# Robust Real-time Audio-Visual Speech Enhancement based on DNN and GAN

Mandar Gogate, Kia Dashtipour, Amir Hussain
*School of computing, Edinburgh Napier University*
Edinburgh, United Kingdom

***Abstract*—The human auditory cortex contextually integrates audio-visual (AV) cues to better understand speech in a cocktail party situation. Recent studies have shown that AV speech enhancement (SE) models can significantly improve speech quality and intelligibility in low signal-to-noise ratios ($SNR < -5dB$) environments compared to audio-only (A-only) SE models. However, despite substantial research in the area of AV SE, development of real-time processing models that can generalise across various types of visual and acoustic noises remains a formidable technical challenge. This paper introduces a novel framework for low-latency, speaker-independent AV SE. The proposed framework is designed to generalise to visual and acoustic noises encountered in real world settings. In particular, a generative adversarial network (GAN) is proposed to address the issue of visual speech noise including poor lighting in real noisy environments. In addition, a novel real-time AV SE based on a deep neural network is proposed. The model leverages the enhanced visual speech from the GAN to deliver robust SE. The effectiveness of the proposed framework is evaluated on synthetic AV datasets using objective speech quality and intelligibility metrics. Furthermore, subjective listening tests are conducted using real noisy AV corpora. The results demonstrate that the proposed real-time AV SE framework improves the mean opinion score by 20% as compared to state-of-the-art SE approaches including recent DNN based AV SE models.**

Impact statement: Hearing aids are widely used devices for compensating for hearing loss. However, they present significant challenges for individuals with hearing impairment, as these devices often amplify sounds without fully restoring speech intelligibility in social settings with high background noise. The cognitively inspired technology proposed in this paper overcomes this limitation. With a significant increase in speech intelligibility performance in the presence of multiple competing noise sources, the technology can support communication for hearing aid users in cocktail party environments. Moreover, the proposed technology can be exploited in mobile teleconferencing and extremely noisy environments e.g., situations where ear defenders are worn such as emergency and disaster response.

***Index Terms*—audio-visual, speech enhancement, generative adversarial network**

## I. Introduction

More than 430 million people worldwide currently suffer from hearing loss. These numbers are expected to reach 2.5 billion by 2050 [1, 2]. The most common type of hearing loss is neither curable nor reversible. Studies have shown that, the hearing impaired listeners often find themselves in social isolation leading to depression. Hearing aids and cochlear implants are widely used to compensate for hearing loss. However, even the hearing aids that use state-of-the-art signal processing algorithms pose significant problems for the people with hearing loss as these listening devices fail to restore speech intelligibility in busy social situations [3]. Normal hearing listeners in such environments use the audio-visual (AV) nature of speech to suppress background noise and focus on the target speaker. Therefore, researchers have proposed AV SE methods that extend audio-only (A-only) speech enhancement (SE). SE, aims to separate speech from background noise, has had a huge impact in recent years due to its applications in hearing aids, cochlear implants, speech recognition, mobile communication, and voice activity detection [4]. Despite extensive research advances in AV SE, hearing scenarios are becoming more complex with a wide range of non-stationary acoustic noises, visual speech noises and reverberations in physical space.

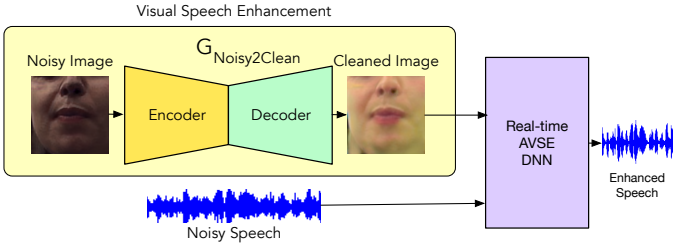



Fig. 1: Proposed Real-time Audio-Visual Speech enhancement Framework

In the literature, extensive studies have been carried out to develop AV SE methods in time-domain and frequency domain [5, 6, 7]. However, despite significant research in the area of AV SE, real-time processing models, with low latency (8-12 ms) remains a formidable technical challenge. Most of the aforementioned methods are non-causal and computationally complex [5]. Therefore, these methods are not suitable for processing streaming data with real-time constraints. The processing latency of SE algorithm is important to the hearing impaired listeners as the delayed processing will result in an echo effect and unsynchronised AV cues leading to poor speech intelligibility.

In this study, a robust real time AV SE framework depicted

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) programme grant: COG-MHEAR (Grant reference EP/T021063/1).

Corresponding author: m.gogate@napier.ac.uk

in Fig 1, that is causal and can process streaming data, is proposed. The framework consists of two components: (1) generative adversarial network (GAN) based visual SE (2) deep neural network based low latency AV SE model. Specifically, the novel visual noise-robust GAN architecture is proposed to enhance the noise present in visual speech. The developed GAN takes in noisy lip images with poor lighting and head movements, and outputs an enhanced version by eliminating visual speech noises, similar to studio recordings. The real-time AV SE model ingests the enhanced visual images predicted by the GAN architecture for more robust SE with a variety of visual and acoustic noises. To the best of our knowledge, this study is the first to propose a framework that jointly enhances the visual and acoustic speech signal for robust AV SE. The comparative simulation results in terms of objective speech quality and intelligibility metrics (including PESQ, STOI and SI-SDR) and subjective listening tests (using real noisy ASPIRE [8] and VISION [9] corpora) show significant performance improvement of the proposed framework as compared to state-of-the-art and DNN based audio-only and AV SE approaches including spectral subtraction (SS) [10], Linear minimum mean square error (LMMSE) [11], SE Generative Adversarial Network (SEGAN) [12] and CochleaNet [8].

In summary, this paper presents two major contributions:

1) A generative adversarial network architecture is proposed to address the main limitation of visual imperfection in AV SE models. To the best of our knowledge, this paper is first to introduce a framework that takes into account both acoustic and visual speech noise for AV SE.
2) A real-time AV SE model is proposed for low-latency inference on streaming audio-visual data. In addition, the computational latency of individual blocks in the proposed framework is presented, demonstrating the effectiveness of the model for causal SE.

The rest of the paper is organised as follows: Section II presents an overview of audio-visual speech enhancement models proposed in the literature. Section III presents visual speech enhancement using generative adversarial network. Section IV introduces real-time audio-visual speech enhancement framework. Section V presents the experimental results. Finally, section VI concludes the work and presents possible future research directions.

## II. Related Work

This sections presents the related work in the area of AV SE.

Afouras et al. [13] presented a deep neural network model to separate the speaker's voice using lip region features. The model is trained using studio-quality LRS2 and VoxCeleb2 datasets to predict magnitude and phase of the target signal. Similarly, Gogate et al. [14] proposed a speaker independent AV SE model based on deep neural networks. The model is trained and evaluated using synthetic GRID-CHIME3 dataset. However, the main limitation is that, the model is evaluated on a limited vocabulary GRID corpus. On the other hand, Hou et al. [15] proposed a deep denoising autoencoder based on convolutional neural network (AVDCNN) for SE, which combines both audio and visual modalities. Comparative simulation results show that, the proposed AVDCNN outperforms state-of-the-art A-only approaches including logMMSE.

Lu et al. [16] introduced a speaker-independent speech separation model based on AV deep clustering. The model learns time-frequency (T-F) embeddings for AV speech features. The model was trained using GRID and TCD-TIMIT corpora. The experimental results demonstrate that the proposed AV model outperforms A-only deep clustering and other state-of-the-art approaches. In addition, Furthermore, Michelsanti et al. [18] performed a set of experiments to understand the impact of Lombard effect on AV SE. The empirical results indicated the benefits of training system with Lombard AV GRID corpus on speech quality and intelligibility in low SNR environments.

In addition, Gogate et al. [8] proposed speaker independent AV deep neural network for ideal binary mask estimation to remove the noise from the speech in low SNR environments. The model is trained using GRID corpus and evaluated using a range of speaker, noise, and language independent test sets.

More recently, Arriandiaga et al. [20] proposed a method based for SE in multi-talker conversation environment. The facial landmarks are used as an alternative to visual lip images which is widely used in AV SE literature. The model is trained using GRID corpus to predict ideal amplitude masks in order to filter noisy audio. The experiments show that the method achieved better performance with low latency and computational cost. Moreover, Gao et al. [21] developed a method to learn cross-modal speaker embeddings and SE in the multi-task setting. The model integrates lip motion features, facial attribute embedding for robust SE. The LRS2, VoxCeleb1 and VoxCeleb2 corpora are used to measure the performance of the approach. The results indicated that the approach can generalise well in real-world scenarios.

Furthermore, Chuang et al. [24] introduced an AV SE approach for car-driving scenarios based on deep learning. The approach addresses three main issues which are often encountered in developing AV SE approaches, additional cost of processing data, AV synchronization, and low-quality visual data. The Taiwan Mandarin speech used to evaluate the performance of the approach. The initial experimental results confirm that the approach is suitable for real-world scenarios where the high quality of data is not always available. Finally, Gogate et al. [23] proposed an AV SE model based on Temporal Convolutional Networks which exploit the privacy-preserving lip-landmark features for SE in multi-talker cocktail party environments. The model was trained using GRID and TCD-TIMIT corpora. The experimental results reveal the effectiveness of the approach as compared to benchmark A-only and AV approaches.

Recently, Zhu et al. [25] presented AV-E3Net, a low-latency real-time AV end-to-end SE model. The model incorporates a multistage gating-and-summation (GS) fusion module to combine speech and vision modalities. The AVSpeech, VoxCeleb, and LRS3 datasets were used to evaluate the model's

TABLE I: Summary of the State-of-the-art AV SE Approaches.
SI - speaker independent, RTL - real-time latency

| Paper | Year | Input | Output | Dataset | Model | SI | Causal | RTL | Limitation |
|---|---|---|---|---|---|---|---|---|---|
| [13] | 2018 | Raw pixels (Lip) | Complex Ratio Mask | LRW | CNN | Yes | No | - | Model is non-causal and sensitive to AV synchoronisation |
| [14] | 2018 | Raw pixels (Lip) | Ideal binary mask | GRID | CNN, LSTM | Yes | Yes | 20 ms | Limited vocabulary dataset is used for training and evaluation |
| [15] | 2018 | Raw pixels (Lip) | Raw spectrogram | Taiwan MHINT | CNN | No | No | - | Model is speaker-dependent |
| [16] | 2019 | Raw pixels (Lip) and Optical Flow | Ideal ratio mask | GRID, TCD-TIMIT | CNN, BiLSTM | Yes | No | - | Limited vocabulary dataset is used for training and evaluation |
| [17] | 2019 | Raw pixels (Face) | Complex Ratio Mask | LRS3 | CNN, BiLSTM | Yes | No | - | Model is non-causal |
| [18] | 2019 | Raw pixels (Lip) | Ideal Amplitude Mask | Lombard GRID | CNN | No | No | - | Model is speaker-dependent |
| [19] | 2020 | Raw pixels (Lip) | Raw waveform | GRID | CNN, LSTM | Yes | No | - | Limited vocabulary dataset is used for training and evaluation |
| [8] | 2020 | Raw pixels (Lip) | Ideal binary mask | GRID, TCD-TIMIT, Mandarin | CNN, LSTM | Yes | Yes | 25 ms | Phase is not considered |
| [20] | 2021 | Event-driven motion features | Ideal Amplitude Mask | GRID | BiLSTM | Yes | No | - | Limited vocabulary dataset is used for training and evaluation |
| [21] | 2021 | Raw pixels (Face and Lip) | Complex Ratio Mask | VoxCeleb2, TCD-TIMIT, LRS2 | CNN | Yes | No | - | Model is non-causal |
| [22] | 2021 | Raw pixels (Lip) | Ideal ratio mask | GRID | CNN | Yes | No | - | Model is non-causal and Limited vocabulary dataset is used for evaluation |
| [23] | 2022 | Lip Landmark | Ideal Amplitude Mask | GRID, TCD-TIMIT | CNN | Yes | Yes | 11 ms | Landmark features cannot be extracted in the presence of visual noise |
| [24] | 2022 | Raw pixels (Lip) | Raw spectrogram | TMSV | CNN, LSTM | Yes | No | - | Model is non-causal |
| [25] | 2023 | Raw pixels (Lip), noisy signals | Enhanced signal | AVSpeech, VoxCeleb, and LRS3 | CNN, LSTM | Yes | Yes | - | One of the proposed model cannot be used for real-time processing because of the model's dependency on a pre-trained video encoder |
| [26] | 2023 | Raw pixels (Lip), Noisy Mel-spec | Enhanced signal | AVSpeech | GAN | Yes | No | - | Model is non-causal |

performance. The experimental results indicate that the proposed AV-E3Net has excellent potential for real-world video communication applications, offering a low-latency real-time solution. In addition, Mira et al. [26] introduced a two-stage approach for predicting mel-spectrograms from noisy AV speech using a transformer-based architecture. The predicted mel-spectrograms are then converted into waveform audio using a neural vocoder. The performance of this approach was evaluated using the AVSpeech dataset.

Table 1 presents a summary of the aforementioned state-of-the-art approaches for AV SE. It can be seen that, most of the aforementioned AV SE models are non-causal and hence cannot be used for real-time SE. In addition, the models are trained and evaluated using corpora (e.g. GRID, TCD-TIMIT, LRS2) recorded in an ideal (studio-like) environment with no visual imperfections. However, in real world environments visual speech is often degraded by poor lighting, occlusions and head movements. This limits the ability of AV SE models to generalise in real noisy environments where both visual and acoustic speech is mixed with a range of noises.

## III. Visual speech enhancement using Generative Adversarial Networks

In the literature, it has been shown that the performance of the AV SE model is severely affected when visual speech is degraded with noise including poor lighting, occlusions, and head movements. In order to address the aforementioned issues, this section proposes a Generative Adversarial Networks (GAN) architecture to denoise the visual speech imperfections encountered in the real-world environment.

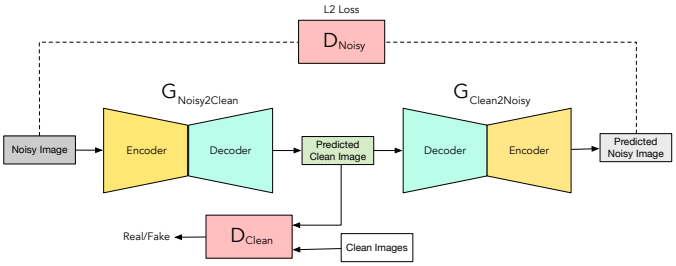



Fig. 2: Proposed visual speech enhancement framework adopted from [27]

In order to train the visual speech denoising model, the set of input and output images required for training. Collection of paired real noisy data (studio quality images and images with visual imperfections) is infeasible for the task of visual SE. This limitation can be addressed by using CycleGAN [27], that exploits cycle consistency loss to enable training without the need for paired data. The model can translate from one domain to another using unpaired samples from individual domains. The CycleGAN architecture, shown in Fig. 2, consists of two generator and discriminator networks. The first generator maps

the input image from domain A to B and the second generator maps from domain B to A. The two discriminator is used to estimate the distance between the predicted samples and actual samples from each domain for model training. The domain A to B generator can be used separately for the task of image translation once the training is complete.

In this paper, the CycleGAN architecture is adopted for the task of robust visual SE. The GAN architecture learns the mapping between noisy visual speech (consisting of poor lighting, exposure, and contrast) and clean visual speech images. The predicted visual speech is then fed to a real-time AV model, presented in section IV, for more robust SE.

### A. Data Representation

**Input features**: The GAN framework uses noisy cropped lip region input. The cropped $96 \times 96$ lip region is extracted from real noisy VISION [9] and were used as noisy image samples.
**Output**: The GAN framework generates the clean cropped lip region. It is worth to mention that, the GRID corpus is recorded in the studio environment and it can be used without any modification as model output.

### B. Network Architecture

The previously proposed CycleGAN architecture [27] was adopted for the task of visual SE. The generator network consists of three basic building blocks: (1) encoder (2) resnet (3) decoder. The encoder block consists of three convolutional layers with 64, 128, and 256 filters respectively that mitigates the illustration by one fourth of actual image size. Encoder output is ingested to a resnet consisting of 9 residual blocks. The decoder block consists of 3 deconvolutional layers with 128, 64, and 3 filters respectively to regenerate the original size of input image ($3 \times 96$). For the discriminator network, the $48 \times 48$ pixel PatchGAN [28] architecture used which is able to classify if $48 \times 48$ overlapping image patches are real or fake. This reduces the number of parameters as compared to full image discriminator. PatchGAN discriminator network consists of 4 convolutional layers with 64, 128, 256 and 512 filters. Each convolution has filter of size $4 \times 4$, stride of $1 \times 1$ and is followed by instance normalisation layer and LeakyReLU activation with slope of 0.2.

## IV. Real-time Audio-visual Speech enhancement Framework

This section presents the steps involved in end-to-end processing of the proposed real-time AV SE framework shown in Fig. 3. The proposed real-time AVSE can be used in web communication applications like Microsoft Teams and Zoom, as the model can process streaming AV data to focus on the target speaker.

### A. Data Representation

*a) Input features:* The deep neural network used both the audio and visual as the input features. Three seconds lip embeddings are used for the batch training and the cropped $96 \times 96$ lip region is extracted from the video and fed to lip embedding network to generate $75 \times 512$ dimension vector of lip embeddings for three second of video (assuming 25fps sampling rate). The audio input is divided into windows and a short-time Fourier transform (STFT) of audio segment is calculated. The magnitude of STFT is fed to the models as noisy input. The model is trained on 3 second segments and can be used for inference of arbitrary lengths of noisy video.

*b) Output:* The ideal binary mask is used as the output of the network. IBM is a multiplicative spectrogram mask that shows the time-frequency (T-F) relationship between the source audio and interfering noise. The IBM has a value of 1 where the local SNR is higher than local criterion (LC) and zero otherwise. The LC is calculated using the source audio and interfering noise.

TABLE II: Noisy Audio Feature Extraction

| | conv1 | conv2 | conv3 | conv4 | conv5 |
|---|---|---|---|---|---|
| Num filters | 64 | 64 | 64 | 64 | 4 |
| Filter size | $5 \times 5$ | $5 \times 5$ | $5 \times 5$ | $5 \times 5$ | $1 \times 1$ |
| Dilation | $1 \times 1$ | $1 \times 1$ | $1 \times 1$ | $1 \times 1$ | $1 \times 1$ |

### B. Network Architecture

A detailed outline of the proposed framework is shown in Fig.3. The individual components are explained in the subsequent sections.

*1) Noisy audio feature extraction:* The number of convolution filters, strides and dilation used in the acoustic feature extraction is detailed in Table II.

*2) Lip Embedding Network:* The lip embedding network shown in Fig. 3 consists of a 3D-convolutional network, RESNET-18 and temporal convolutional network (TCN). The 3D-CNN consists of single filter with size of $5 \times 7 \times 7$. The 3D-CNN features are fed to standard RESNET-18 architecture. The output of the residual network is fed to a multi-scale TCN as described in [29]. TCN output is fed to a fully-connected layer for word classification. The model is trained end-to-end with LRW dataset for lip reading [30]. The model weights were frozen and the fully connected network in the trained model was removed for extracting lip embedding features.

*3) Multimodal Fusion:* The sampling rate for visual features (512-D) is 25 frames per second and 75 vectors per second for audio features. In order to match the audio STFT sampling rate (75 fps), the visual frames (sampled at 25 fps) are upsampled using the repetition of the element three times in the temporal dimension. The acoustic features extracted using the final convolutional layer ($T \times 1028$) and upsampled visual features ($T \times 512$) are combined across the time dimension ($T \times 1540$) and fed to a LSTM layer with 257 units. The LSTM output is fed to two fully connected layers with 257 neurons and a ReLU activation function. It is to be noted that, the fully connected layer weights are shared across time dimensions. Finally, the extracted features are fed to a fully connected layer with 257 neurons and sigmoid activation function.

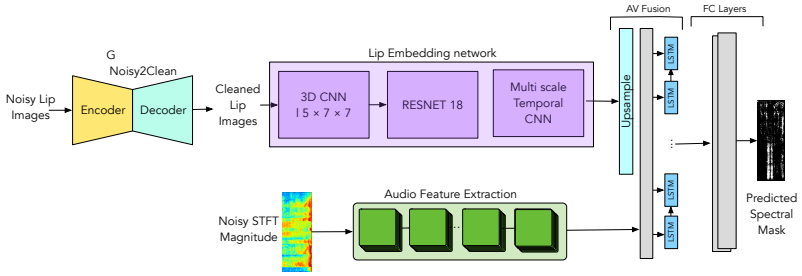



Fig. 3: Proposed Real-time Audio-Visual Speech enhancement Framework

### C. Speech Resynthesis

The model predicts the time-frequency binary mask when the noisy spectrogram and lip embeddings are fed into the network. The enhanced speech is obtained by multiplying the predicted magnitude mask with noisy magnitude spectrogram and inverse STFT. It is to be noted that, the phase of noisy signal and masked magnitude is used to resynthesize enhanced speech.

## V. Experimental Results

### A. Synthetic AV Corpora

The synthetic AV GRID corpus [31] is used for the training and evaluation of the model. All the thirty-three speakers with one thousand utterances for each speaker used. The GRID utterances are mixed with the noises from CHIME3 [32], which comprises of different types of noises including bus, street, cafeteria at SNR ranging from -12 to 9 dB with step size of 3dB. In total, the training, evaluation and test set consists of 25000, 4000 and 4000 utterances respectively. It is to be noted that, there is no overlap of speakers between train, test and validation set. In addition, the GRID corpus consists of limited vocabulary (51 words).

The TCD-TIMIT [33] corpus is used to understand the performance of the model in large vocabulary settings. TCD-TIMIT consists of 56 speakers with 5488 utterances. Each of the utterances are mixed randomly with the noises from MUSAN. In addition, for language independent evaluation Mandarin dataset [15] consists of 320 utterances combined with different types of noise from NOISEX-92 [34].

### B. Data preprocessing

*1) Audio:* In order to preprocess the audio signals, the 16 kHz mono-channel has been used. After re-sampling the audio, the signals were segmented into N 32 millisecond frames and also the 25% increment rate. The hanning window is applied to the frame to generate 257-bin STFT magnitude spectrogram.

*2) Video:* The TCD-TIMIT and GRID corpora were recorded at 25 frames per second (fps). In addition, the Mandarin dataset [15] were recorded at 30 fps, ffmpeg was used to downsample Mandarin dataset to 25 fps. In order to extract the lip images at 25 fps for the speakers dlib landmark detection model was employed. A square region around the lip-centre was extracted using landmark points. The extracted lip region is resized to a square of size 96 pixels. The cropped lip region was fed into the GAN followed by lip embedding network to extract 512 dimensional embedding for each lip image.

### C. Experimental Setup

The model was developed using PyTorch library and trained using a with Intel i9 processor, 64 GB RAM and 2 NVIDIA 2080Ti GPUs with 12 GB memory. It is to be noted that, the GAN and AV SE models were trained separately. First, the GAN is trained for 200 epochs and Adam optimiser with learning rate of 0.002 is used. The lr was not changed for the first 100 epochs. The lr was linearly decayed to 0 over the next 100 epochs. Second, the generator weights were frozen and real-time AV SE model was trained for 50 epochs with Adam optimiser (lr = $3e-4$) and batch size 8. The lr was divided by two when the validation binary cross entropy stops reducing for three consecutive epochs.

It is to be noted that, grid search was used to find the optimal hyperparameters for the GAN framework, including filter size of convolutional layers ($2^{4to10}$), stride of convolution (1 to 4), patch size for the discriminator network (16, 32, 48) and number of convolutional and deconvolutional layers (2 to 6) used in the encoder and decoder blocks respectively. The architecture of noisy audio feature extraction is adopted from CochleaNet [8] architecture by changing filter size from 96 to 64.

### D. Baseline Systems

The performance of the proposed AV SE framework with GAN (proposed AV + GAN) was compared with conventional A-only SE approaches (including SS [10] and LMMSE [11]), deep learning based A-only SE models (including SEGAN+ [12] and A-only CochleaNet [8]) and CochleaNet AV SE model [8] in a range of synthetic and real noisy scenarios. In addition, for further evaluation and ablation study two models were used: A-only version of the proposed model (proposed A-only), and AV version of proposed framework without GAN (proposed AV). Finally, oracle IBM [14] is used for comparison to understand the maximum performance the model can achieve as the framework is trained to estimate oracle IBM. It is to be noted that, the oracle IBM can only be calculated for synthetic AV datasets as perfect estimate of noise and speech spectrum is required to calculate it [14].

### E. Objective evaluation on Synthetic mixtures

In order to evaluate the performance of SE models, subjective listening tests are conducted to ask users to listen and

TABLE III: Comparison of PESQ scores for resynthesised speech. (The bold values indicates the best results a model can achieve without a perfect estimate of noise. Oracle IBM can only be calculated for synthetic AV datasets)

(a) Speaker independent test set (GRID - CHIME3)

| | dB | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Models** | -12 | -9 | -6 | -3 | 0 | 3 | 6 | 9 |
| Noisy | 1.31 | 1.40 | 1.55 | 1.70 | 1.88 | 2.08 | 2.28 | 2.46 |
| SS [10] | 1.13 | 1.23 | 1.41 | 1.60 | 1.83 | 2.09 | 2.35 | 2.58 |
| LMMSE [11] | 1.36 | 1.52 | 1.74 | 1.96 | 2.17 | 2.40 | 2.59 | 2.76 |
| SEGAN+ [12] | 0.83 | 1.07 | 1.45 | 1.80 | 2.12 | 2.38 | 2.58 | 2.76 |
| CochleaNet A-only [8] | 1.85 | 2.04 | 2.24 | 2.40 | 2.53 | 2.64 | 2.74 | 2.81 |
| CochleaNet AV [8] | 1.87 | 2.07 | 2.23 | 2.37 | 2.48 | 2.59 | 2.68 | 2.76 |
| Proposed A-only | 1.88 | 2.07 | 2.24 | 2.37 | 2.54 | 2.63 | 2.75 | 2.82 |
| Proposed AV | 1.91 | 2.09 | 2.25 | 2.40 | 2.55 | 2.65 | 2.77 | 2.88 |
| **Proposed AV + GAN** | **1.95** | **2.11** | **2.28** | **2.42** | **2.55** | **2.66** | **2.78** | **2.89** |
| Oracle IBM [14] | 2.03 | 2.20 | 2.34 | 2.47 | 2.59 | 2.70 | 2.82 | 2.91 |

(b) Large vocabulary test set (TCD-TIMIT + MUSAN)

| | dB | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Models** | -12 | -9 | -6 | -3 | 0 | 3 | 6 | 9 |
| Noisy | 1.48 | 1.56 | 1.62 | 1.69 | 2.23 | 2.33 | 2.44 | 2.50 |
| SS [10] | 1.08 | 1.13 | 1.19 | 1.27 | 1.79 | 1.89 | 2.06 | 2.18 |
| LMMSE [11] | 1.44 | 1.45 | 1.61 | 1.62 | 2.04 | 2.15 | 2.25 | 2.34 |
| SEGAN+ [12] | 1.60 | 1.74 | 1.77 | 1.84 | 2.32 | 2.44 | 2.58 | 2.66 |
| CochleaNet A-only [8] | 1.81 | 1.90 | 2.02 | 2.13 | 2.37 | 2.47 | 2.52 | 2.56 |
| CochleaNet AV [8] | 1.90 | 2.00 | 2.12 | 2.18 | 2.48 | 2.56 | 2.62 | 2.66 |
| Proposed A-only | 1.92 | 2.03 | 2.13 | 2.44 | 2.50 | 2.59 | 2.62 | 2.68 |
| Proposed AV | 1.98 | 2.09 | 2.20 | 2.52 | 2.59 | 2.63 | 2.67 | 2.70 |
| **Proposed AV + GAN** | **2.05** | **2.17** | **2.27** | **2.60** | **2.67** | **2.70** | **2.74** | **2.78** |
| Oracle IBM [14] | 2.16 | 2.29 | 2.39 | 2.74 | 2.81 | 2.84 | 2.89 | 2.92 |

(c) Language-independent test set (Hou et al. [15] + NOISEX92)

| | dB | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Models** | -12 | -9 | -6 | -3 | 0 | 3 | 6 | 9 |
| Noisy | 1.04 | 1.25 | 1.29 | 1.31 | 1.40 | 1.49 | 1.55 | 1.61 |
| SS [10] | 0.63 | 1.06 | 0.99 | 0.98 | 1.28 | 1.23 | 1.36 | 1.34 |
| LMMSE [11] | 1.21 | 1.42 | 1.39 | 1.40 | 1.40 | 1.44 | 1.61 | 1.44 |
| SEGAN+ [12] | 1.14 | 1.30 | 1.16 | 1.45 | 1.59 | 1.66 | 1.71 | 1.74 |
| CochleaNet A-only [8] | 1.28 | 1.42 | 1.56 | 1.53 | 1.66 | 1.72 | 1.79 | 1.74 |
| CochleaNet AV [8] | 1.23 | 1.45 | 1.44 | 1.46 | 1.66 | 1.68 | 1.74 | 1.75 |
| Proposed A-only | 1.32 | 1.53 | 1.59 | 1.61 | 1.70 | 1.73 | 1.79 | 1.78 |
| Proposed AV | 1.39 | 1.55 | 1.62 | 1.63 | 1.71 | 1.74 | 1.79 | 1.80 |
| **Proposed AV + GAN** | **1.44** | **1.57** | **1.65** | **1.58** | **1.70** | **1.75** | **1.79** | **1.81** |
| Oracle IBM [14] | 1.55 | 1.69 | 1.77 | 1.70 | 1.83 | 1.88 | 1.92 | 1.95 |

TABLE IV: Comparison of STOI scores for resynthesised speech (The bold values indicates the best results a model can achieve without a perfect estimate of noise. Oracle IBM can only be calculated for synthetic AV datasets)

(a) Speaker independent test set (GRID - CHIME3)

| | dB | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Models** | -12 | -9 | -6 | -3 | 0 | 3 | 6 | 9 |
| Noisy | 0.41 | 0.45 | 0.49 | 0.54 | 0.59 | 0.64 | 0.68 | 0.72 |
| SS [10] | 0.36 | 0.40 | 0.45 | 0.50 | 0.56 | 0.61 | 0.67 | 0.71 |
| LMMSE [11] | 0.39 | 0.43 | 0.48 | 0.53 | 0.58 | 0.63 | 0.68 | 0.71 |
| SEGAN+ [12] | 0.31 | 0.39 | 0.49 | 0.58 | 0.65 | 0.70 | 0.74 | 0.76 |
| CochleaNet A-only [8] | 0.51 | 0.57 | 0.59 | 0.63 | 0.70 | 0.74 | 0.76 | 0.77 |
| CochleaNet AV [8] | 0.53 | 0.58 | 0.62 | 0.66 | 0.70 | 0.73 | 0.75 | 0.77 |
| Proposed A-only | 0.55 | 0.60 | 0.62 | 0.65 | 0.69 | 0.74 | 0.76 | 0.78 |
| Proposed AV | 0.57 | 0.61 | 0.64 | 0.67 | 0.71 | 0.75 | 0.77 | 0.78 |
| **Proposed AV + GAN** | **0.59** | **0.63** | **0.66** | **0.69** | **0.72** | **0.76** | **0.78** | **0.78** |
| Oracle IBM [14] | 0.61 | 0.65 | 0.68 | 0.71 | 0.74 | 0.77 | 0.79 | 0.80 |

(b) Large vocabulary test set (TCD-TIMIT + MUSAN)

| | dB | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Models** | -12 | -9 | -6 | -3 | 0 | 3 | 6 | 9 |
| Noisy | 0.31 | 0.34 | 0.43 | 0.50 | 0.60 | 0.64 | 0.70 | 0.73 |
| SS [10] | 0.30 | 0.35 | 0.42 | 0.46 | 0.59 | 0.62 | 0.70 | 0.73 |
| LMMSE [11] | 0.46 | 0.48 | 0.53 | 0.55 | 0.66 | 0.69 | 0.73 | 0.76 |
| SEGAN+ [12] | 0.44 | 0.47 | 0.52 | 0.55 | 0.65 | 0.68 | 0.73 | 0.77 |
| CochleaNet A-only [8] | 0.48 | 0.51 | 0.54 | 0.61 | 0.67 | 0.70 | 0.75 | 0.78 |
| CochleaNet AV [8] | 0.51 | 0.55 | 0.60 | 0.61 | 0.71 | 0.73 | 0.76 | 0.79 |
| Proposed A-only | 0.50 | 0.52 | 0.59 | 0.62 | 0.72 | 0.74 | 0.78 | 0.80 |
| Proposed AV | 0.60 | 0.62 | 0.65 | 0.69 | 0.73 | 0.74 | 0.79 | 0.81 |
| **Proposed AV + GAN** | **0.64** | **0.66** | **0.68** | **0.72** | **0.74** | **0.75** | **0.80** | **0.81** |
| Oracle IBM [14] | 0.72 | 0.74 | 0.77 | 0.79 | 0.81 | 0.82 | 0.84 | 0.85 |

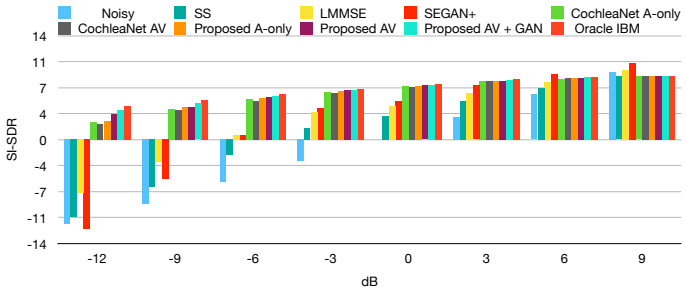
(c) Language-independent test set (Hou et al. [15] + NOISEX92)

| | dB | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Models** | -12 | -9 | -6 | -3 | 0 | 3 | 6 | 9 |
| Noisy | 0.54 | 0.71 | 0.68 | 0.73 | 0.78 | 0.78 | 0.85 | 0.86 |
| SS [10] | 0.42 | 0.58 | 0.58 | 0.62 | 0.70 | 0.72 | 0.77 | 0.80 |
| LMMSE [11] | 0.52 | 0.70 | 0.66 | 0.71 | 0.76 | 0.75 | 0.83 | 0.82 |
| SEGAN+ [12] | 0.52 | 0.66 | 0.58 | 0.70 | 0.76 | 0.76 | 0.82 | 0.85 |
| CochleaNet A-only [8] | 0.54 | 0.73 | 0.70 | 0.76 | 0.81 | 0.82 | 0.86 | 0.88 |
| CochleaNet AV [8] | 0.56 | 0.73 | 0.70 | 0.75 | 0.81 | 0.80 | 0.85 | 0.87 |
| Proposed A-only | 0.58 | 0.75 | 0.71 | 0.77 | 0.82 | 0.81 | 0.87 | 0.88 |
| Proposed AV | 0.65 | 0.77 | 0.74 | 0.79 | 0.84 | 0.85 | 0.87 | 0.88 |
| **Proposed AV + GAN** | **0.70** | **0.80** | **0.77** | **0.81** | **0.86** | **0.86** | **0.87** | **0.88** |
| Oracle IBM [14] | 0.81 | 0.88 | 0.87 | 0.90 | 0.92 | 0.92 | 0.94 | 0.94 |

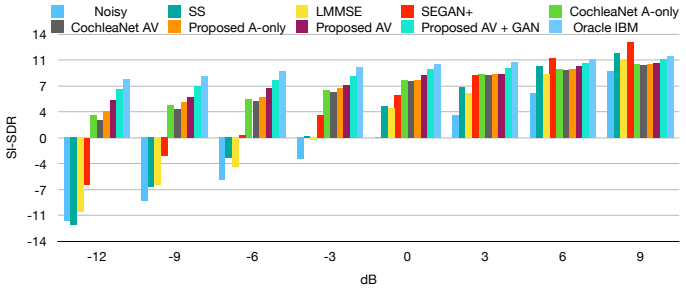
compare the speech quality difference between the processed and unprocessed audio samples. However, conducting subjective listening tests is time consuming for large datasets and the results may not accurately represent the actual distribution. In such scenarios, PESQ [35], STOI [36], and SI-SDR [37] are used as objective evaluation metrics to approximate subjective listening tests. The proposed model has been compared with the baseline systems presented in section V-D. It is to be noted that, GRID + CHIME3, TCD TIMIT + MUSAN and Hou et al. [15] + NOISEX-92 are used for speaker independent, large-vocabulary and language independent evaluation of the proposed AV SE model.

*1) Perceptual Evaluation of Speech quality (PESQ) comparison:* PESQ [35] is one the most well-known evaluation metric used to predict the subjective listening test scores in the SE and preliminary results display that correlate well with the subjective listening tests [38]. The PESQ scores for speaker-independent, large-vocabulary and language independent test set are presented in Table IIIa, IIIb, IIIc respectively. The PESQ scores shows that, the proposed AV model outperforms all state-of-the-art SE models including AV CochleaNet [8], SEGAN [12], A-only version of proposed models, and AV version of the proposed model without GAN. It can be seen that, AV + GAN outperforms A-only model in low SNR particularly $SNR < 0$ dB, mainly where the AV model achieved PESQ score of 1.95 (-12 dB), 2.11 (-9 db), and 2.28 (-6db). Whereas, A-only model achieved PESQ score of 1.88 (-12 dB), 2.07 (-9 db), and 2.24 (-6dB) for speaker independent GRID CHIME3 test set. On the other hand, in the high SNR ($SNR >= 0$ dB) AV + GAN performs similar to A-only model. Specifically, AV achieved 2.55, 2.66 and 2.78 PESQ scores for 0dB, 3dB and 6dB respectively. The A-only model achieved PESQ score of 2.54, 2.63, and 2.75 for 0dB, 3dB and 6dB respectively.

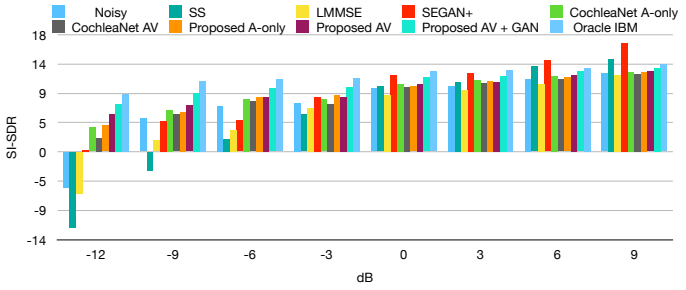
*2) Short Term Objective Intelligibility (STOI) comparison:* STOI is one of the most widely used alternative to PESQ that shows high correlation with subjective listening tests [36]. The STOI scores for speaker-independent, large-vocabulary and language independent test set is presented in Table IVa, IVb, IVc respectively. The STOI scores shows that the proposed AV model outperforms state-of-the-art SE models including DNN based CochleaNet [8], SEGAN [12], A-only version of proposed models, and AV version of the proposed




(a) Speaker independent test (GRID + CHIME3)




(b) Large vocabulary test set (TCD-TIMIT + MUSAN)




(c) Language independent test set (Hou et al. [15] + NOISEX-92)

Fig. 4: Comparison of SI-SDR scores for resynthesised speech

model without GAN. It can be seen that, AV + GAN model obtained STOI score of 0.59 (-12 dB), 0.63 (-9 dB) and 0.66 (-6 dB), as compared to 0.55 (-12 dB), 0.60 (-9 dB) and 0.62 (-6 dB) STOI score obtained by A-only model for speaker independent test set. However, at high SNRs (i.e., $SNR >= 0$ dB) proposed AV + GAN model performs similar to A-only model, where AV + GAN model achieved STOI score of 0.69 (0 dB), 0.74 (3 dB), and 0.76 (6 dB), as compared to 0.65 (0 dB), 0.69 (3 dB), and 0.74 (0 dB) achieved by A-only model for speaker independent test set. In addition, it has been shown that the A-only and AV + GAN model significantly outperform A-only and AV CochleaNet.

*3) Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) comparison:* SI-SDR is used to predict the distortion introduced by the separated signal and usually defined as the ratio between clean signal and distortion energy. The higher SDR shows the better quality of SE and less distortion in the enhanced speech. Fig. 4a, 4b, 4c depict the SI-SDR for speaker independent, large vocabulary and language independent test set. The SI-SDR scores shows that, the proposed AV model outperforms state-of-the-art SE models including DNN based

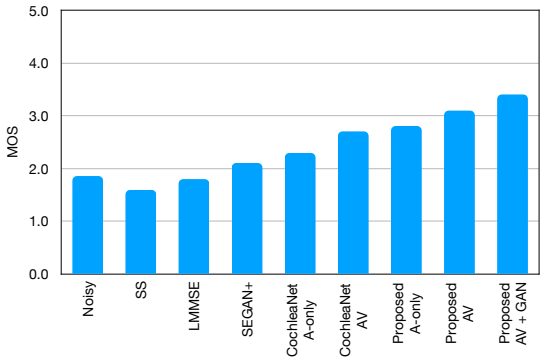



Fig. 5: Subjective evaluation using mean-opinion score listening test using real noisy ASPIRE and VISION corpora for the resynthesised speech

CochleaNet [8], SEGAN+ [12], A-only version of proposed models, and AV version of the proposed model without GAN. It can be seen that, for the lower SNR ($SNR < 0$ dB) the proposed AV outperforms A-only model especially, where AV + GAN model achieved the SI-SDR score of 3.50 (-12 dB), 4.45 (-9 dB), and 5.04 (-6dB) as compared to 2.53 (-12 dB), 4.39 (-9 dB), and 5.62 (-6 dB) SI-SDR score obtained by A-only model for speaker independent test set. However, AV + GAN perform similar to A-only model at high SNRs levels (i.e. $SNR >= 0$ dB), where AV model achieved SI-SDR score of 7.35 (0 dB), 8.01 (3 dB) and 8.35 (6dB), as compared to 7.25, 7.97 and 8.35 achieved by A-only model for speaker independent test set. It is worth mentioning that, the proposed A-only and AV + GAN model perform better than A-only and AV CochleaNet.

### *F. Subjective listening tests on ASPIRE and VISION corpus*

The subjective speech quality can be computationally approximated using state-of-the-art objective evaluation metrics including PESQ and STOI [35, 36]. However, human listening tests need to be conducted to accurately understand subjective speech quality. In this study, the mean opinion score (MOS) type listening test was conducted for comparative subjective evaluation. The data used for these tests include ASPIRE [8] and VISION corpora [9] recorded in a real noisy environment with a range of visual and acoustic noises. Twenty native English speakers (twelve men and eight women) with normal hearing volunteered to participate in the listening test. The listeners were first trained with five utterances and the purpose of the study was explained. In each individual listening test, twenty utterances chosen at random from the ASPIRE and VISION corpora were played. Listeners were asked to rate the enhanced speech quality on a scale of 0 to 5, with 0 representing incomprehensible, 1 representing very annoying, 2 representing annoying, 3 representing slightly annoying, 4 representing perceptible but annoying, and 5 representing perceptible. SEGAN, SS, LMMSE, A-only and AV CochleaNet, proposed A-only, proposed AV and proposed AV+GAN are comparatively evaluated along with noisy audio as reference. The speech quality scores for the aforementioned models are presented in Fig.5. It can be seen that, the proposed A-only

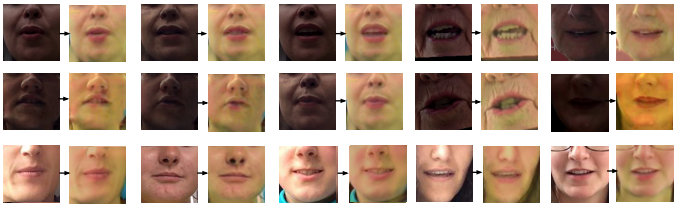


Fig. 6: Visual Speech Enhancement Evaluation on Unseen speakers from VISION Corpus

and AV models significantly outperform SS, LMMSE, A-only and AV CochleaNet. It can be seen that, the proposed AV model can handle the visual imperfections present in VISION corpora as compared to state-of-the-art CochleaNet model.

### *G. Qualitative evaluation of Proposed GAN model*

The GAN model is evaluated on the unseen utterances from real Noisy VISION corpus. The real noisy VISION corpus consists of a number of visual imperfections like improper lighting, exposure and occlusions. Fig. 6 shows some of the unprocessed and processed sample frames from real noisy VISION corpus. It can be observed that, the model significantly enhanced visual speech noise present in the VISION corpus. The developed visual SE model will enable deployment of real-time AV SE models in a variety of real-world settings.

### *H. Effect of Visual Speech Enhancement on Intelligibility*

In real world settings, visual speech is often degraded by a range of visual imperfection including poor lighting and head movements. In order to evaluate the behaviour of proposed model in such settings, the visual speech from test set of GRID-CHIME3 was degraded using random transformations including changing brightness, contrast and exposure. The noisy visual speech images and audio are fed to the proposed AV SE model. For further evaluation, then randomly replaced a percentage of noisy lip image with corresponding enhanced lip image generated using GAN. The results for the effect of visual SE on intelligibility are depicted in Fig. 7. It can be seen that, for the case of both -6 dB and -12 dB, the model achieves performance similar to visually intact data when all of the degraded images are enhanced using the proposed GAN. It should be noted that, the model performance is slightly worse than A-only model when the noisy visual speech images are fed without denoising.

### *I. Spectrogram comparison*

Fig. 8 illustrates the spectrogram for the resynthesised speech signal of a randomly selected utterance from GRID - CHIME3 AV corpus using proposed A-only and AV models as well as state-of-the-art approaches including SS [10], LMMSE [11], SEGAN+ [12] A-only CochleaNet, and AV CochleaNet [8]. In addition, spectrogram for clean and noisy speech signals is shown for comparison. It is to be noted that, the speech is completely swamped with street noise and the performance of the proposed model is closer to clean spectrogram. The state-of-the-art A-only approaches were unable to

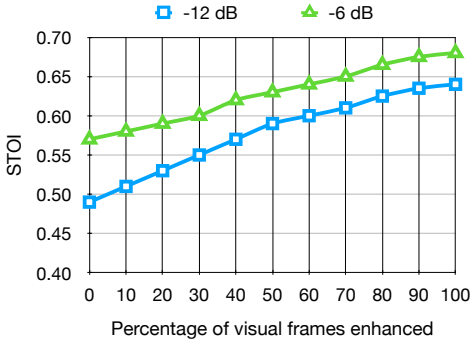



Fig. 7: Short-Time Objective Intelligibility (STOI) for different percentage of visual frames enhanced

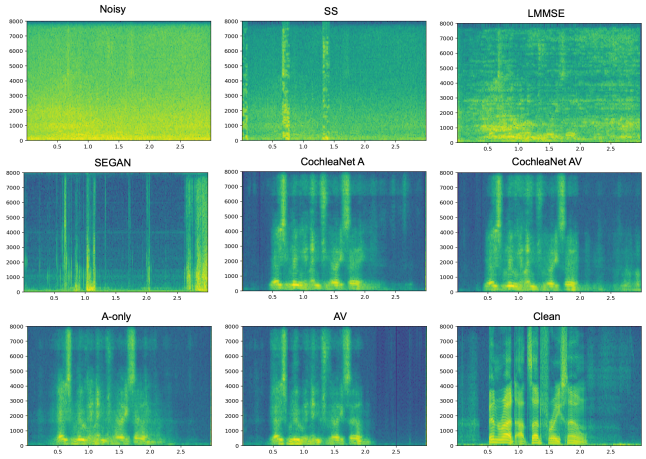



Fig. 8: Spectrogram comparison of a randomly selected utterance at -12 dB SNR from GRID-CHIME3 corpus. It can be seen that proposed AV model recovers frequency components better than state-of-the-art SE models

recover the speech components from the noisy signal. Finally, it can be seen the AV models were able to better reconstruct target speech as compared to A-only models specifically in silent speech regions.

### *J. Processing latency*

The processing latency for a SE algorithm can be defined as the difference between the time of arrival of speech and the time when the model finishes processing the noisy speech. The latency is generally measured in milliseconds (ms). The ideal processing latency of a listening device depends on the severity of the hearing loss. The latency of most commercial hearing aids generally ranges from 8-12 ms. The processing latency of the proposed model is 15 ms. The latency is dependent upon the Fourier window shift (8 ms), STFT (0.5ms), ISTFT (0.5ms), and model prediction time (6 ms). These values are calculated with a M1 Macbook Pro with 16 GB RAM. The processing latency is mainly affected by the shift of Fourier window and the model processing time. The model processing time can be further optimised using mixed precision processing and quantization. In addition, the window shift can be optimised to further decrease the latency. Currently, the model

is being tested for noise suppression in video conferencing scenario.

### K. Limitations

The limitations with the proposed framework are: (1) the framework is not privacy preserving. Privacy preserving visual features could be explored (2) the resynthesised speech using IBM ignores phase estimation and result in invalid STFT problem (3) the STFT used for transforming time-domain signals to frequency domain has fixed temporal resolution. Other types of transform including discrete wavelet transform [39], discrete Tchebichef transform [40] and the discrete Krawtchouk transform [40] could be explored to address this limitation (4) the model currently do not consider the noise level in the environment for contextual switching between A-only and AV model (5) the enhanced speech cannot be localised in the space as the model only supports processing single channel data

## VI. Conclusion

In this paper, a novel framework is presented for robust real-time AV SE that contextually takes account of both visual and acoustic speech noises. Specifically, a GAN architecture is developed to enhance the visual speech imperfections encountered in real noisy environments. Further, the GAN is integrated with a real-time AV SE model to contextually exploit noisy visual and acoustic speech to suppress noise dominant regions and enhance speech dominant regions. The model is evaluated on benchmark visual speech noises from real world recordings that consists of noisy speech recorded in the presence of multiple competing background sources. Preliminary performance evaluation in terms of objective metrics and subjective listening tests demonstrates significant improvement of the proposed AV SE framework compared to the state-of-the-art A-only (including SS, LMMSE) approaches as well as DNN based AV approaches (including benchmark SEGAN and CochleaNet models). Comparative experimental results indicate that, the proposed framework has the ability to work effectively in a range of SNRs with both visual and acoustic noise and can be deployed in real-world environments. Ongoing work involves optimisation of trade offs between generalisation, latency and energy for deployment of proposed framework in web communication scenario. In future, we will address the current limitations of the proposed model and further investigate its generalisation capability using datasets collected in the wild. In addition, we will explore alternatives to Fourier transform including wavelet transform, and short-time discrete cosine transform for robust AV SE. Finally, privacy concerns associated with AV SE based assistive devices will be addressed by exploiting privacy preserving visual features and homomorphic encryption.

## VII. Acknowledgements

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) programme grant: COG-MHEAR (Grant reference EP/T021063/1).

## References

[1] World Health Organization et al. “Hearing screening: considerations for implementation”. In: (2021).

[2] Nathalie Moermans et al. “Recovery of natural hearing after cochlear implantation in a case of ISSNHL”. In: *Laryngo-Rhino-Otologie* 102 (2023), S296–S296.

[3] Nicholas A Lesica. “Why do hearing aids fail to restore normal auditory perception?” In: *Trends in neurosciences* 41.4 (2018), pp. 174–185.

[4] Chris Donahue, Bo Li, and Rohit Prabhavalkar. “Exploring speech enhancement with generative adversarial networks for robust speech recognition”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5024–5028.

[5] Daniel Michelsanti et al. “An overview of deep-learning-based audio-visual speech enhancement and separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1368–1396.

[6] Guinan Li et al. “Audio-visual multi-channel speech separation, dereverberation and recognition”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6042–6046.

[7] Tassadaq Hussain et al. “A Novel Temporal Attentive-Pooling based Convo-lutional Recurrent Architecture for Acoustic Signal Enhancement”. In: *IEEE Transactions on Artificial Intelligence* (2022).

[8] Mandar Gogate et al. “CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement”. In: *Information Fusion* 63 (2020), pp. 273–285.

[9] Mandar Gogate, Kia Dashtipour, and Amir Hussain. “Visual Speech In Real Noisy Environments (VISION): A Novel Benchmark Dataset and Deep Learning-Based Baseline System”. In: *Interspeech 2020*. ISCA. 2020, pp. 4521–4525.

[10] S Boll. “A spectral subtraction algorithm for suppression of acoustic noise in speech”. In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’79.* Vol. 4. IEEE. 1979, pp. 200–203.

[11] Yariv Ephraim and David Malah. “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator”. In: *IEEE transactions on acoustics, speech, and signal processing* 33.2 (1985), pp. 443–445.

[12] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. “SEGAN: Speech Enhancement Generative Adversarial Network”. In: *Proc. Interspeech 2017* (2017), pp. 3642–3646.

[13] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. “The Conversation: Deep Audio-Visual Speech Enhancement”. In: *Proc. Interspeech 2018* (2018), pp. 3244–3248.

[14] Mandar Gogate et al. “DNN Driven Speaker Independent Audio-Visual Mask Estimation for Speech Sepa-

ration". In: *Interspeech 2018*. ISCA. 2018, pp. 2723–2727.

[15] Jen-Cheng Hou et al. "Audio-visual speech enhancement using multimodal deep convolutional neural networks". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 2.2 (2018), pp. 117–128.

[16] Rui Lu, Zhiyao Duan, and Changshui Zhang. "Audio–visual deep clustering for speech separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.11 (2019), pp. 1697–1712.

[17] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. "My Lips Are Concealed: Audio-Visual Speech Enhancement Through Obstructions". In: *Proc. Interspeech 2019* (2019), pp. 4295–4299.

[18] Daniel Michelsanti et al. "Deep-learning-based audio-visual speech enhancement in presence of Lombard effect". In: *Speech Communication* 115 (2019), pp. 38–50.

[19] Ahsan Adeel, Mandar Gogate, and Amir Hussain. "Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments". In: *Information Fusion* 59 (2020), pp. 163–170.

[20] Ander Arriandiaga et al. "Audio-visual target speaker enhancement on multi-talker environment using event-driven cameras". In: *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2021, pp. 1–5.

[21] Ruohan Gao and Kristen Grauman. "Visualvoice: Audio-visual speech separation with cross-modal consistency". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2021, pp. 15490–15500.

[22] Tassadaq Hussain et al. "Towards intelligibility-oriented audio-visual speech enhancement". In: *Proc. Clarity Workshop on Machine Learning Challenges for Hearing Aids*. 2021.

[23] Mandar Gogate, Kia Dashtipour, and Amir Hussain. "Towards real-time privacy-preserving audio-visual speech enhancement". In: *Proc. 2nd Symposium on Security and Privacy in Speech Communication*. 2022, pp. 7–10. DOI: 10.21437/SPSC.2022-2.

[24] Shang-Yi Chuang, Hsin-Min Wang, and Yu Tsao. "Improved lite audio-visual speech enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 1345–1359.

[25] Zirun Zhu et al. "Real-Time Audio-Visual End-To-End Speech Enhancement". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[26] Rodrigo Mira et al. "LA-VocE: Low-SNR Audio-visual Speech Enhancement using Neural Vocoders". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[27] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[28] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[29] Brais Martinez et al. "Lipreading using Temporal Convolutional Networks". In: *ICASSP*. 2020.

[30] J. S. Chung and A. Zisserman. "Lip Reading in the Wild". In: *Asian Conference on Computer Vision*. 2016.

[31] Martin Cooke et al. "An audio-visual corpus for speech perception and automatic speech recognition". In: *The Journal of the Acoustical Society of America* 120.5 (2006), pp. 2421–2424.

[32] Jon Barker et al. "The third 'CHiME'speech separation and recognition challenge: Dataset, task and baselines". In: *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE. 2015, pp. 504–511.

[33] N. Harte and E. Gillen. "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech". In: *IEEE Transactions on Multimedia* 17.5 (May 2015), pp. 603–615. ISSN: 1520-9210. DOI: 10.1109/TMM.2015.2407694.

[34] Andrew Varga and Herman JM Steeneken. "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems". In: *Speech communication* 12.3 (1993), pp. 247–251.

[35] Antony W Rix et al. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2. IEEE. 2001, pp. 749–752.

[36] Cees H Taal et al. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2125–2136.

[37] Jonathan Le Roux et al. "SDR–half-baked or well done?" In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 626–630.

[38] Yi Hu and Philipos C Loizou. "Evaluation of objective quality measures for speech enhancement". In: *IEEE Transactions on audio, speech, and language processing* 16.1 (2007), pp. 229–238.

[39] Wissam A Jassim, Raveendran Paramesran, and Muhammad SA Zilany. "Enhancing noisy speech signals using orthogonal moments". In: *IET Signal Processing* 8.8 (2014), pp. 891–905.

[40] Li Wang et al. "Denoising speech based on deep learning and wavelet decomposition". In: *Scientific Programming* 2021 (2021), pp. 1–10.