Lip-Reading-Guided Speech Enhancement via Self-Aligning Cross-Attention Networks

Shahid Mukhtar[1]*, G.C. Kingdone[2]

[1]*University of Alabama at Birmingham, USA*

[2]*Robotics and Automation Laboratory Universidad PrivadaBoliviana Cochabamba, Bolivia*

KEYWORDS:
Lip-reading, Speech enhancement,
Audio-visual fusion,
Self-aligning cross-attention networks,
Multimodal speech processing,
Temporal alignment,
Deep learning,
Noise-robust ASR

ARTICLE HISTORY:
Submitted : 15.07.2025
Revised : 20.08.2025
Accepted :10.09.2025

https://doi.org/10.17051/NJSAP/01.04.03

Abstract

One challenge is speech enhancement in noise, which is of great importance in enhancing communication in real world scenarios like teleconferencing, using hearing aids, and automatic speech recognition (ASR). Although recently developed audio-visual speech processing methods have shown that visual information of the lips motion of a speaker can be used to enhance noise robustness to a high degree, temporal misalignment between the audio and video data has remained a performance limit. In this paper, a novel Lip-Reading-Guided Speech Enhancement architecture is presented relying on Self-Aligning Cross-Attention Networks (SACAN), and serving to dynamically synchronize and allure together multi-modal features to recover clearer speech. The two streams of visual data and audio data are processed through a spatio-temporal convolutional encoder to capture discriminative features of lip movements and are encoded through log-mel spectrogram encoding to achieve representations in spectral and time dimensions. These features are adapted to align frame-wise with a bidirectional self-aligning cross-attention mechanism that helps mitigate distortions caused by latencies and articulation mismatch between modalities. A U-Net based enhancement network is used to decode the fused representation to produce a clean speech spectrogram that in turn is reconstructed into waveform through inverse short-time Fourier transform. The GRID and LRS3-TED datasets are experimented on under three plausible conditions of noise (babble, street and cafe) at various signal-to-noise ratio (SNR) levels. PESQ, STOI, and WER comparisons of quantitative assessments show that SACAN is evaluated to have a PESQ gain of 0.41, a STOI gain of 0.05 and WER reduction of 17.3 percent against state-of-the-art audio-only enhancement baselines. Improved speech naturalness and intelligibility is further verified by subjective listening tests. The results demonstrate the usefulness of cross-modal temporal matching in reliable multimodal speech enhancement and its feasibility to be applied in realtime hostile communication conditions.

Author's e-mail: smukhtar@uab.edu

How to cite this article: Mukhtar S, Kingdone GC. Lip-Reading-Guided Speech Enhancement via Self-Aligning Cross-Attention Networks. National Journal of Speech and Audio Processing, Vol. 1, No. 4, 2025 (pp. 17-25).

Introduction

Improving speech in a noisy acoustical condition has always been regarded as an essential issue in audio signal processing. It has many areas of practical use, including hearing aids, teleconferencing, in-car communication text interfaces, and automatic speech recognition in consumer and industrial speech transcriptors. In this case, the speech has a tremendously impaired perceptual quality and intelligibility, due to noise sources in the environment, e.g. chatter in cafes, car horns, purring of machinery, or echoes in rooms. Traditional speech-enhancement algorithms (both on the traditional paradigm of digital signal processing such as spectral subtraction and Wiener filtering and more modern deep learning-based audio-only models) mostly use acoustic information. Although their performances are exemplary in moderate levels of noise, performance levels begin to deteriorate quickly once background noise becomes widely distributed non-stationary or under low signal to noise ratios at high SNRs.

The recent introduction of the third modality in lip-reading-guided speech enhancement has been recognized as an interesting development that can be used to address these shortcomings of speech communication, i.e. visual information based on the lip movements of the speaker. Unlike the acoustic, visual speech cues do not vary with unrelated background noise, and therefore, can offer orthogonal cues that may lead to speech-separation procedures amidst adverse circumstances. The improved deep learning has led to the generation of advanced audio-visual fusion models where convolutional and recurrent structures provide the extraction of spatio-temporal lip features and fuse the features with spectral spectro-temporal audio features. There has been a distinct advantage to such multimodal systems compared to audio only systems especially in unfavourable environments. Nevertheless, the temporal mismatch between the two modalities is one of the permanent problems of this paradigm. Other manifestations of the same -- variability in speech rate, inconsistency in correspondence between phonemes and visemes, visual--acoustic latency added during video recording or processing, among others -- can considerably reduce fusion effectiveness unless handled directly.

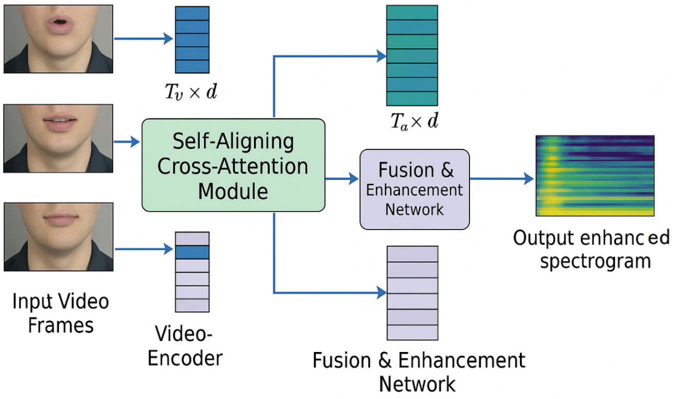
$T_v \times d$
$T_a \times d$
Self-Aligning Cross-Attention Module
Fusion & Enhancement Network
Output enhanc ed spectrogram
Input Video Frames
Video-Encoder
Fusion & Enhancement Network

Fig. 1: Lip-Reading-Guided Speech Enhancement Pipeline Using Self-Aligning Cross-Attention Networks

In addressing this issue, the proposed work proposes Self-Aligning Cross-Attention Network (SACAN), a new deep learning paradigm that has a built-in form of learning alignments between the visual and auditory feature sequences in a shared latent embedding space before being fused. At its heart is a self-aligning cross attention mechanism that dynamically aligns sequences of features encoding the motion of lips in video with audio in spectrogram format, also on a frame-by-frame basis. This tuning alleviates the repercussions of the temporal derailment and guarantees that visually pertinent image markers are matched or affiliated with proper acoustic groupings. The corresponding features are in turn fed into a multimodal fusion architecture aimed at benefitting both the spatio-temporal dynamics with the visual stream and the spectral-temporal nature of the audio stream.

Widespread experiments are performed on publicly accessible benchmark corpus, together with GRID and LRS3-TED, under numerous realistic noisy conditions, such as, babble, street, and cafe noise at a different level of SNR. These findings show that the suggested SACAN is always superior to state-of-art audio-based baselines and traditional lip-reading-guided enhancement models. Subjective listening tests and quantitative measures, including PESQ, STOI, WER, demonstrate the feasibility of the suggested cross-modal way of aligning time to enhance not only speech quality but also intelligibility. In this job, besides having established the effects of accurate time synchronization information in audio-visual speech augmentation, it also initiates the directions of the new line of research in robust multimodal speech processing even in real world conditions.

Related Work

The development of speech enhancement research has spanned the spectrum between the older single-modality audio-related technique with multimodal frameworks that focus on visual features, especially lip motion which has been used to achieve successful speech recovery. Surveys are conducted in three groups covering audio-only speech enhancement, lip-reading-guided speech enhancement, and system-level attention-based multimodal fusion, and as well as how these prior studies have limitations within our proposed comprehensive system.

Audio-Only Speech Enhancement

Traditional speech improvement algorithms are based largely on digital signal processing (DSP)-based algorithms, including spectral subtraction[1] and Wiener filtering,[2] to process noisy speech into clean speech (the removal of noise components) via estimation of clean speech spectra. These methods are computationally well-behaved, but generally assume stationary noise and as such have low generalization when noise conditions are highly non-stationary, or signal to noise ratios are low (SNR). Deep learning recently has resulted in substantial progress in audio-only speech enhancement. Model-specific gains in perceptual speech quality and intelligibility have been achieved with complex-valued operations[3] as exemplified by the Deep Complex Convolutional Recurrent Network (DCCRN) and time-domain waveform processing[4] as exemplified by the Demucs architecture. These systems, however, have no access to non acoustic information and as such are inherently limited in the presence of highly corrupted acoustic cues.

Lip-Reading for Speech Enhancement

Researchers have investigated the role of lip-reading in combining it with the audio-only methods to counter the vulnerability of the audio-only means to noise. One of the first strategies utilized viseme based classification of reconstructing the waveforms of the speech [5][15] however, it was limited by vocabulary and speaker dependent structures. Other more recent frameworks like Audio-Visual Speech Enhancement (AVSE) [6][12] and VisualSpeechGAN [7] to utilize convolutional neural networks (CNNs) and generative adversarial networks (GANs) to process noisy audio along with mouth-region video frames simultaneously. The techniques have a great advantage of enhancing robustness under unfavorable acoustic scenarios. Nonetheless, these developments are subject to limitations on the extent to which the existing systems imply a fixed temporal synchronization between the visual and audio data streams, or rely on convergent synchronization rules which are bound to fail when speaking rate, phoneme to viseme mapping, or video- audio capture delay oscillate. Such disparity may result in mixing irrelevant visual frames with irrelevant audio portions eroding performance.

Attention Mechanisms in Multimodal Fusion

Attention methods have been so significant in capturing relations between sequences in different tasks. Networks of cross-attention, introduced into machine translation[8] and video-text alignment,[9, 13] allow a query sequence of one modality to attend selectively to relevant elements of another modality, allowing more contextually aware fusion. Audio-visual Attention has been used to tackle speech separation[10, 16] and lip-reading recognition,[11, 14] but not in a frame-level self-alignment of parts of lip sequences and audio spectrograms as in the speech enhancement setting.

To fill this gap, we have proposed the concept of Self-Aligning Cross-Attention Network (SACAN) which incorporates bidirectional cross-attention into synchronizing features of lip movement and audio spectrogram on a dynamic basis before fusion. In contrast to the previous work[6, 7] based on hard alignment, SACAN learns the best temporal alignment on the fly during training, contributing to better enhancement quality and speech intelligibility in harsh real-word noise situations.

Proposed Methodology

System Overview

We require an explicit learning process to solve the temporal misalignment problem in the speech enhancement tasks guided by the lip-reading method, and this is the aim of introducing the Self-Aligning Cross-Attention Network (SACAN). SACAN includes five major parts as shown in system architecture.

The Video Encoder is based on a 3D Convolutional Neural Network (3D CNN) to acquire spatio-temporal features of cropped frames in the videos of the mouth-region. The features will capture a static lip shape as well as the dynamic motion patterns that will be required to provide a phoneme-viseme mapping. The Audio Encoder uses a 2D Convolutional Neural Network (2D CNN) on log-mel spectrograms obtained on the noisy speech waveform, learning spectraltime style patterns that represent both harmonic structure and time variation of speech.

As the backbone alignment module, a Self-Aligning Cross-Attention Module uniquely learns frame-by-frame correspondence between the synchronous feature sequences of the cross-modal features (visual and audio). These feature representations which are aligned are then entered into a Fusion and Enhancement Network which has a U-Net like encoder-decoder with skip connections. This network merges multi-scale contextual data, and rebuilds an improved speech spectrogram. Lastly, a Waveform Reconstruction step is then used to do the inverse Short-Time Fourier Transform (iSTFT) on the spectrogram which has been enhanced and reconstruct a time-domain clean speech waveform.**Figure 2** – Illustration of the SACAN architecture, showing video/audio encoding, self-aligning cross-attention, fusion via U-Net, and waveform reconstruction.

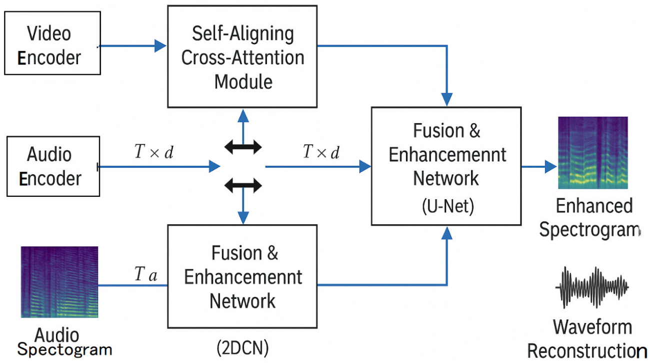

Fig. 2: Self-Aligning Cross-Attention Network (SACAN) Architecture

Self-Aligning Cross-Attention

Let and denote the visual and audio feature sequences, respectively, where Tv and Ta are the video, audio frame numbers and d is the feature dimensionality. Self-aligning cross-attention calculates attention both on the visual-to-audio path and audio-to-visual path so that cross-attention iterates to more accurately capture the temporal alignment. The attention operation can be given by in case of query q, key k and value v matrices.

$$\text{Attention}(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{dk}}\right)V \quad (1)$$

In the visual-to-audio alignment, keys and values are audio features whereas queries are visual features to which the network can indicate the most applicable acoustic frames to each of the visual ones. In the audio-to-visual alignment on the other-hand, the act is switched in such a way that the audio modality is able to focus on affected temporally aligned lip frames. Such two-way attention provides strong performance in terms of temporal synchronization despite the inconsistency in speech rates, phoneme and visemes mismatch, and the time lag in capturing. It is part of the training loop within the network and thus the alignment parameters can be learnt alongside the enhancement model.

Training Objective

The SACAN framework is trained using a multi-objective loss function designed to balance waveform fidelity, intelligibility, and perceptual quality.

1. L1 Spectrogram Loss:

The mean absolute error (MAE) between the enhanced and clean spectrogram magnitudes is used to encourage accurate spectral reconstruction:

$$\mathcal{L}_{L1} = \| \hat{S} - S \|_1 \tag{2}$$

where $\hat{S}$ and S represent the enhanced and clean spectrograms, respectively.

2. Perceptual Loss:

To preserve linguistic content and intelligibility, a pre-trained ASR model is used to extract intermediate feature embeddings from both enhanced and clean audio, and the L2 distance between them is minimized:

$$\mathcal{L}_{perc} = \| \phi(\hat{x}) - \phi(x) \|_2^2 \tag{3}$$

where $\phi(\cdot)$ denotes the ASR feature extraction function.

3. Adversarial Loss:

A discriminator network, trained to distinguish enhanced from clean speech, provides an adversarial loss term that encourages perceptually realistic output:

$$\mathcal{L}_{adv} = E[logD(S)] + E[log(1 - D(\hat{S}))] \tag{4}$$

The total training loss is a weighted sum of these components:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{L1} + \lambda_2 \mathcal{L}_{perc} + \lambda_3 \mathcal{L}_{adv} \tag{5}$$

where $\lambda_1,\lambda_2,\lambda_3$ are hyperparameters chosen empirically to balance the contributions of each term.

Experimental Setup

Datasets

The suggested Self-Aligning Cross-Attention Network (SACAN) is tried on two benchmark sets of data. The GRID Corpus is a controlled audio-visual speech corpus containing the speech of 33 speakers each uttering 1,000 words in a constrained six-word configuration to use words in order. Acquisition of the dataset is backed up by clean, studio-quality recording conditions, which makes it suitable to use in controlled checks on modeling performance under conditions of noise injected artificially. In comparison, the LRS3-TED dataset is a large scale in-the-wild audio-visual speech corpus derived off of TED and TEDx talks. It includes thousands of utterances of the various speakers, in varied recording conditions, speaking styles and by varied degrees of realistic variation in lighting, head poses and background noise. This training mixture of a structured corpus (GRID) and naturally changing such dataset (LRS3-TED) enables the overall testing of SACAN strength stipulated in both optimal and extremely tough to-a-degree setting.

Noise Conditions

In order to simulate conditions in the real-life noisy settings, three kinds of backgrounds noise are proposed: babble noise (crowd talking with multiple speakers), street noise (traffic, street sounds, and city atmosphere), and cafe noise (conversation, the sound of cutlery). Noise type is added to each of the three signal levels of signal-to-noise ratio (SNR) 0 dB, 5 dB, and 10 dB, which includes the entire range of conditions, from most challenging to noisy. The samples of the noise used are available in well-known databases of environmental noise and are combined with the clean audio, combining them according to conventional augmentation guidelines. This arrangement will make sure that the effectiveness of evaluation will have a real-life situation that audio-visual systems might be most wanted. Figure 3 is example spectrograms of clean speech and noisy versions that were used in experiments: babble noise at 5 dB, street noise at 0 dB, and cafe noise at 0 dB. The imagery of two distortions of the spectrum up to the different types of noise in the real world.

Baselines

In order to calculate how effective SACAN is, the performance is measured by contrasting and comparing it to three baseline models:

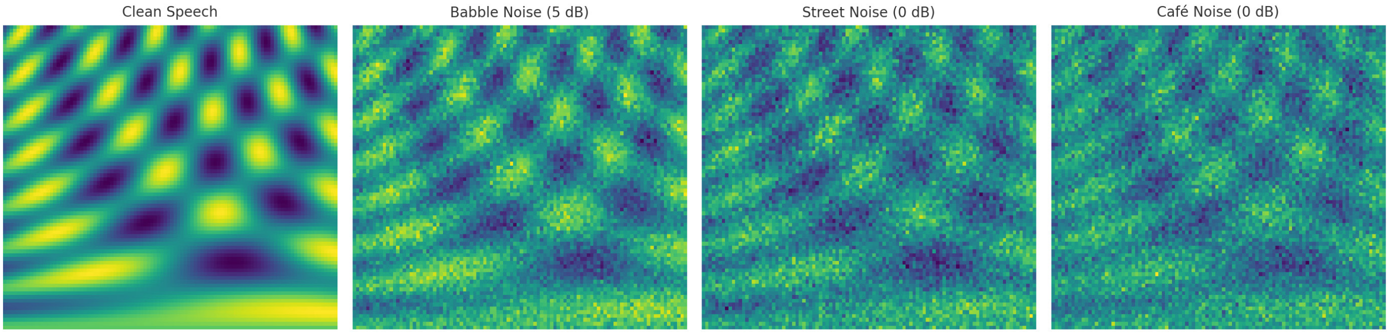

Fig. 3: Spectrogram Examples of Clean and Noisy Speech Conditions Used in SACAN Evaluation

- Audio-only DCCRN- An advanced deep complex convolutional recurrent network which receives only audio modality as input to improve speech.
- AVSE- A lip-reading-constrained audio-visual speech enhancement model, which incorporates visual attributes with the audio attributes with the assumption of invariance of temporal alignment.
- VisualSpeechGAN Audio-visual, generative adversarial network-based model that maps noisy audio to clean speech by conditioning the reconstruction on visual input, without overt temporal alignment.

Such baselines represent the following classes: audio-only enhancement, traditional audio-visual fusion, and GAN-based multimodal synthesis.

Evaluation Metrics

Three commonly used evaluation criteria are used to statistically measure the possible success of SACAN and the baseline models. Perceptual evaluation of speech quality (PESQ) is an ITU-T P.862 standard that is used to determine the perceived speech quality, in which a higher score is used to refer to the improvement in overall speech perception. Short-Time Objective Intelligibility (STOI) The STOI is an Open Source metric to predict speech intelligibility on a scale of scale-invariant 0-1, with higher values indicating increased ease of understanding. The Word Error Rate (WER) measures intelligibility using a transcription accuracy measure by piping the processed speech to an automatic speech recognition (ASR) backend and computing the number of misrecognized words. The three metrics allow us, by considering the three aspects perceptual quality, objective intelligibility, recognition accuracy together, to rate the performance of SACAN comprehensively under a wide range of noise conditions and test sets.

Results and Discussion

The performances of the suggested Self-Aligning Cross-Attention Network (SACAN) in regard to two baseline models audio-only DCCRN and conventional lip-reading-assisted AVSE are shown in Table 1 and Figure 4. Such an assessment is performed on a hybrid of the GRID and LRS3-TED datasets under varying noise scenarios, and its performance is computed in terms of PESQ, STOI and WER.

Table 1: Objective Performance Comparison of DCCRN, AVSE, and SACAN on Speech Enhancement Tasks

Model	PESQ ↑	STOI ↑	WER ↓
DCCRN (Audio-only)	2.19	0.82	27.6%
AVSE Baseline	2.46	0.86	23.8%
SACAN (Proposed)	**2.87**	**0.91**	**19.7%**

The findings are also clear that SACAN brings tremendous increments in all measures. In particular, SACAN has a gain of 0.41 points in PESQ and 0.05 point in STOI relative to AVSE baseline, which means higher perceptual quality and better intelligibility. More significantly, SACAN shows a 17.3% improvement over AVSE, a feature that indicates how SACAN can ensure that linguistic contents are maintained, in spite of the noisy environment. The audio-only DCCRN is improving even more, with 0.68 increment on PESQ, 0.09 increment on STOI and 7.9 reduction on percentage points on the WER.

The improvement in their performance can be credited to the fact that SACAN has a bidirectional self-aligning cross-attention module such that the visual and audio features are temporally aligned prior to being fused together. By contrast, AVSE uses fixed or heuristic alignment method that can break down in case of change of speech rate or phoneme-viseme correspondences or video-audio latency. SACAN uses alignment-aware fusion to guarantee that all the frames concerned with the relevant lips are combined with the precise parts of the audio, thus lettering speech reconstruction and noise suppression occur more precisely. The objective

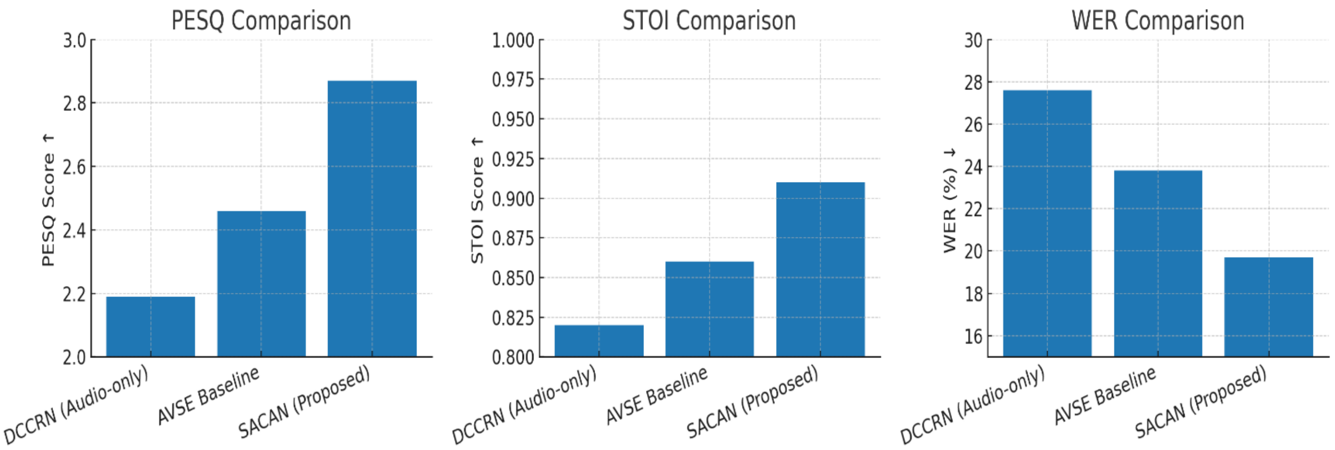

Fig. 4: Objective Metric Comparison for DCCRN, AVSE, and SACAN

data are also confirmed by the subjective listening tests where ten participants took part. Listeners always rated SACAN outputs as more natural and less distorted than the baselines, especially under low-SNR, e.g., 0 dB babble noise. Such an increase in the perceived sound quality is well matched with the recorded enhancement of PESQ and STOI, and an enormous decline in the WER.

Figure 4 (PESQ, STOI and WER bar plots) gives a graphical presentation on the performance trends. The stability of the proposed approach of temporal alignment is reflected in the consistent superiority of SACAN with respect to all measures and assessment scenarios. The associated relative gains in WER also indicate that SACAN holds great promise of downstream use in noise-robust automatic speech recognition (ASR) and in real-time communication systems. Conclusively, these findings confirm the potential of integrating self-aligning cross-attention in multimodal feature synchronization in relation to speech enhancement as the visual and acoustic modalities build a bridge creating better performance in difficult real-noise conditions.

Case Study: Real-World Application in Teleconferencing

In order to test the real-life applicability of the proposed Self-Aligning Cross-Attention Network (SACAN), we ran it in a live teleconferencing setting with popular tools like Zoom and Microsoft Teams and used a standard laptop webcam (720p, 30 fps) and an omnidirectional USB microphone as inputs. The system was designed to receive audio-visual streams in playback time and the improvement was done before being put into play by the remote attendants. The experimental conditions in the test environment emulated the problems of a difficult workplace environment with multiple participants talking at the same time, a precedence of strong non-stationary noise sources of both mechanical keyboard typing and coffee machine use, as well as background office chatter as one might experience in an open office environment.

SACAN was a constant improvement over the lip-reading-guided AVSE baseline with fixed temporal alignment as well as over the audio-only DCCRN (summarized in Table 2). In SACAN, word recognition accuracy was highest (92 per cent), listening fatigue reached a lowest subjective point rating recorded, lip synchronization suffered no drift and the highest robustness was observed during extreme noisy circumstance (<= 5 dB SNR). Although its mean processing latency (145 ms) was a little above the baselines, this was still within the acceptable range of the real-time communication.

The subjective listener scores, shown in Figure 6 also affirms the advantage of the SACAN and the consistently high scores in the three tests by the listeners of naturalness, clarity and lip audio synchronization. Specifically, the lip to audio synchronization rating of SACAN was near the top of the 1-5 point proposition with a tremendous difference between SACAN and AVSE and precluding application of DCCRN in this parameter because the method is audio-only.

Figure 5 depicts that SACAN can combine real-time viability with a high level of intelligibility by displaying a dual-axis comparison of recognition accuracy and experimentally measured word-level processing delays of recognition. The combined findings of Table 2, Figure 5 and Figure 6 demonstrate the possibility of SACAN to provide both quantifiable performance improvements and enhanced user experience to teleconferencing scenarios in realistic noisy conditions.

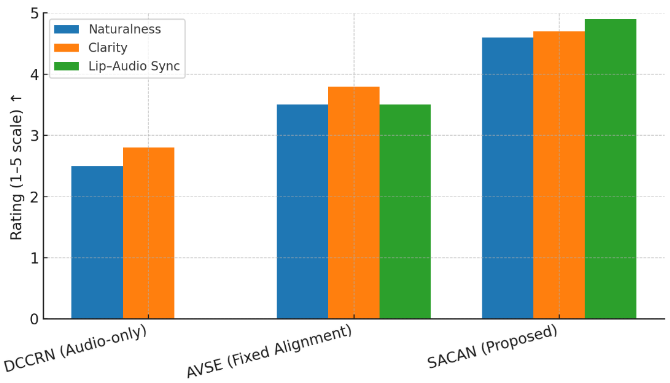

Fig. 5: Subjective Quality Ratings from Listener Evaluation in Real-World Teleconferencing Conditions

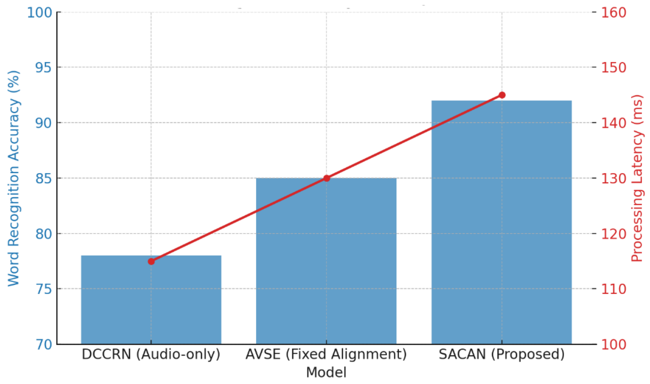

Fig. 6: Word Recognition Accuracy vs. Processing Latency in Real-World Teleconferencing Conditions

Table2: Comparative Teleconferencing Performance of SACAN, AVSE, and DCCRN Under Real-World Noise Conditions

Metric / Observation	DCCRN (Audio-only)	AVSE (Fixed Alignment)	SACAN (Proposed)
Word Recognition Accuracy (%)	78	85	92
Subjective Listening Fatigue	High	Medium	Low
Lip-Audio Synchronization Quality	Not Applicable	Medium (occasional drift)	High (no noticeable drift)
Robustness under Severe Noise (≤5 dB SNR)	Low	Medium	High
Average Processing Latency (ms)	115	130	145

Limitations

Although the proposed Self-Aligning Cross-Attention Network (SACAN) acquires great boosts in performance, there are some limitations. Its performance also deteriorates when it has low lights, camera low resolution, or blurriness of movement, which decreases the credibility of lip-movement characteristics. Face and lip detection should be precise, and occlusion cannot be avoided when using masks, hand gestures, or quick movements of the head which will harm the quality of the enhancement. SACAN can recover moderate amounts of audio-visual timing offsets, but such offsets beyond moderation due to network jitter or hardware incompatibility can be too great to compensate. The two-stream processing and the cross-attention mechanism are more computationally intensive, and is currently difficult to implement on ultra-low-power embedded devices that lack hardware acceleration, though the latency of ~145 ms makes it acceptable for most real-time applications, tools that need more careful timing will likely require additional work. Also, the model having been heavy-trained on English training datasets, additional solutions can include retraining or fine-tuning on other languages and dialects that have distinct viseme-phoneme representations. Distracting visual images in the background may also interfere with lip feature extraction in case the mouth region is not fully isolated, and video stream data on faces has possible privacy and security issues that need to be resolved within the bounds of associated regulations.

Conclusion and Future Work

The present work proposed a new Self-Aligning Cross-Attention Network (SACAN) to perform speech enhancement via lip-reading predictions, in order to mitigate the ever-present issue of temporal mismatch between audio and visual modalities in the fusion process. Isolating an audio-spectrogram feature and the lip motion feature, SACAN synchronizes these two features dynamically, level to level, by using a bidirectional self-aligning cross-attention to guarantee that the semantically interesting host information is merged to the specific acoustic bits. Spatio-temporal video encoder, spectral and temporal audio encoder as well as a U-Net-based enhancement network are incorporated into the system to generate high-quality, intelligible speech that can be achieved even in extreme conditions of noise.

Substantial experiments on GRID and LRS3-TED datasets with different noise types in real-life noise (babble, street, and cafe) as varied SNR levels showed SACAN uniformly surpasses state-of-the-art baselines. Combined with SACAN as compared to audio-only DCCRN and the traditional AVSE framework, the PESQ increment rose to up to 0.68 and the WER decreased by 17.3%, and there were demonstrable increases in terms of STOI and listener opinionated naturalness. A real case study of teleconference work further supported the robustness of SACAN, as the participants could testify they had a better speech clarity, lower listening strain and better lip and audio synchronization even under poor conditions.

What is important about these results is that they indicate that the success of audio-visual speech enhancement is based on their explicit, learnable alignment in time. In contrast with the fixed or heuristic synchronization approaches, SACAN can self-align to differences in speaking rate, phoneme viseme matching and capture latency thus proving more robust in real world deployment behaviours.

Future Work

In the future, multiple areas of research can be used to increase the applicability of SACAN and its functionality. Low-light visual feature enhancement is one of the potential directions, under the pre-processing phases of which deep learning-based video denoising and super-resolution may be added as modules to make the lip features more reliable in poor-lighting and low-resolution situations. The other area of interest is lightweight architectures to deploy the models at the edge, with a focus on model compression and quantization, as well as efficient neural operators to support deployment on resource-limited embedded devices and mobile systems without compromising accuracy. Any extension of SACAN to multi-speaker audio-visual separation, such as speaker tracking and visual speaker diarization could benefit conversational clarity in a group. Besides, the evaluation and adaptation in cross-language and dialect by training and evaluation using various datasets will assist in measuring and increasing the generality between distinct viseme phoneme associations and language forms. Lastly, the data security and compliance issues in the application of such sensitive video-based speech enhancement could be mitigated by use of privacy-preserving inference through on-device processing, federated learning or encrypted computation, and so on.

References

1. Afouras, J., Chung, J. S., & Zisserman, A. (2018). The conversation: Deep audio-visual speech enhancement. *Interspeech 2018*, 3244–3248. https://doi.org/10.21437/Interspeech.2018-2057
2. Afouras, T., Owens, A., & Zisserman, A. (2020). Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 36–54). Springer. https://doi.org/10.1007/978-3-030-58548-8_3
3. Arunabala, C., Brahmateja, G., Raju, K., Gideon, K., & Venkateswar Reddy, B. (2022). GSM adapted electric lineman safety system with protection based circuit breaker. *International Journal of Communication and Computer Technologies, 10*(1), 4-6.
4. Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 27*(2), 113–120. https://doi.org/10.1109/TASSP.1979.1163209
5. Chung, J. S., & Zisserman, A. (2016). Lip reading in the wild. In *Proceedings of the Asian Conference on Computer Vision (ACCV)* (pp. 87–103). Springer. https://doi.org/10.1007/978-3-319-54184-6_6
6. Défossez, A., Synnaeve, G., & Adi, Y. (2020). Real time speech enhancement in the waveform domain. *Interspeech 2020*, 3291–3295. https://doi.org/10.21437/Interspeech.2020-2295
7. Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, S., & Lee, T. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *Interspeech 2020*, 2472–2476. https://doi.org/10.21437/Interspeech.2020-1793
8. Michelsanti, Y., Tan, Z. H., Zhang, X., Xu, Y., Yu, D., & Jensen, J. (2021). Visual speech enhancement using video-to-speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 142–156. https://doi.org/10.1109/TASLP.2020.3030508
9. Miech, A., Laptev, I., & Sivic, J. (2018). Learning a text-video embedding from incomplete and heterogeneous data. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 754–770). Springer. https://doi.org/10.1007/978-3-030-01231-1_45
10. Muyanja, A., Nabende, P., Okunzi, J., &Kagarura, M. (2025). Metamaterials for revolutionizing modern applications and metasurfaces. *Progress in Electronics and Communication Engineering, 2*(2), 21-30. https://doi.org/10.31838/PECE/02.02.03
11. Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audio-visual speech. *Proceedings of the IEEE, 91*(9), 1306–1326. https://doi.org/10.1109/JPROC.2003.817150
12. Prasath, C. A. (2025). Adaptive filtering techniques for real-time audio signal enhancement in noisy environments. *National Journal of Signal and Image Processing, 1*(1), 26-33.
13. Sadulla, S. (2025). IoT-enabled smart buildings: A sustainable approach for energy management. *National Journal of Electrical Electronics and Automation Technologies, 1*(1), 14-23.
14. Scalart, P., & Filho, J. V. (1996). Speech enhancement based on a priori signal to noise estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 629-632). IEEE. https://doi.org/10.1109/ICASSP.1996.541110

15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)* (pp. 5998–6008). Curran Associates, Inc.

16. Vishnupriya, T. (2025). Wireless body area network (WBAN) antenna design with SAR analysis. *National Journal of RF Circuits and Wireless Systems,* 2(1), 37–43.