apart from RecognAVSE's positive results, its attention mechanism, embedded within the DCU-Net architecture, could not fully exploit the temporal relationships between audio and visual features. Further, the attention mechanism was not explicitly designed to account for the time alignment between the audio and video streams, limiting the model's ability to exploit both modalities' sequential and temporal information. As a result, additional projections between audio and video features were required, but these projections did not fully address the need for precise temporal synchronization. Additionally, RecognAVSE imposes a high computational burden, which implies more expensive, robust, and specialized hardware for training such models, as well as considerable energy consumption and a larger carbon footprint.

To tackle such problems, this paper proposes RecognAVSE-V2, which extends RecognAVSE by introducing a time-

[1] https://cogmhear.org/

synchronized cross-attention mechanism that exploits temporal correlations between audio and video features. In a nutshell, RecognAVSE-V2 employs a video encoder module to extract spatiotemporal features from raw video frames and an STFT-based audio encoder to capture spectral features of the audio signal. Finally, the model uses a time-synchronized cross-attention mechanism to align audio and video features. Experiments conducted over CHiME3 [12] and AVSEC [7] datasets confirm the proposal's effectiveness for AVSE tasks, achieving competitive results with a less complex, faster, and 20 times smaller model. Thus, the contributions of this paper are provided as follows:

- to propose RecognAVSE-V2, a novel multimodal approach for AVSE;
- to overcome the issues observed in our prior method, i.e., RecognAVSE, regarding the inability to exploit temporal relationships between audio and visual features thoroughly;
- to provide a smaller, faster, and more efficient model capable of outperforming prior results in some cases with a reduced and less complex solution.

The remainder of this paper is described as follows. Section II introduces the related works, while Section III defines the proposed RecognAVSE-V2. Further, Section IV describes the datasets and the methodology. Furthermore, experimental results and discussions are provided in Section V. Finally, the conclusions are stated in Section VI.

II. Related Work

This section provides a brief literature review concerning audio-visual speech enhancement. In this context, Chuang et al. [13] proposed a lightweight model, iLAVSE, to address privacy concerns in facial data. Experimental results showed that iLAVSE could overcome AVSE systems on these issues, also serving as an alternative for real-world scenarios where high-quality audio-visual sensors may not always be available.

Recently, Foroushi et al. [14] introduced the AV-RVAE, a variational autoencoder-based model for AVSE that fuses visual lip motion extracted through a dedicated video encoder with the audio, using a recurrent neural network for temporal modeling. The method outperformed existing methods, showing improvements in several metrics. Meanwhile, Zheng et al. [15] incorporated ultrasound tongue images to improve the performance of lip-based AVSE systems. The authors employed knowledge distillation to obtain tongue-related information without directly inputting ultrasound tongue images. Further, the model aligned lip and tongue modalities, creating a lip-tongue key-value memory network that enables the retrieval of tongue features from lip features. Experimental results demonstrate improvements over traditional lip-based AVSE baselines.

Finally, the AVSE community has witnessed rapid growth in the field due to the work promoted by the AVSE Challenge [7], which encourages the development of dozens of novel, high-quality solutions each year. Among such solutions, one can refer to the work proposed by Gogate et al. [16], which implements an end-to-end framework for low-latency real-time AVSE that combines a lightweight visual feature extraction network and a recurrent neural network to separate the speakers' voices from interferences, demonstrating substantial improvement compared to the state-of-the-art AVSE models. Wahab et al. [17] also proposed an interesting model for the challenge using a multimodal dual-transformer architecture that employs the attention mechanism to capture correlations between modalities, obtaining satisfactory results. Meanwhile, Fourati et al. [18] proposed a binary multiobjective particle swarm optimization-based approach called AVSE-Pruner that is capable of learning efficient pruning strategies dynamically, optimizing multimodal Convolutional Neural Network (CNN)-based models for the task of AVSE, maintaining high performance while significantly reducing computational burden for real-time embedded applications. Later on, Fourati et al. [19] proposed an efficient and sustainable AVSE method using a latency-aware pruning approach. Apart from such contributions, RecognAVSE [8], [9], which served as the basis for the method developed in this work, was also proposed for the challenge.

III. Proposed Method

This section briefly describes the RecognAVSE [8], the basis upon which RecognAVSE-V2 was built, as well as the proposed RecognAVSE-V2.

A. RecognAVSE

RecognaAVSE was designed to solve AVSE problems by employing a DCU-Net to enhance audio signals and an S3DNN-based approach to capture relevant visual information correlated with speech. It effectively combines information from audio and video features using a similar strategy adopted by [20]

Further, the model implements a cross-attention mechanism to connect the innermost layer of the DCU-Net with the latent embedding of the S3DCNN. This mechanism integrates contextual visual information with audio features, thus enhancing speech clarity and intelligibility. Such a mechanism ensures the contextual alignment of information from distinct modalities, thus enhancing audio signals based on visual context.

B. Proposed RecognAVSE-V2

This paper introduces RecognAVSE-V2, an AVSE approach built to overcome the limitations observed in RecognAVSE. RecognAVSE-V2 introduces a Time-Synced Cross-Attention Mechanism explicitly designed to exploit the temporal correlations between audio and video features by directly modeling the temporal relationships between the two modalities, enabling the model to attend to video frames that are temporally relevant for each audio frame. This approach ensures that the model can fully exploit the temporal alignment between the audio and video streams without requiring intermediary projections between the features. Introducing this temporal attention architecture improves the accuracy and effectiveness

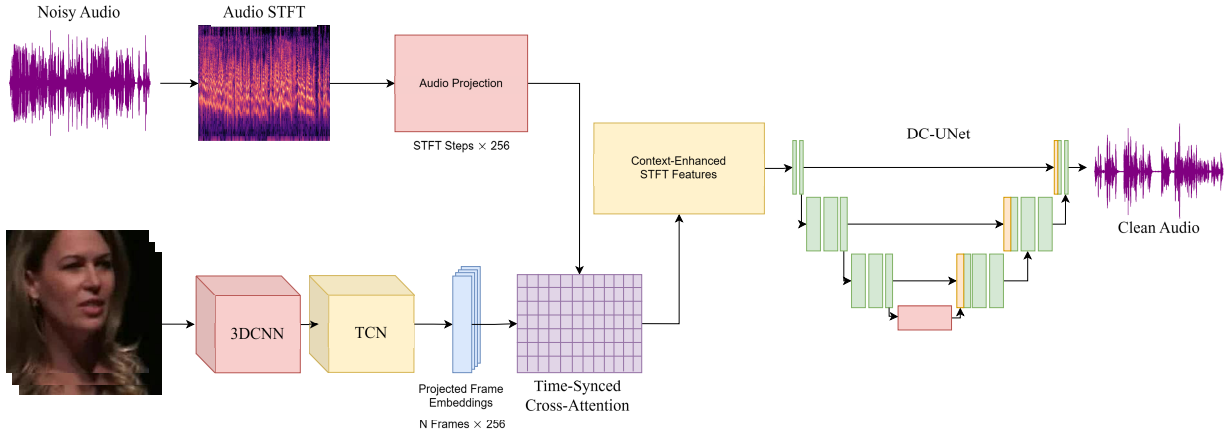

Fig. 1. Depiction of the overall pipeline of our method: RecognAVSE-V2. A TCN network extracts video embeddings. To combine audio and video features, our time-synced cross-attention mechanism is used between audio and video projections in order to obtain a context-enhanced version of the audio STFT that a DCU-Net can use for audio filtering.

of multimodal fusion, resulting in a more robust model for tasks requiring precise audio-visual synchronization.

RecognAVSE-V2's architecture consists of three key components: (i) a video encoder that extracts spatiotemporal features from raw video frames, (ii) an audio encoder that captures spectral features of the audio signal via the Short-Time Fourier Transform (STFT), and (iii) a time-synced cross-attention module that enables explicit temporal alignment between the audio and video features.

The video encoder processes the input video sequence and produces latent representation $V \in \mathcal{R}^{B\times T_v\times d_v}$, where B is the batch size, T_v is the number of video frames, and d_v is the dimensionality of the video features.

The audio signal is transformed into a complex spectrogram using STFT, resulting in a tensor $A \in \mathcal{R}^{B\times T_a\times 2F}$, where T_a is the number of STFT frames and $2F$ represents the real and imaginary components of each frequency bins.

The main contribution to this work is the Time-Synced Cross-Attention Mechanism, which directly synchronizes the audio and video features in the temporal domain. Such module facilitates the fusion of the two modalities by attending to relevant video frames for each audio frame within a localized temporal window. The whole pipeline of the method is illustrated in Figure 1

C. Time-Synced Cross-Attention

The Time-Synced Cross-Attention Mechanism treats audio features as queries Q and video features as keys K and values V. Using the scaled dot-product attention, the attention scores are computed as:

$$\text{Attention Scores}_{ij} = \frac{Q_i \cdot K_j^\top}{\sqrt{\frac{d}{h}}},$$

where d is the dimensionality of the feature space, and h is the number of heads in the multi-head attention mechanism. The attention scores are then used to compute the attention probabilities, which are applied to the video features to produce the context vector for each audio frame.

To ensure that each audio frame attends to the relevant video frames in the temporal domain, we introduce a temporal attention mask $\mathbf{M} \in \{0,1\}^{T_a\times T_v}$. This mask restricts the attention mechanism to only consider video frames within a local temporal window around each audio frame.

Given the difference in temporal resolution between the audio and video streams, we compute the corresponding video frame for each audio as follows:

$$\text{center\_frame}_i = \text{int}\left(\frac{i\cdot \text{window\_shift}}{\text{sampling\_rate}}\cdot \text{video\_fps}\right), \quad (1)$$

where i represents the STFT frame index, window_shift is the hop length, sampling_rate is the audio sample rate, and video_fps is the video frame rate.

For each audio frame, we define a temporal window $[-w,+w]$ centered around the computed center_frame$_i$, within which video frames are attended to:

$$\mathbf{M}_{i,j} = \begin{cases} 1 & \text{if } |j - \text{center\_frame}_i| \le w \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Such a procedure ensures that the attention mechanism only attends to video frames that are temporally relevant to each audio frame. The relationship is represented in Figure 2, which illustrates how each frame attends each STFT step in the audio.

The temporal mask is applied to the attention scores during the attention computation. Scores corresponding to irrelevant video frames (where $\mathbf{M}_{i,j} = 0$) are set to $-\infty$, preventing them from contributing to the final attention probabilities.

$$\text{Attention Scores}_{ij} = \begin{cases} \text{Attention Scores}_{ij} & \text{if } \mathbf{M}_{i,j} = 1 \\ -\infty & \text{if } \mathbf{M}_{i,j} = 0. \end{cases} \quad (3)$$

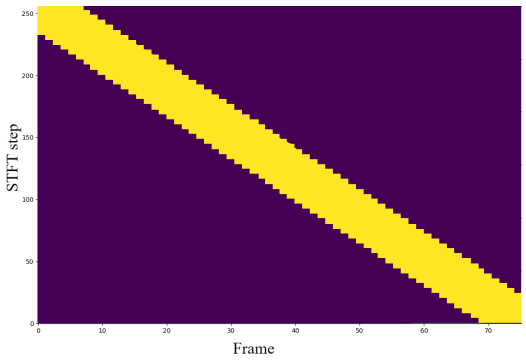

Fig. 2. Visualization of the temporal attention mask applied during audio-visual alignment. The mask ensures that each audio frame (STFT step) attends only to the temporally relevant video frames, restricting attention to a diagonal band in the matrix. Video frames outside this region have masked attention scores, preventing irrelevant contributions to the final attention probabilities.

IV. Methodology

This section presents the datasets and the experimental setup considered in the experiments.

A. AVSE Challenge Dataset

The AVSE challenge dataset [7] aims to enhance a target speech signal mixed with a competing speaker or noise, given the audio and visual information of the target speaker. The dataset is built upon the LRS3 dataset [21], which comprises thousands of sentences/phrases randomly selected from TED and TEDx lectures. The audio post-processing consists of a resampling at 16 kHz and 16 bits to further introduce the interference to audio. In short, the dataset task consists of reducing or removing this interference.

The interferer insertion is conducted by combining the audio extracted from three distinct datasets: (i) the DNS Challenge dataset [22], comprising environmental sounds[2], (ii) the DEMAND dataset [23], which consists of noise from soundscapes, and (iii) Clarity Challenge dataset [24], composed of domestic noises.

The AVSE challenge dataset is composed of disjoint subsets for training and development. Table I outlines an overview of the data distribution, such that each mix regards a single sentence mixed with an interference signal, i.e., competing speaker and noise at a specific signal-to-noise ratio.

TABLE I
Training and Development Sets Distribution.

Sets	# Mixes	# Target Speakers	Interferers
Train	34,524	605	405 competing speakers and 7,346 noise files.
Dev	3,306	85	30 competing speakers and 1,825 noise files.

[2]Environmental sounds extracted from https://freesound.org/.

B. CHiME3 Dataset

CHiME3 [12] is an AVSE dataset based on the benchmark audio-visual GRID corpus. It comprises a range of audiovisual vectors in 2D discrete cosine transform visual features, whose visual vectors were extracted from the lip region of five speakers (two females and three males) over 1,000 videos each. The audio vector was extracted by windowing the audio signal and transforming each frame into a log-filterbank vector. The visual signal was then interpolated to match the audio, and several large datasets were created, with the frames shuffled randomly to prevent bias and with different pairings. Table II provides a summary of the dataset.

TABLE II
Grid Corpus Aligned Sentences Distributed by Speakers and Divided into Train, Validation, and Test.

Speaker	Removed	Used	Train	Val	Test
Speaker 1	11	989	692	99	198
Speaker 2	164	836	585	84	167
Speaker 3	71	929	650	93	186
Speaker 4	9	991	693	99	199
Speaker 5	11	989	692	99	198
All	266	4,734	3,312	474	948

C. Experimental Setup

RecognAVSE-V2 was trained using videos of 75 frames, resized to an image size of 224x224. The model architecture was designed to extract temporal and spatial features from video sequences efficiently. It begins with a 3D convolutional frontend, which applies a $5\times7\times7$ kernel over the input frames to capture spatiotemporal dependencies. This layer reduces the feature dimensionality while retaining important motion information.

After the initial convolutional layers, the extracted feature maps are passed through a Temporal Convolutional Network (TCN), responsible for capturing long-range dependencies across frames. The TCN consists of three layers with progressively increasing channels (64, 128, 256), resulting in a set feature embedding of size 256 for each frame for the output, which is then used for multimodel alignment as described in section III-B.

The audio used to produce the enhanced context was divided into random clips of 40,800 samples, randomly clipped for each video during training, and sampled at 16 kHz. The audio waveform was then converted to the time domain using the Fast Fourier Transform (FFT), utilizing a Hann window of size 400, a window shift of 160, and a hop length of 512. After obtaining the contextual information from the video and fusing information in the attention layer, a DCU-Net was used to filter the noise. The DCU-Net was built with five layers of downsampling and five layers of upsampling. The last downsampling layer of the network was flattened and then used to combine and integrate video features through the cross-attention layer.

The experiments used a modified version of the Scale-Invariant Signal-to-Noise Ratio (SI-SNR) loss [25] as the loss

function. The original loss is a widely used metric for evaluating and optimizing the performance of audio enhancement models. This metric measures the similarity between the clean target signal and the enhanced signal produced by the model by decomposing the enhanced signal into two components: one aligned with the target signal (the desired component) and one orthogonal (the noise component). The SI-SNR loss is then defined as the ratio between the magnitude of each element, providing a reliable measure of the model's performance.

In our work, we used a modified version of the loss to include a soft limit for the SNR loss values [26]. The primary objective of this modification is to prevent well-separated examples from dominating the training process. This is achieved by incorporating a parameter τ corresponding to a maximum SNR value (SNRmax), which acts as a soft threshold. The modified loss function is defined as:

$$L(y, \hat{y}) = -10 \log_{10} \left(\frac{\|y\|^2}{\|y - \hat{y}\|^2 + \tau \|y\|^2} \right),$$

where τ is set to $10^{-\text{SNRmax}/10}$. We empirically set SNRmax to 30 dB to balance training stability and performance. By applying this threshold, the loss function effectively limits the influence of examples where the SI-SNR exceeds this value, ensuring that the model focuses on more challenging examples during training. Our loss function was also mean-normalized to maintain scale invariance, ensuring stability across different input levels.

Concerning the evaluating metrics, this work presents the results considering the Short-Time Objective Intelligibility (STOI) [27], which measures the intelligibility Setupof degraded speech signals, the Perceptual Evaluation of Speech Quality (PESQ), which measures audio quality considering its sharpness, background noise, clipping, and audio interference, and the scale-invariant signal-to-distortion ratio (SISDR) [28], which considers the scaling of the separated sources to measure the quality of a source sound.

Our results were compared against the previous version, i.e., RecognAVSE, as well as the baseline provided by the 3rd COG-MHEAR Audio-Visual Speech Enhancement Challenge, which will be referred to as *baseline* in our comparisons. In addition, we also provide a comparison to the original noisy audio without any processing, referred to as *noisy*, during our analysis.

Further, the parameters are optimized using the ADOPT optimizer [29] with a learning rate of 10^{-3}, reduced by a factor of 0.5 on a plateau with the patience of 10 during 100 epochs considering a batch size of 16. RecognAVSE-V2 is implemented in Python using Pytorch framework [30] and the code is available in GitHub[3]. The model comprises 8M parameters, 11.45 GFlops, and presents an average training time of 1h43m per epoch running on an Intel® Xeon® Bronze 3204 CPU with 1.90GHz, 48GB of RAM, and a Tesla T4 Nvidea GPU with 16GB of memory, a reduction from the original RecognAVSE version, containing 164M parameters, 12.81GFlops and an average training time of 3h16m.

[3]Available at: hidden due to double-blind policy.

V. Experimetal Results

This section evaluates the effectiveness of RecognAVSE-V2 using quantitative metrics and qualitative analysis of the spectrograms, considering the CHIME3 and AVSE Challenge datasets. It also compares the architectures of RecognAVSE-V2 and RecognAVSE.

A. CHIME3

In this context, Table III compares the proposed RecognAVSE-V2 against its prior version, i.e., RecognAVSE, and the noisy set over the CHIME3 dataset.

TABLE III
QUANTITATIVE RESULTS OBTAINED ON THE TEST SET OF THE CHIME DATASET, COMPARING THE PERFORMANCE OF RECOGNAVSE-V2 AGAINST THE NOISY AUDIO AND RECOGNAVSE USING PESQ, STOI, AND SISDR METRICS. BOLD INDICATES THE BEST RESULTS.

Method	PESQ↑	STOI↑	SISDR↑
Noisy	1.13	0.35	−45.13
RecognAVSE	1.36	0.56	1.75
RecognAVSE-V2	**1.46**	**0.59**	**4.97**

Such results assert the robustness of the model and the proposed Time-Synced Cross-Attention Mechanism, whose fusion of temporally correlated audio and visual features enabled the achievement of the highest values over all three metrics.

Figure 3 allows a visual inspection of the RecognAVSE-V2 spectrogram over the CHiME3 dataset. The method is compared against the spectrograms of RecognAVSE, the clean audio, the noise, and the noisy signal composed by the clean audio interfered with the noise. Such images show that RecognAVSE-V2 is more efficient in removing noise and irrelevant signals than RecognAVSE. On the other hand, it performs a more incisive signal cut, also pruning high frequencies in the clean signal.

B. AVSE Challenge

Table IV compares RecognAVSE-V2 against RecognAVSE, the baseline, namely AVSEC [7], and the noisy audio over the development set of AVSE Challenge Dataset.

TABLE IV
QUANTITATIVE RESULTS OBTAINED ON THE DEVELOPMENT SET FOR THE AVSE CHALLENGE DATASET, COMPARING THE PERFORMANCE OF RECOGNAVSE-V2 AGAINST RECOGNAVSE, THE NOISY AUDIO, AND THE BASELINE USING STOI, SISDR, AND PESQ METRICS. BOLD INDICATES THE BEST RESULTS.

Method	PESQ↑	STOI↑	SISDR↑
Noisy	1.16	0.62	−4.33
AVSEC [7]	1.33	0.68	2.13
RecognAVSE	1.32	**0.70**	**2.35**
RecognAVSE-V2	**1.36**	0.61	−0.02

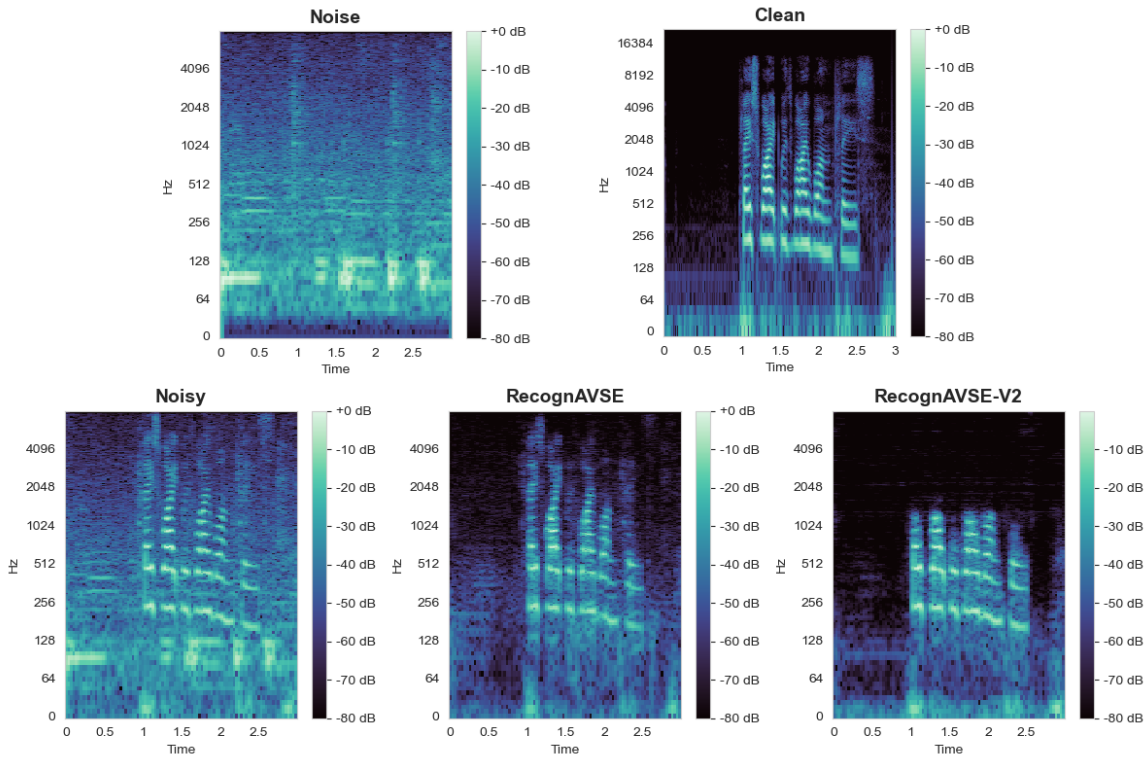

Fig. 3. CHiME3 dataset sample spectrograms of noise and clean audio (Top), noisy audio (clean audio + noise), RecognAVSE, and RecognAVSE-V2 predictions (Bottom).

Such results show that RecognAVSE-V2 outperformed all techniques concerning the PESQ while presenting lower results over the STOI and SISDR metrics. Such behavior is likely attributed to the Time-Synced Cross-Attention fusion mechanism, which shows itself very efficient for noise removal at the cost of reducing the intelligibility of some pronounced words due to the partial prune of high-frequency information from clean audio. The result also shows that the AVSEC dataset poses a more challenging scenario than CHiME3 since it comprises noise and multiple speaker interference, making it harder for the model to separate the main speaker's voice from background noise and voices.

Similar behavior is observed in the test set results, provided in Table V. Again, RecognAVSE-V2 obtained the most accurate results considering the PESQ metric, reinforcing the ability to prune noise while, on the other hand, maintaining artifacts of additional speakers that compromise the prediction clarity.

Figure 4 compares the RecognAVSE-V2 and RecognAVSE spectrograms over the AVSE Challenge dataset. RecognAVSE-V2 depicts similar behavior to RecognAVSE until frequencies slightly above $1,024$ Hz. Higher frequencies are pruned by RecognAVSE-V2, which probably impacted the results and the low values of STOI and SISDR over this dataset.

TABLE V
QUANTITATIVE RESULTS OBTAINED ON THE TEST SET FOR THE AVSE CHALLENGE DATASET, COMPARING THE PERFORMANCE OF RECOGNAVSE-V2 AGAINST RECOGNAVSE, THE NOISY AUDIO, AND THE BASELINE USING PESQ, STOI, AND SISDR METRICS. BOLD INDICATES THE BEST RESULTS.

Method	PESQ↑	STOI↑	SISDR↑
Noisy	1.17	0.61	−4.88
AVSEC [7]	1.29	0.65	**0.80**
RecognAVSE	1.28	**0.65**	0.40
RecognAVSE-V2	**1.31**	0.63	−1.32

C. Computational Complexity

The computational efficiency of RecognAVSE-V2 is computed in terms of the number of parameters (Par #), Giga Flops per second (GFlops), inference time (IT), the number of Multiply-Accumulate operations (MAC), whose higher values indicate more complex models, and the Peak Memory Usage Inference over the inference (PMUi) and the training (PMUt). The IT, PMUi, and PMUt metrics were computed using the AVSE Challenge dataset.

Table VI shows that RecognAVSE-V2 is more than 20 times smaller than RecognAVSE in parameter numbers and requires almost 11% less Floating-point Operations Per Second. It also presents an inference time reduced by 26% and a MAC 10.6% smaller. Concerning memory usage, RecognAVSE-V2

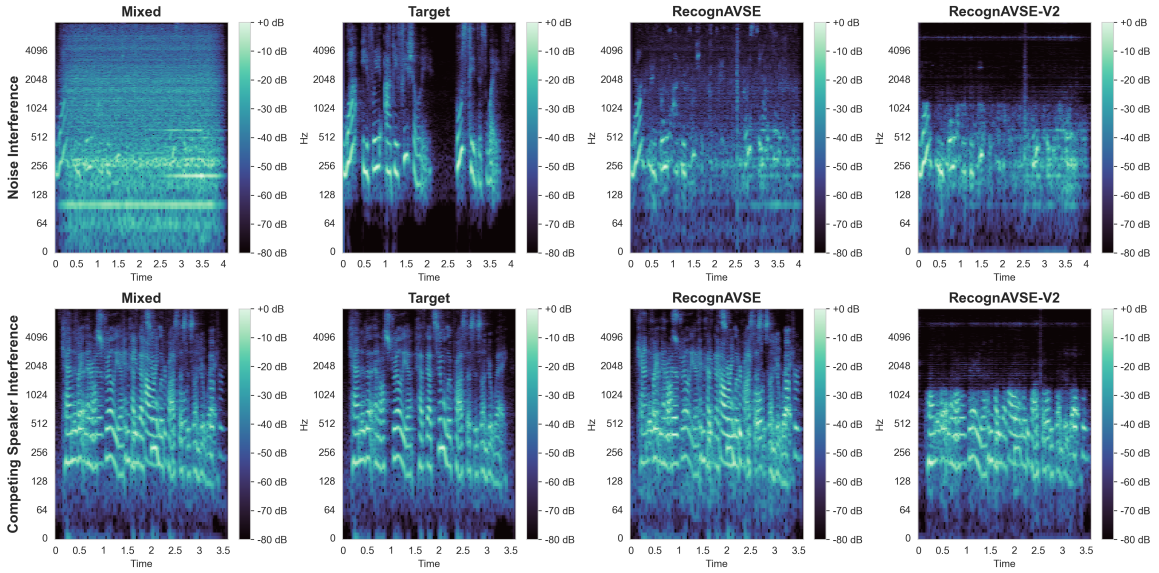

Fig. 4. Comparison of Spectrograms computed over the AVSE dataset considering noise (top) and competing speaker (bottom) interference. The first column (Mixed) illustrates the spectrogram of a target signal mixed with the interference, while the second column (Target) depicts the expected clean signal. The third and fourth columns present the spectrograms of the outputs generated by RecognAVSE and RecognAVSE-V2, respectively.

outperforms its prior version in terms of PMU by showing itself $3,6$ times more efficient during training and inference.

Such results reinforce the model's suitability. It outperformed the prior version in many scenarios and obtained similar results in others while requiring a fraction of the computational resources.

TABLE VI
COMPLEXITY COMPARISON.

Method	Par #↓	GFlops↓	IT↓	MAC↓	PMUi↓	PMUt↓
RecognAVSE	164M	12.81	89.85	6.40	1058.97	2585.60
RecognAVSE-V2	8M	11.45	66.20	5.72	293.63	729.79

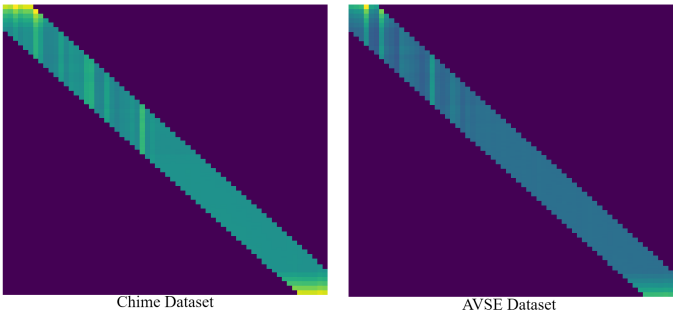

Fig. 5. Comparison of average attention weights learned by RecognAVSE-V2 on the CHiME3 dataset (left) and the AVSE dataset (right). While both exhibit a temporal alignment pattern, the AVSE dataset shows a more uniform and smoothed attention distribution, suggesting weaker adaptability to complex signal variations.

D. Attention Scores

Figure 5 compares the average attention scores learned by RecognAVSE-V2 over the CHiME3 and AVSE datasets. While both datasets exhibit a strong diagonal structure, ensuring temporal alignment with the attention mask, there are key differences in attention distribution. The CHiME3 dataset demonstrates sharper variations in attention weights, suggesting that the model dynamically adjusts its focus to handle background noise. On the other hand, the attention score computed over the AVSE dataset shows a more diffused and uniform pattern, indicating a reduced capacity to prioritize informative frames. This is likely due to the nature of the AVSE dataset, which includes concurrent speakers as a noise source. Unlike stationary or environmental noise, overlapping speech presents a more complex interference pattern, making it harder for the model to separate the primary speaker from competing voices.

VI. CONCLUSION

This work proposes RecognAVSE-V2, an audio-visual speech enhancement architecture built upon RecognAVSE that implements an attention mechanism capable of overcoming RecognAVSE's limitations related to sequential and temporal information and the computation burden. RecognAVSE-V2 implements a light architecture that exploits the temporal relationships between audio and visual features and promotes the alignment between the audio and video streams, enabling the model to exploit both modalities' sequential and temporal information.

Experimental results showed that RecognAVSE-V2 outperformed its prior version and the baseline when considering the PESQ, STOI, and SISDR metrics over the CHiME3 dataset. It also obtained the most accurate results when considering the STOI metric over the AVSE Challenge dataset. Such results

were obtained when considering a less complex, faster, and 20 times smaller architecture.

Regarding future work, we aim to investigate the attention mechanism further, particularly focusing on the normalization aspects of the data, which we suspect may contribute to performance degradation in frequencies above 1024 Hz. Given the observed attention distribution differences across datasets, we plan to analyze whether normalization inconsistencies between the training and evaluation data affect the model's ability to attend to high-frequency components effectively.

ACKNOWLEDGMENTS

João R. Manesco, Leandro A. Passos, and João P. Papa are grateful to the São Paulo Research Foundation (FAPESP) grants $2025/13172-1$, $2024/00789-8$, $2013/07375-0$, $2023/14427-8$, and $2023/01374-3$, as well as the National Council for Scientific and Technological Development (CNPq) grant $308529/2021-9$ for their financial support. Rahma Fourati has received funding from the Ministry of Higher Education and Scientific Research of Tunisia under grant agreement number LR11ES48. Amir Hussain acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC) (Grants No. EP/M026981/1, EP/T021063/1, EP/T024917/1).

REFERENCES

[1] W. H. Organization *et al.*, *Hearing screening: Considerations for implementation*. World Health Organization, 2021.

[2] W. Noble, *Self-assessment of hearing and related function*. Wiley-Blackwell, 1998.

[3] A. R. Huang, J. A. Deal, G. W. Rebok, J. M. Pinto, L. Waite, and F. R. Lin, "Hearing impairment and loneliness in older adults in the united states," *Journal of Applied Gerontology*, vol. 40, no. 10, pp. 1366–1371, 2021.

[4] L. A. Passos, J. P. Papa, A. Hussain, and A. Adeel, "Canonical cortical graph neural networks and its application for speech enhancement in audio-visual hearing aids," *Neurocomputing*, vol. 527, pp. 196–203, 2023.

[5] L. A. Passos, J. P. Papa, J. Del Ser, A. Hussain, and A. Adeel, "Multimodal audio-visual information fusion using canonical-correlated graph neural network for energy-efficient speech enhancement," *Information Fusion*, vol. 90, pp. 1–11, 2023.

[6] M. Raza, L. A. Passos, A. Khubaib, and A. Adeel, "Multimodal speech enhancement using burst propagation," *arXiv preprint arXiv:2209.03275*, 2022.

[7] A. L. A. Blanco, C. Valentini-Botinhao, O. Klejch, M. Gogate, K. Dashtipour, A. Hussain, and P. Bell, "Avse challenge: Audio-visual speech enhancement challenge," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 465–471.

[8] J. R. R. Manesco, L. A. Passos, R. Fourati, J. Papa, and A. Hussain, "RecognAVSE: An audio-visual speech enhancement approach using separable 3d convolutions and deep complex u-net," in *3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC)*, 2024, pp. 11–15.

[9] J. R. Manesco, L. Passos, R. Fourati, J. P. Papa, and A. Hussain, "Tackling reverberation and binaural data on audio-visual speech enhancement through recognavse," in *Proc. AVSEC 2025*, 2025, pp. 6–9.

[10] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 305–321.

[11] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[12] A. Adeel, M. Gogate, and A. Hussain, "Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments," *Information Fusion*, vol. 59, pp. 163–170, 2020.

[13] S.-Y. Chuang, H.-M. Wang, and Y. Tsao, "Improved lite audio-visual speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1345–1359, 2022.

[14] Z. Foroushi and R. Dansereau, "Dynamic audio-visual speech enhancement using recurrent variational autoencoders," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2024, pp. 60–64.

[15] R.-C. Zheng, Y. Ai, and Z.-H. Ling, "Incorporating ultrasound tongue images for audio-visual speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[16] M. Gogate, K. Dashtipour, and A. Hussain, "A lightweight real-time audio-visual speech enhancement framework," in *Proc. AVSEC 2024*, 2024, pp. 19–23.

[17] F. Wahab, N. Saleem, A. Hussain, R. Ullah, and M. B. Hossen, "Multi-model dual-transformer network for audio-visual speech enhancement," in *Proc. AVSEC 2024*, 2024, pp. 1–5.

[18] R. Fourati, J. Tmamna, N. Kouka, M. Gogate, K. K. Dashtipour, L. A. Passos, J. Papa, T. Arslan, and A. Hussain, "AVSE-Pruner: Filter pruning of audio-visual speech enhancement system using multi-objective binary particle swarm optimization," in *3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC)*, 2024, pp. 24–29.

[19] R. Fourati, J. Tmamna, J. R. Manesco, L. A. Passos, J. P. Papa, and A. Hussain, "Efficient and sustainable audio-visual speech enhancement through latency-aware pruned model," in *Proc. AVSEC 2025*, 2025, pp. 10–13.

[20] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer *et al.*, "Perceiver io: A general architecture for structured inputs & outputs," in *International Conference on Learning Representations*, 2021.

[21] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[22] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6623–6627.

[23] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. AIP Publishing, 2013.

[24] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. Viveros Munoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2, 2021.

[25] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[26] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised speech separation using mixtures of mixtures," in *ICML 2020 Workshop on Self-supervision in Audio and Speech*, 2020.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[28] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[29] S. Taniguchi, K. Harada, G. Minegishi, Y. Oshima, S. C. Jeong, G. Nagahara, T. Iiyama, M. Suzuki, Y. Iwasawa, and Y. Matsuo, "Adopt: Modified adam can converge with any β_2 with the optimal rate," in *Advances in Neural Information Processing Systems*, 2024.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.