CochleaNet: A Robust Language-independent Audio-Visual Model for Speech Enhancement

Mandar Gogate[a], Kia Dashtipour[b], Ahsan Adeel[c], Amir Hussain[a,*]

[a] *Edinburgh Napier University, School of Computing, Edinburgh, EH10 5DT, UK*
[b] *University of Stirling, Division of Computing Science and Maths, Stirling, FK9 4LA, UK*
[c] *University of Wolverhampton, School of Mathematics and Computer Science, Wolverhampton, UK*

Abstract

Noisy situations cause huge problems for suffers of hearing loss as hearing aids often make speech more audible but do not always restore the intelligibility. In noisy settings, humans routinely exploit the audio-visual (AV) nature of speech to selectively suppress the background noise and focus on the target speaker. In this paper, we present a language, noise and speaker independent AV deep neural network (DNN) architecture for causal or real-time speech enhancement (SE). The model jointly exploits the noisy acoustic cues and noise robust visual cues to focus on the desired speaker and improve speech intelligibility. The proposed SE framework is evaluated using a first of its kind AV binaural speech corpus, called ASPIRE, recorded in real noisy environments including cafeteria and restaurant. We demonstrate superior performance of our approach in terms of objective measures and subjective listening tests over the state-of-the-art SE approaches as well as recent DNN based SE models. In addition, our work challenges a popular belief that, scarcity of multi-language large vocabulary AV corpus and a wide variety of noises is a major bottleneck to build a robust language, speaker and noise independent SE systems. We show that a model trained on synthetic mixture of Grid corpus (with 33 speakers and a small English vocabulary) and ChiME 3 Noises (consisting of bus, pedestrian, cafeteria,

*Corresponding author
Email address: a.hussain@napier.ac.uk (Amir Hussain)

and street noises) generalise well not only on large vocabulary corpora, wide variety of speakers/noises but also on completely unrelated language (such as Mandarin).

Keywords: `Audio-Visual, Speech Enhancement, Speech Separation, Deep Learning, Real Noisy Audio-Visual Corpus, Speaker Independent, Causal`

---

1. Introduction

The human brain integrates the available heterogeneous information received from the five sensory organs (ears, eyes, nose, tongue and skin) with prior contexts to perform day-to-day cognitive tasks including vision, hearing, logic and reasoning. In the literature, the integration of multiple modalities have shown significant performance improvements as compared to unimodal systems [1] in terms of acquiring cognitive functionalities. For example, during busy social gatherings, human brain integrates the acoustic and visual cues in order to better perceive speech. This multi-sensory hearing phenomenon was first demonstrated in the McGurk effect [1] where a visual $/ga/$ with a voiced $/ba/$ is perceived as $/da/$ by most subjects. In particular, the visual cues provide information on the place of articulation [2] and muscle movements which can often aid to differentiate between speech with similar acoustic sounds (e.g. the unvoiced consonants $/p/$ and $/k/$) or phonological ambiguities [3]. In addition, further studies have shown the importance of visual cues in improving the speech intelligibility as well as speech detection in noisy environments [4, 5].

In the recent years, speech enhancement (SE) has attracted wide attention due to the noise reducing ability that helps hearing impaired listen better in noisy social situations and opened the doors for speech processing systems (such as speech recognition and voice activity detector systems) in noisy environments [6, 7]. SE approaches can be categorised into statistical analysis based noise reduction models such as spectral subtraction (SS), linear minimum mean square error (LMMSE), Wiener filtering and computational auditory scene anal-

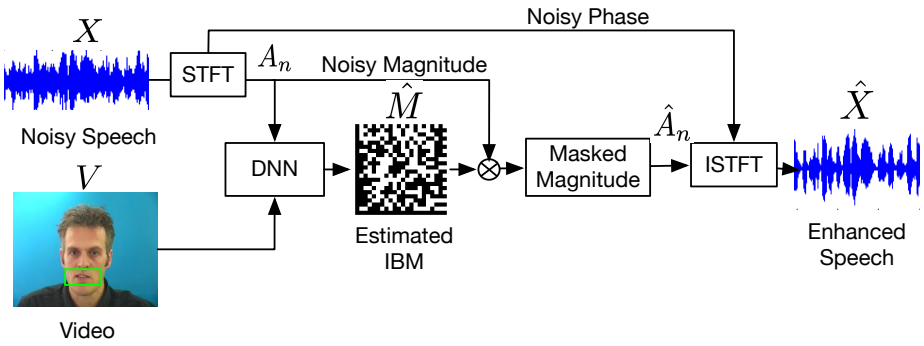


Figure 1: CochleaNet Framework: Audio-Visual Mask Estimation based Speech Enhancement

ysis (CASA) [8]. It has been observed that the statistical methods fail to achieve improved speech intelligibility in some scenarios due to introduction of distortions such as musical noises. In contrast, CASA has shown to be more effective in stationary and non-stationary noises [9]. In CASA, the speech is separated from interfering background noise by using a time-frequency (T-F) spectral mask to the T-F representation of noisy speech. The T-F spectral mask is used to enhance speech dominant regions and suppress the noise-dominant regions.

In the literature, extensive research has been carried out to develop audio-only (A-only) and audio-visual (AV) SE methods. Researchers have proposed several SE models such as deep neural network (DNN) based spectral mask estimation models [10, 11], DNN based clean spectrogram estimation models [12, 13], Wiener filtering based hybrid SE models [14, 15, 16], and time-domain SE models [17, 18, 19]. However, limited work has been conducted to develop robust language, noise and speaker independent AV SE models for low SNRs (< -3 dB) observed in everyday social environments (such as cafeteria, railway stations and restaurants) where traditional A-only hearing aids fail to improve the speech intelligibility. The few attempts to develop such robust models have been limited to speaker-dependent scenarios [13] and small scale (< 5 speakers) speaker independent scenarios [11, 16].

In addition, none of the aforementioned AV SE studies have conducted listen-

ing tests on real noisy mixtures that often consists of speech signal reverberantly mixed with multiple competing background noise sources [20]. Finally, studies have shown that a pretrained DNN based SE model does not generalise well on new languages [21]. The model can be fine-tuned on large AV corpus consisting of wide variety of languages such as AVSPEECH [10] (consisting of 1500 hours recording) to potentially achieve the language-independent performance given enough model capacity. However, training on corpora like AVSPEECH requires a large number of graphics processing units (GPUs) or tensor processing units (TPUs) that are often unavailable in academic research environments.

In this paper, we present a language, noise and speaker independent AV model to focus on a target speaker by selectively suppressing the background noise. More specifically, we design and train a cross-modal DNN architecture, called CochleaNet, that ingests the noisy sound mixture and cropped images of speakers lip as an input and output a T-F mask to selectively suppress and enhance each T-F bin. In addition, the model contextually exploits the available AV cues to estimate the spectral mask independent of the SNRs.

The proposed AV SE model is evaluated using, ASPIRE, a first of its kind high quality AV binaural speech corpus recorded in real noisy settings such as cafeteria and restaurant. It is to be noted that, most of the aforementioned AV SE methods used a synthetic mixture of clean speech and noises for model evaluation. However, the synthetic mixture do not reflect the real noisy mixtures as speech is often reverberantly mixed with multiple competing noise background sources. Therefore, the ASPIRE corpus can be used by speech and machine learning communities as a benchmark resource to support reliable evaluation of AV SE technologies.

We demonstrate superior speech quality and intelligibility of proposed approach over the state-of-the-art A-only SE approaches (including SS, LMMSE) as well as recent DNN based SE models (including SEGAN) using real noisy ASPIRE corpus. In addition, we show that a model trained on a synthetic mixture of Grid corpus [22] (with only 33 speakers and a small English vocabulary) and ChiME 3 [20] noises (consisting of bus, pedestrian, cafe, and street noises)

generalise well on real noisy ASPIRE corpus, large vocabulary corpora (such as TCD-TIMIT [23]), other languages (such as Mandarin [13]) and wide variety of speakers and noises [24, 25]. An overview of our proposed AV SE model is shown in Figure 1.

In summary, this paper presents four major contributions:

(i) A language, noise and speaker independent AV DNN driven model for causal or real-time SE is proposed. To the best of our knowledge, our paper is first to propose a model that generalises on different languages even after training on a small English vocabulary Grid corpus. In the literature, it has been shown that a pretrained SE model trained on a single language does not perform well on new languages [21].

(ii) A first of its kind AV corpus, consisting of high quality binaural speech recorded in real noisy environments such as cafeteria and restaurant, is collected to evaluate the performance of the proposed model in challenging real noisy settings. In the literature, a synthetic mixture of clean speech and noise is generally used to evaluate the AV SE methods. However, the synthetic mixtures do not depict the real noisy mixtures as in real mixtures the speech is often reverberantly mixed with multiple competing noise background sources.

(iii) We perform extensive evaluation of our proposed approach, using real noisy ASPIRE corpus, with state-of-the-art A-only SE approaches (including SS, LMMSE) as well as recent DNN based SE models (including SEGAN) using objective measures (PESQ, SI-SDR, and ESTOI) and subjective MUSHRA listening tests.

(iv) We critically analyse and compare the performance of audio-only model with the audio-visual counterpart to empirically identify the role visual cues plays in the performance of audio-visual model. Specifically, we study the behaviour of the audio-only and audio-visual models in silent speech regions as well as we conduct listening tests to gauge the model perfor-

mances on different phonemes. We hypothesise that the model performs better on visually distinguishable phonemes as compared to visually indistinguishable phonemes. Finally, we study the behaviour of the trained AV model, in terms of objective metrics, when the visual cues are temporarily or permanently absent for random duration of time due to occlusions.

The rest of the paper is organised as follows: Section 2 briefly reviews the related work, section 3 presents the ASPIRE corpus collection setup and the postprocessing involved. Section 4 presents, CochleaNet, an AV Mask Estimation model for SE. Section 5 discuss the experimental setup and results. Section 6 concludes this work and propose future research directions.

2. Related work

This section briefly reviews the related works in the area of A-only and AV SE.

2.1. Audio-Visual Speech Enhancement

Ephrat et al. [10] proposed a speaker independent AV DNN for complex ratio mask estimation to separate speech from overlapping speech and background noises. The model is trained on, AVSPEECH, a new large AV corpus consisting of 1500 hours recording with wide variety of languages, people and face pose. The main limitation, with the aforementioned study, is that the model is trained and evaluated on a fixed SNR. Similarly, Gogate et al. [11] presented a speaker independent AV DNN for IBM estimation to separate speech from background noises. However, the model is trained and evaluated using a limited vocabulary Grid corpus [22] and can help in achieving superior performance. In addition, Hou et al. [13] proposed a speaker-dependent based SE model, trained and evaluated on a single speaker, that predicts the enhanced spectrogram from the noisy spectrogram using multimodal deep convolutional network. However, the model was trained and evaluated on a single speaker corpus. On the other hand, Gabbay et al. [12] trained a convolutional encoder-decoder architecture to

estimate the spectrogram of the enhanced speech from noisy speech spectrogram and cropped mouth regions. However, the model fails to work when the visuals are occluded. Adeel et al. [15, 16] proposed a visual-only and AV SE models by integrating an enhanced visually-derived wiener filter (EVWF) and DNN based lip reading regression model. The preliminary evaluation demonstrated the effectiveness to deal with spectro-temporal variations in any wide variety of noisy environments. Owens et al. [26] proposed a self-supervised trained network to categorise whether audio and visual streams are temporally aligned. The model is then used for feature extraction to condition an on/off screen speaker source separation model. Afouras et al. [27] trained a DNN to predict both magnitude and phase of denoised speech spectrograms. Finally, Zhao et al. [28] presented a model to separate the sound of multiple objects from a video (e.g. musical instruments).

2.2. Audio-only Speech Enhancement

Hershey et al. [29] proposed deep clustering that exploits discriminatively trained speech embeddings to cluster and separate the different sources. For time-domain SE, Rethage et al. [17] proposed a non-causal Wavenet based SE model that operates on raw audios to address the invalid short-time fourier transform (STFT) problem [30] in spectral mask based models. Similarly, Pandey et al. [18] and Luo et al. [19] proposed a fully-convolutional time-domain SE model that address the shortcomings of separation in the frequency domain, including the decoupling of phase and magnitude, and high latency of calculating the STFT.

A fundamental problem with A-only SE and separation is the label permutation problem [29] i.e. there is no easy way to associate a mixture of audio sources with the corresponding speakers or instruments [31]. In addition, the main limitation with most of the aforementioned A-only and AV SE approaches is that the developed model is either evaluated on high SNRs ($SNR > 0$ dB) or on a fixed SNR. In addition, none of the aforementioned AV approaches have used an AV speech corpus recorded in real noisy settings for evaluation of the

Table 1: Grid Corpus Sentence Structure e.g. place blue in A 9 soon

command	**colour**	**preposition**	**letter**	**digit**	**adverb**
bin	blue	at	A-Z	1–9	again
lay	green	by	minus W	zero	now
place	red	in			please
set	white	with			soon

proposed system with commonly observed multiple competing dynamic noise sources.

3. ASPIRE Corpus

In the literature, extensive research has been carried out to develop A-only real noisy mixtures that often consists of speech signal that is reverberantly mixed with multiple competing noise background sources [20]. However, to the best of our knowledge, no such AV corpus recorded in real noisy settings is available. In this section, we present ASPIRE, a first of its kind, AV speech corpus recorded in real noisy environments (such as cafeteria and restaurant) to support reliable evaluation of AV SE technologies.

3.1. Sentence design

ASPIRE corpus follows the same sentence format as the AV Grid corpus as shown in Table 1. The six words sentence consists of command, colour, preposition, letter, digit and adverb. The letter "w" was excluded because it is the only multi-syllabic letter. Each speaker produced all combinations of colour, letter and digit leading to 1000 utterances per talker in both real noisy settings and acoustically isolated booth. As a result each talker recorded 2000 utterances.

3.2. Speaker population

Three speakers (one male and two female) contributed to the corpus. The speakers age ranged from 23 to 55. All the speakers have spent most of their

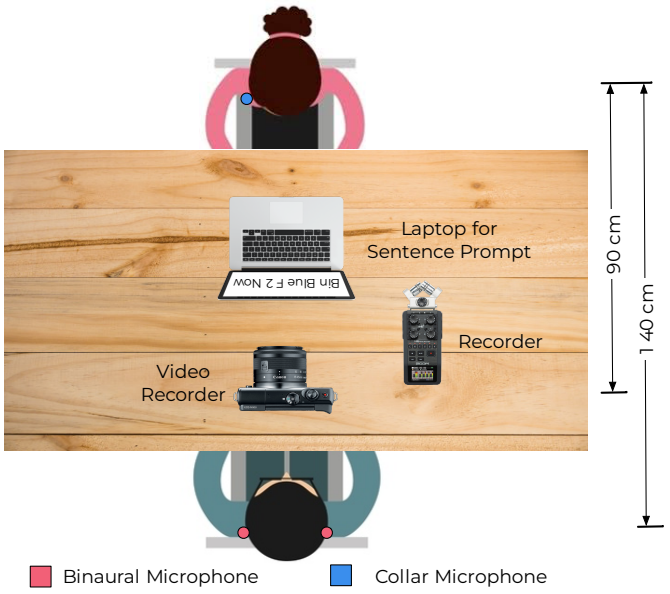


Figure 2: Plan of ASPIRE recording setting showing location of listener, speaker, audio recorder, video recorder, sentence prompter and binaural/collar microphone

lives in the United Kingdom and together encompassed a range of mixed English accents. All the participants were paid for their contribution. In total, the corpus consists of total 6000 utterances (3000 recorded in real noisy settings, 3000 in acoustically isolated booth).

3.3. Collection

The ASPIRE corpus is recorded in real noisy settings specifically the university cafeteria and restaurant during busy lunchtimes (11.30 to 1.30) as well as in an acoustically isolated booth. The recording setup is shown in Figure 2. Apple iPad mini 2, placed at an eye level to avoid noise and distraction from the video apparatus, was used to record the video (the distance between iPad and speaker was 90 centimetres) at 30 frames per second (fps) and 1080p resolution. A collar microphone was also connected to the iPad. The high quality binaural audio from speaker is recorded using Zoom H4n pro recorder at a sampling rate of 44100 Hz and binaural microphone. The listener was wearing the binaural

Figure 3: Sample video frames from ASPIRE corpus

microphone at an approximate distance of 140 centimetres.

The listener and speaker were sitting opposite to each other on the fixed chairs. Speaker was initially trained with few utterances and the purpose of research is also explained in detail. Periodic breaks were given to the speakers during the recording to avoid fatigue and each sentence was mandatory to be read correctly without any interruption. The sentences as detailed in section 3.1 were presented to the speaker on a laptop in random order and speaker was allowed to repeat the sentence if the sentence recording is interrupted or sentence is incorrectly uttered. In addition, the speaker repeated the utterance if any mistake is spotted by the listener. In total, 2000 utterances per speaker (1000 utterances in real noisy settings and 1000 utterances in the booth) around 2% and 4% of the utterances were re-recorded in booth and real noisy settings respectively.

3.4. Postprocessing

Audio postprocessing. Audio and video data were continuously collected throughout a session. The drift between audio and video data was calculated by synchronising the claps. The utterance start and end times were identified using Gentle (a robust forced-aligner built on Kaldi), speech recorded from the collar microphone and the presented transcriptions. Finally, all the segmented utterances were manually checked to correct any additional alignment errors.

Video postprocessing. The raw videos recorded in busy restaurant and cafeteria consists of a few clearly identifiable people except the speaker itself. Therefore, to ensure the privacy, we estimate the speaker area for the first frame using

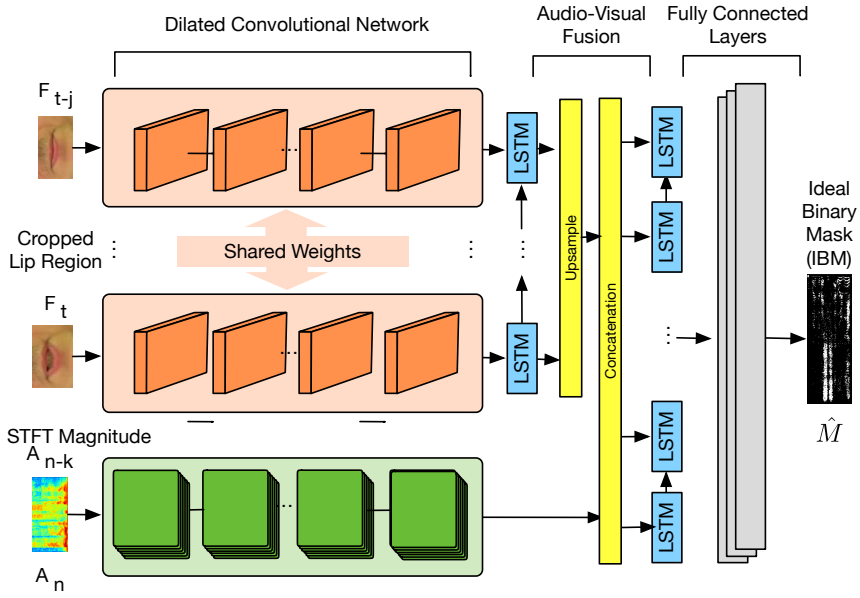


Figure 4: CochleaNet DNN Architecture Overview: Audio-Visual Speech Enhancement

a segmentation model and pixelate the non-speaker area for the complete utterance using the estimated segmentation mask. This is possible because the speaker is sitting in a single position throughout an utterance. Figure 3 shows some sample video frames from the ASPIRE corpus.

4. CochleaNet

This section presents the stages involved in end-to-end processing of the proposed model to output enhanced speech given noisy input speech. Specifically, input feature preprocessing, output feature representation, DNN architecture and speech resynthesis pipeline is described.

4.1. Data Representation

Input features. The DNN ingests both audio and visual as input. For batch training, 3 second video clips are considered. A cropped 80 x 40 lip region is extracted from the video and is used as a visual input (75 cropped lip images

Table 2: Audio Feature Extraction

	conv1	conv2	conv3	conv4	conv5
Num filters	96	96	96	96	96
Filter size	5 x 5	5 x 5	5 x 5	5 x 5	1 x 1
Dilation	1 x 1	2 x 1	4 x 1	8 x 1	1 x 1

for 3 second clip recorded at 25 fps). For audio input, we compute STFT of audio segments and a magnitude spectrogram is used. The trained model can be applied to both streaming data as well as data of arbitrary lengths during inference time.

Output. The output of our network is an IBM, a multiplicative spectrogram mask, that describes the T-F relationship between clean audio and background noise. The IBM assigns zero to a T-F unit if the local SNR is lower than the local criterion (LC), and unit value otherwise. IBM is defined as follows:

$$IBM(t,f) = \begin{cases} 0 & \text{if } SNR(t,f) \leq \text{LC} \\ 1 & otherwise. \end{cases} \tag{1}$$

The IBM has shown to improve the speech quality and intelligibility for the hearing impaired and normal hearing listeners [32, 33, 34]. The IBM cannot be calculated using equation 1 in real-world scenarios because the target speech and interfering background noise cannot be estimated with high accuracy. However, IBM estimation can be modelled as a data-driven optimisation problem that jointly exploits noisy speech and visual face images for the spectral mask estimation. In literature, it has been shown that the multiplicative masks perform better than direct prediction of time-domain waveform and clean spectrogram magnitudes [35, 36].

Table 3: Visual Feature Extraction

	conv1	conv2	maxpool1	conv3	conv4	maxpool2	lstm1
Num filters	32	48		64	96		
Size	3 x 3	3 x 3	2 x 3	3 x 3	3 x 3	2 x 3	256
Dilation	1 x 1	1 x 1		2 x 2	3 x 3		

4.2. Network Architecture

This section describes the network architecture of the proposed AV SE model. Figure 4 depicts a high-level overview of the multi-stream modules present in the network. The subsequent subsections describes each module in detail.

4.2.1. Audio Feature Extraction

The audio feature extraction consist of dilated convolutional layers as detailed in Table 2. Each layer is followed by a ReLU activation for non-linearity.

The dilated convolutions have shown to aggregate the multi-scale contextual information required for dense prediction problem (i.e. assigning 1 or 0 to the individual bin) without losing resolution [37]. In addition, dilated convolutions allow exponential expansion of receptive fields without loss of coverage or resolution [37].

4.2.2. Visual Feature Extraction

The visual feature extraction consist of dilated convolutional, max pooling and long short-term memory (LSTM) layer as detailed in Table 3. Each convolutional layer is followed by a ReLU activation for non-linearity.

As discussed in the aforementioned section, the dilated convolutions are also used to aggregate the multi-scale contextual information from visual lip images. It is to be noted that, the convolution weights are shared across each visual frame. Finally, a LSTM layer is used to exploit the temporal correlation between the extracted visual features using the dilated convolutional network.

4.2.3. Multimodal Fusion

The visual features are sampled at 25 fps while the audio feature sampling rate is 75 vectors per second (VPS). Visual features were upsampled to match the audio vector per second rate and to compensate for the sampling rate discrepancies. This is done using simple repetition of each element 3 times in the temporal dimension. After upsampling, the audio and visual features are concatenated across time dimension and are fed to a LSTM layer consisting of 622 units. The LSTM output is then fed to two fully connected (FC) layers with 622 neurons and ReLU activation. The weights of the FC layers are shared across the time dimension. Finally, the extracted features were fed to a FC layer with 622 neurons and sigmoid activation.

The LSTM layer exploits the joint temporal correlation between the concatenated visual and audio cues to learn the long-term temporal dependencies between the multimodal input. Finally, the FC layers map the integrated AV features to the output IBM. The binary cross-entropy between the estimated and the actual IBM is used as a loss function. It is to be noted that, no thresholding was applied to the predicted mask and the sigmoidal outputs were considered as the estimated mask.

4.3. Mathematical Representation

The framework, shown in Fig. 1, ingests noisy speech (X) and video (V) to output and enhanced speech ($\hat{X}$). Let $A_n, A_{n-1}, ..., A_1$ be the noisy STFT features obtained from x, $\hat{A}_n, \hat{A}_{n-1}, ..., \hat{A}_1$ be the enhanced STFT features,$F_t, F_{t-1}, ..., F_1$ be the cropped images of speakers lips extracted from v of time instance $t_n, t_{n-1}, ..., t_1$ where t is the current time instance and n the is current window frame. Let M be the IBM and $\hat{M}$ be the estimated IBM. The framework can be represented as follows:

$$\hat{X} = f(X) \tag{2}$$

$$\hat{A}_n = \hat{M} \bigodot A_n \tag{3}$$

where $\bigodot$ represents the element wise multiplication.

The DNN, shown in Fig. 4, ingest the noisy STFT features $(A_1, A_2, ..., A_n)$ and cropped images of speakers lips $(F_1, F_2, ..., F_t)$ as an input to output a multiplicative T-F mask $(\hat{M}_n)$ for the current time instance. The DNN can be represented as follows:

$$\hat{M}_n = g(A_1, A_2, ..., A_n, F_1, F_2, ..., F_t) \tag{4}$$

i.e.

$$\hat{M}_1 = g(A_1, F_1), \hat{M}_2 = g(A_1, A_2, F_1, F_2), \tag{5}$$

It can be seen that, the model can be used for causal or real-time mask estimation as the predicted mask for the current time instance (t_n) depends only on the past $(t_{n-1}, t_{n-2}, ..., t_{n-j})$ and current inputs but not future inputs $(t_{n+1}, t_{n+2}, ..., t_{n+k})$.

The network is trained to minimise the binary cross entropy between the M_n and $\hat{M}_n$. The loss function can be represented as follows:

$$Loss(\hat{M}_t, M_t) = -\frac{1}{N}\sum_{i=1}^{N} M_t \cdot log(\hat{M}_t) + (1 - M_t) \cdot log(1 - \hat{M}_t) \tag{6}$$

The mathematical formulation for the dilated convolutional network, and LSTM is detailed in [38] and [39] respectively·

4.4. Post-processing: Speech Resynthesis

The model estimates a T-F IBM when a noisy spectrogram and cropped lip images are fed. The estimated multiplicative spectral mask is applied to the noisy magnitude spectrum. The masked magnitude is then combined with the noisy phase to get the enhanced speech using ISTFT. Figure 1 depicts an overview of speech resynthesis.

5. Experiments and Results

We qualitatively and quantitatively evaluated our proposed approach with other state-of-the-art A-only and AV SE in real noisy environments and a range of synthetic AV corpora.

5.1. Synthetic AV Corpora

This section presents the synthetic AV corpora used for training and testing of CochleaNet.

5.1.1. Grid + ChiMe 3

In our experiments, benchmark Grid corpus [22] is used for the training and evaluation of the proposed framework. All 33 speakers with 1000 utterances each are considered. The sentence format is depicted in Table 1. The Grid corpus is randomly mixed with non-stationary noises from 3rd CHiME challenge (CHiME 3)[20], consisting of bus, cafeteria, street, and pedestrian noises, for SNRs ranging [-12, 9] dB with a step size of 3 dB. It is to be noted that, the trained model is SNR-independent i.e. the utterances at all SNRs were combined for training and evaluation. For training, 21000 utterances from 21 speakers were employed. The model was validated and tested on 4000 and 8000 utterances from 4 and 8 speakers respectively. It is to be noted that, the experiments are repeated 10 times for different set of train and validation speakers for statistically significant comparison with other state-of-the-art methods.

5.1.2. TCD-TIMIT + MUSAN

For large vocabulary generalisation analysis, we used benchmark TCD-TIMIT [23] corpus. Specifically, 5488 utterances from 56 speakers are mixed with randomly selected non-speech noises from MUSAN noises [25]. The MUSAN noises include technical noises (e.g. dialtones, fax machine noises etc.) as well as ambient sounds (e.g. thunder, wind, footsteps, animal noises etc.). It to be noted that, all the 5488 utterances were used as a test set to asses the model performance on large vocabulary, speaker and noise independent settings.

5.1.3. Mandarin AV Corpus + NOISEX-92

For language-independent generalisation testing, a Mandarin dataset [13] based on Taiwan Mandarin Hearing in Noise Test (MHINT) with 320 utterances is mixed with randomly selected noise from NOISEX-92 [24] consisting of voice

babble, factory radio channel and various military noises including fighter jets, engine room, operations room, tank and machine gun.

5.2. Data Preprocessing

5.2.1. Audio Preprocessing

The audio signals were resampled at 16 kHz and a mono channel is used for processing. The resampled audio signal was segmented into N 78 millisecond (ms) frames and 17% increment rate to produce 75 fps. A hanning window and STFT is applied to produce 622-bin magnitude spectrogram.

5.2.2. Video Preprocessing

The Grid and TCD-TIMIT corpora are recorded at 25 fps. However, the Mandarin dataset [13], recorded at 30 fps, is downsampled to 25 fps using ffmpeg [40]. A dlib face detector [41] is used to locate the faces in each frame of a video clip (75 face cropped images assuming 3 second clip recorded at 25 fps). The speakers lip images are extracted out of the 25 fps faces video using a minified dlib [41] model optimised for extracting the lip landmarks. A region of aspect ratio 1:2 centred at lip-centre is extracted using the lip landmark points. The extracted region is resized to 40 pixels x 80 pixels and converted to a greyscaled image. It is to be noted that, the lip sequences are extracted at 25 fps and audio features are extracted at 75 VPS.

5.3. Experimental Setup

For the AV features fusion and mask estimation, the network is trained using TensorFlow library and NVIDIA Titan Xp GPUs. A subset of speakers from Grid ChiME 3 corpus (as described in section 5.1) are used for training/validation of the neural network and rest of the speakers are used to test the performance of the trained neural network in speaker independent scenario (25% testing dataset). The preprocessed training set of Grid ChiME 3 corpus consists of around 25000 utterances, that are split into 21000 and 4000 utterances for training and validation respectively. It is to be noted that, there was

no overlap between the speakers and the noises present in the train, validation and test set for ensuring the speaker and noise independent criteria. When a missing visual frame is encountered a vector of zeros is used in lieu of the lip image. The preprocessed dataset consists of cropped lip images and noisy audio spectrogram as input and IBM as an output. The network is trained for 50 epochs using backpropagation with Adam optimiser [42] with learning rate 0.0003. The learning rate is divided by 2 when the validation error stops reducing for 3 consecutive epochs. Finally, early stopping is used if the validation error stops decreasing for 6 consecutive epochs.

5.4. Objective testing on Synthetic mixtures

In the past, the quality of the speech processing system can only be evaluated by conducting subjective listening tests (i.e. by asking listeners to compare between different speech systems) or by conducting intelligibility test (where a listener writes down the intelligible words and metrics such as word error rate are used). However, as the size of data increases, conducting a subjective listening test take more time and the test results may not represent the actual distribution present in the data. Therefore, researchers have proposed methods such as PESQ [43], STOI [44], and SI-SDR [45] to computationally approximate the subjective listening tests. In this section, the proposed model is compared with the state-of-the-art using the following objective metrics.

5.4.1. Perceptual Evaluation of Speech quality (PESQ) comparison

PESQ [43] is one of the most commonly used objective assessment metric to predict the subjective listening test scores in the SE literature and has shown to correlate well with the subjective listening tests [46]. PESQ is computed as a linear combination of the average disturbance value and the average asymmetrical disturbance values between a reference signal and modified signal. However, PESQ only measures the effect of one-way speech distortion and noise speech quality, and the effect related to two-way interaction including loudness, loss, delay, sidetone, and echo are not reflected in the PESQ score.

PESQ score ranges from $[-0.50, 4.50]$, indicating the minimum and maximum possible reconstructed speech quality. The PESQ scores for A-only and AV CochleaNet, SEGAN, SS, and LMMSE with Grid + ChiME 3, TCD TIMIT + MUSAN and Hou et al [13] + NOISEX-92 for different SNRs are presented in Table 4, 5, 6 respectively. The variety of datasets ensure speaker and noise independent criteria, large vocabulary corpus as well as language-independent scenario. It is to be noted that, the model trained on Grid + ChiME 3 corpus is used for evaluation. It can be seen that, at low SNRs, AV CochleaNet and A-only CochleaNet outperformed SS [47], LMMSE [48], and SEGAN [49] based SE methods. In addition, AV perform better than A-only CochleaNet especially for low SNR ranges (i.e. $SNR < 0$ dB), where AV CochleaNet model achieved the 1.98, 2.18, and 2.33 PESQ score at SNR levels, of -12dB, -9dB, and -6 dB respectively, as compared to 1.85, 2.05, and 2.24 PESQ score achieved by A-only CochleaNet model for Grid ChiME 3 speaker independent test set. However, at high SNRs (i.e. $SNR >= 0$ dB) AV slightly outperformed A-only mask estimation model, where AV CochleaNet achieved 2.58, 2.69, and 2.78 PESQ score at SNR levels, of 0 dB, 3 dB, and 6 dB respectively, as compared to 2.52, 2.63, and 2.73 achieved by A-only CochleaNet model for Grid ChiME 3 speaker independent test set. The overall PESQ improvement as compared to noisy audio is depicted in Figure 5, where AV CochleaNet outperformed the A-only CochleaNet, and achieved near optimal performance (close to an ideal IBM) for Grid ChiME 3 corpus.

5.4.2. Short Term Objective Intelligibility (STOI) comparison

STOI is a benchmark objective evaluation metric used for speech intelligibility that shows a high correlation with subjective listening test scores [44]. The correlation of short-time temporal envelopes between the clean and modified speech is calculated in STOI with values ranging from $[0, 1]$, and higher value indicates better intelligibility. STOI decomposes signals into T-F regions followed by energy clipping and normalization. The intelligibility predicts are based on cross-correlations between processed and signal across different T-

Table 4: PESQ scores for Grid ChiME 3 speaker independent test set computed from the resynthesised speech using SEGAN+ [49], SS [47], LMMSE [48], Audio-only (A) CochleaNet, Audio-Visual (AV) CochleaNet, and Oracle IBM. The reference PESQ for the unprocessed (Noisy) signal is included for relative comparison. It is to be noted that, the mean and variance is calculated using the model trained on 10 different shuffled splits of the GRID CHiME3 corpus.

dB	-12	-9	-6	-3
Noisy	1.30 ± 0.015	1.41 ± 0.014	1.54 ± 0.013	1.70 ± 0.016
SEGAN+	0.88 ± 0.012	1.05 ± 0.011	1.45 ± 0.011	1.80 ± 0.012
SS	1.17 ± 0.011	1.23 ± 0.013	1.40 ± 0.014	1.60 ± 0.012
LMMSE	1.38 ± 0.014	1.53 ± 0.014	1.73 ± 0.012	1.96 ± 0.015
Proposed A	1.85 ± 0.011	2.05 ± 0.010	2.24 ± 0.011	2.39 ± 0.012
Proposed AV	1.98 ± 0.012	2.18 ± 0.009	2.33 ± 0.012	2.46 ± 0.010
Oracle IBM	2.05 ± 0.007	2.22 ± 0.007	2.33 ± 0.009	2.47 ± 0.008
dB	0	3	6	9
Noisy	1.87 ± 0.014	2.07 ± 0.012	2.27 ± 0.011	2.45 ± 0.010
SEGAN+	2.12 ± 0.011	2.37 ± 0.010	2.58 ± 0.010	2.76 ± 0.011
SS	1.82 ± 0.012	2.08 ± 0.013	2.34 ± 0.011	2.58 ± 0.010
LMMSE	2.17 ± 0.013	2.39 ± 0.011	2.58 ± 0.012	2.75 ± 0.011
Proposed A	2.52 ± 0.010	2.63 ± 0.011	2.73 ± 0.010	2.81 ± 0.009
Proposed AV	2.58 ± 0.009	2.69 ± 0.009	2.78 ± 0.008	2.85 ± 0.008
Oracle IBM	2.58 ± 0.007	2.70 ± 0.006	2.82 ± 0.006	2.90 ± 0.006

F cells. The STOI scores for A-only and AV CochleaNet, SEGAN, SS, and LMMSE with Grid + ChiME 3, TCD TIMIT + MUSAN and Hou et al [13] + NOISEX-92 for different SNRs are presented in Fig 6. It can be seen that, at low SNRs, AV CochleaNet and A-only CochleaNet outperformed SS [47], LMMSE [48], SEGAN [49] based SE methods. In addition, AV performs better than A-only model especially for low SNR ranges (i.e. $SNR < 0$ dB), where AV CochleaNet model achieved the STOI scores of 0.521, 0.560, and 0.607 at SNR levels, of -12dB, -9dB, and -6 dB respectively, as compared to 0.483, 0.513, and

Table 5: PESQ scores ($\mu \pm \sigma$) for large vocabulary TCD-TIMIT + MUSAN AV dataset computed from the resynthesised speech using SEGAN+ [49], SS [47], LMMSE [48], Audio-only (A) CochleaNet, Audio-Visual (AV) CochleaNet, and Oracle IBM. The reference PESQ for the unprocessed (Noisy) signal is included for relative comparison. It is to be noted that, the mean and variance is calculated using the model trained on 10 different shuffled splits of the GRID CHiME3 corpus.

dB	-12	-9	-6	-3
Noisy	1.46 ± 0.015	1.56 ± 0.014	1.64 ± 0.013	1.66 ± 0.015
SEGAN+	1.05 ± 0.012	1.13 ± 0.014	1.16 ± 0.013	1.25 ± 0.012
SS	1.43 ± 0.013	1.44 ± 0.013	1.61 ± 0.014	1.64 ± 0.015
LMMSE	1.61 ± 0.012	1.73 ± 0.013	1.75 ± 0.013	1.83 ± 0.014
Proposed A	1.81 ± 0.011	1.91 ± 0.012	2.04 ± 0.012	2.14 ± 0.013
Proposed AV	1.88 ± 0.012	1.98 ± 0.011	2.11 ± 0.010	2.15 ± 0.011
Oracle IBM	2.55 ± 0.008	2.56 ± 0.007	2.68 ± 0.008	2.73 ± 0.009
dB	0	3	6	9
Noisy	2.23 ± 0.012	2.34 ± 0.015	2.44 ± 0.013	2.51 ± 0.011
SEGAN+	1.76 ± 0.011	1.88 ± 0.012	2.05 ± 0.011	2.15 ± 0.012
SS	2.03 ± 0.012	2.14 ± 0.012	2.25 ± 0.011	2.33 ± 0.012
LMMSE	2.34 ± 0.013	2.43 ± 0.013	2.55 ± 0.012	2.65 ± 0.011
Proposed A	2.35 ± 0.010	2.46 ± 0.009	2.51 ± 0.010	2.56 ± 0.010
Proposed AV	2.45 ± 0.011	2.55 ± 0.010	2.64 ± 0.011	2.65 ± 0.009
Oracle IBM	2.81 ± 0.007	2.84 ± 0.008	2.86 ± 0.007	2.94 ± 0.006

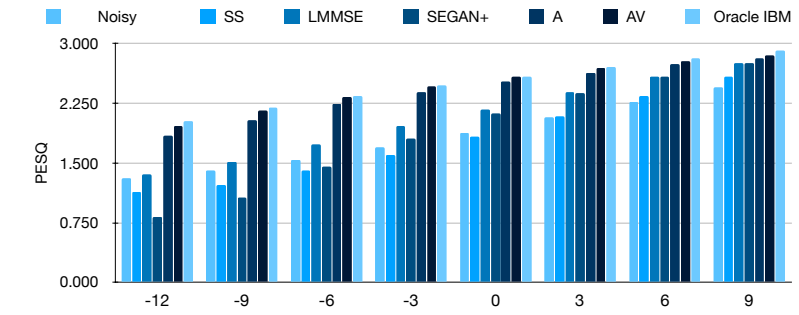
0.544 achieved by A-only CochleaNet model for Hou et al [13] + NOISEX-92 language-independent test set. However, at high SNRs (i.e. $SNR >= 0$ dB) AV slightly outperformed A-only mask estimation model, where AV CochleaNet achieved STOI scores of 0.719, 0.739, and 0.776 at SNR levels, of 0 dB, 3 dB, and 6 dB respectively, as compared to 0.665, 0.701, and 0.752 achieved by A-only CochleaNet model for Hou et al [13] + NOISEX-92 language-independent test set.

Table 6: PESQ scores ($\mu \pm \sigma$) for Hou et al. [13] + NOISEX92 AV language-independent dataset computed from the resynthesised speech using SEGAN+ [49], SS [47], LMMSE [48], Audio-only (A) CochleaNet, Audio-Visual (AV) CochleaNet, and Oracle IBM. The reference PESQ for the unprocessed (Noisy) signal is included for relative comparison. It is to be noted that, the mean and variance is calculated using the model trained on 10 different shuffled splits of the GRID CHiME3 corpus.

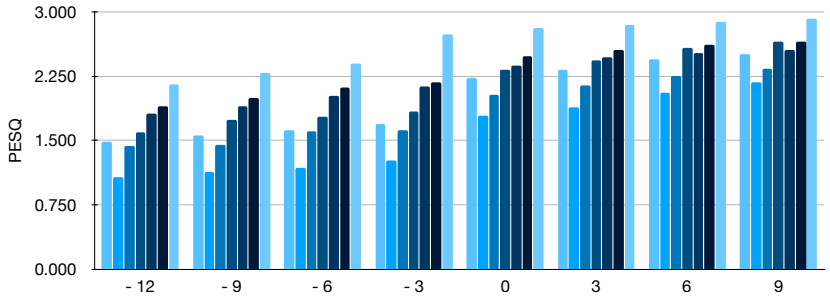
dB	-12	-9	-6	-3
Noisy	1.04 ± 0.016	1.25 ± 0.018	1.28 ± 0.017	1.31 ± 0.014
SEGAN+	0.63 ± 0.014	1.06 ± 0.015	1.36 ± 0.014	1.34 ± 0.016
SS	1.21 ± 0.015	1.44 ± 0.014	1.38 ± 0.014	1.41 ± 0.013
LMMSE	1.14 ± 0.014	1.31 ± 0.014	1.16 ± 0.014	1.45 ± 0.013
Proposed A	1.26 ± 0.014	1.44 ± 0.012	1.56 ± 0.011	1.53 ± 0.012
Proposed AV	1.34 ± 0.012	1.53 ± 0.013	1.54 ± 0.012	1.56 ± 0.011
Oracle IBM	1.55 ± 0.011	1.68 ± 0.0121	1.75 ± 0.011	1.71 ± 0.010
dB	0	3	6	9
Noisy	1.41 ± 0.014	1.48 ± 0.013	1.71 ± 0.014	1.64 ± 0.012
SEGAN+	1.26 ± 0.013	1.23 ± 0.012	1.36 ± 0.013	1.34 ± 0.012
SS	1.41 ± 0.014	1.44 ± 0.010	1.61 ± 0.014	1.44 ± 0.011
LMMSE	1.58 ± 0.012	1.66 ± 0.012	1.71 ± 0.012	1.74 ± 0.012
Proposed A	1.66 ± 0.011	1.74 ± 0.009	1.78 ± 0.008	1.74 ± 0.008
Proposed AV	1.71 ± 0.012	1.74 ± 0.010	1.75 ± 0.009	1.76 ± 0.009
Oracle IBM	1.83 ± 0.010	1.86 ± 0.010	1.94 ± 0.009	1.85 ± 0.008

5.4.3. Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) comparison

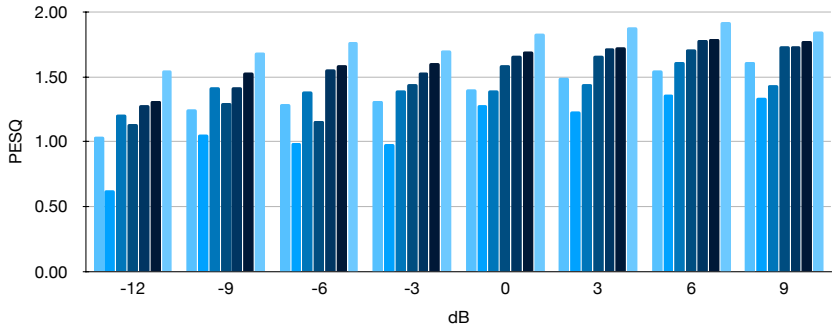
SI-SDR [45] is slightly modified scale invariant version of SDR. SDR is one of the standard speech separation evaluation metrics that measure the amount of distortion introduced by the separated signal and is defined as the ratio between clean signal energy and distortion energy. The higher SDR values indicate better speech separation performance. The SI-SDR scores for A-only and AV CochleaNet, SEGAN, SS, and LMMSE with Grid + ChiME 3, TCD TIMIT + MUSAN and Hou et al [13] + NOISEX-92 for different SNRs are presented



(a) Grid + ChiME 3



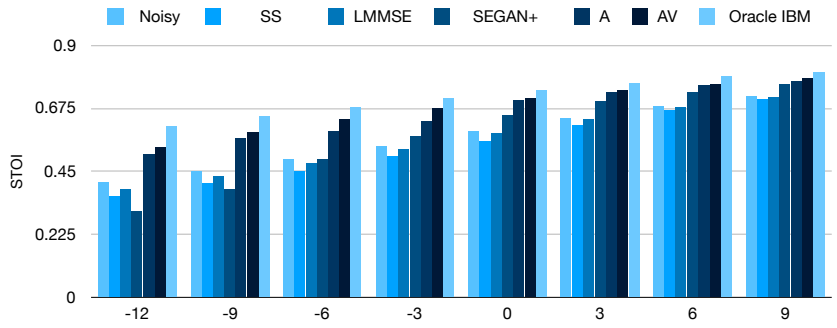
(b) TCD + MUSAN



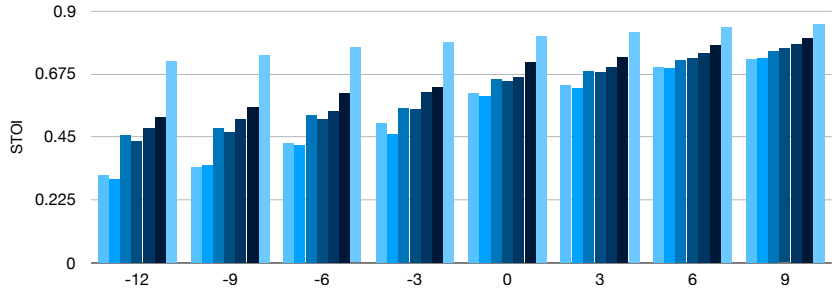
(c) Hou et al. + NOISEX-92

Figure 5: PESQ scores for (a) Grid + ChiME3 (b) TCD + MUSAN (c) Hou et al. [13] + NOISEx-92 AV dataset computed from the resynthesised speech using SEGAN+ [49], SS [47], LMMSE [48], Audio-only (A) CochleaNet, Audio-Visual (AV) CochleaNet, and Oracle IBM. The reference PESQ for the unprocessed (Noisy) signal is included for relative comparison.

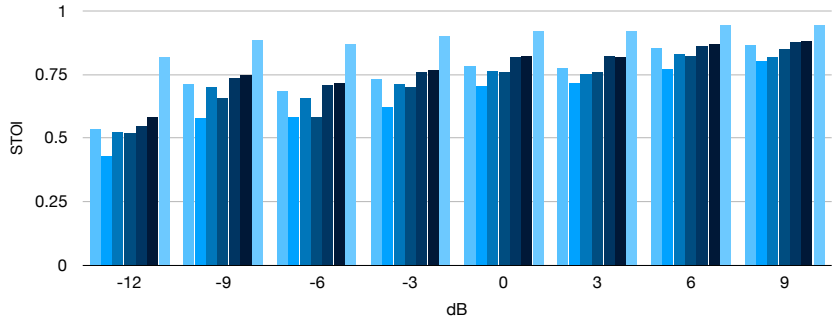
in Fig 7 respectively. It can be seen that, at low SNRs, AV CochleaNet and A-only CochleaNet outperformed SS [47], LMMSE [48], SEGAN [49] based SE methods. In addition, AV performs better than A-only mask estimation model



(a) Grid + ChiME 3



(b) TCD + MUSAN



(c) Hou et al. + NOISEX-92

Figure 6: STOI scores for (a) Grid + ChiME3 (b) TCD + MUSAN (c) Hou et al. [13] + NOISEX-92 AV dataset computed from the resynthesised speech using SEGAN+ [49], SS [47], LMMSE [48], Audio-only (A) CochleaNet, Audio-Visual (AV) CochleaNet, and Oracle IBM. The reference STOI for the unprocessed (Noisy) signal is included for relative comparison.

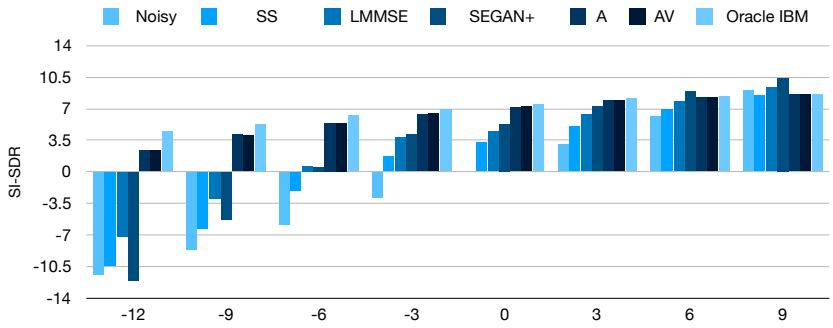
especially for low SNR ranges (i.e. $SNR < 0$ dB), where AV CochleaNet model achieved the SI-SDR scores of 3.62, 4.80, and 5.41 at SNR levels, of -12dB, -9dB,

and -6 dB respectively, as compared to 3.04, 4.41, and 5.29 achieved by A-only CochleaNet model for TCD-TIMIT + MUSAN speaker independent and large vocabulary test set. However, at high SNRs (i.e. $SNR >= 0$ dB) AV slightly outperformed A-only mask estimation model, where AV CochleaNet achieved SI-SDR scores of 7.77, 8.64, and 9.31 at SNR levels, of 0 dB, 3 dB and 6 dB respectively, as compared to 7.76, 8.62, and 9.27 achieved by A-only CochleaNet model for TCD-TIMIT + MUSAN speaker independent and large vocabulary test set.

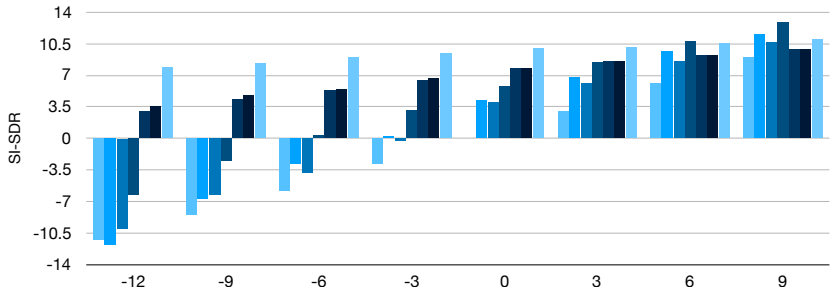
Figure 9 presents the noisy, clean spectrogram and spectrograms for the reconstructed speech signal of a random utterance from GRID + ChiME 3 AV corpus using SS, LMMSE, SEGAN+, A-only CochleaNet, AV CochleaNet and Oracle IBM. It is to be noted that, the speech is completely swamped with background noise and the performance of CochleaNet models can be seen (i.e. close to the Oracle IBM).

5.5. Subjective testing on ASPIRE Corpus

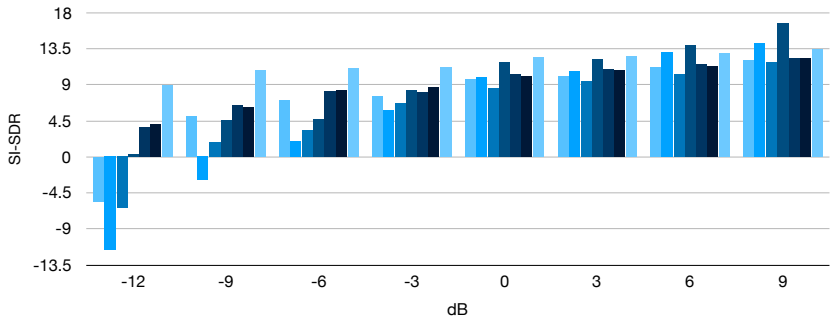
In the literature, the significant number of objective metrics [43, 44, 45] have been proposed to computationally approximate the subjective listening tests. However, the only way to quantify the subjective quality is to ask listeners for their opinions. We used MUSHRA-style [50] listening test method for subjective evaluation, using enhanced speech from real noisy ASPIRE corpus (section 3). A total of 20 native English speakers with normal-hearing participated in the listening test. The individual test consist of 20 randomly selected utterances drawn from the ASPIRE corpus. The first two screens were used to train participants to adjust the volume and to familiarise with the screen and the task. In each screen, the participants were asked to score the quality of each audio sample, on a scale from $[0, 100]$, generated by each SE model for the same sentence. The range from $[80, 100]$ is described as "excellent", from $[60, 80]$ as "good", from $[40, 60]$ as "fair", from $[20, 40]$ as "poor", and from $[0, 20]$ as "bad". Noisy speech was included in the test therefore that participants would have a reference for the degraded speech as well as for checking if participants



(a) Grid + ChiME 3



(b) TCD + MUSAN



(c) Hou et al. + NOISEX-92

Figure 7: SI-SDR scores for (a) Grid + ChiME3 (b) TCD + MUSAN (c) Hou et al. [13] + NOISEX-92 AV dataset computed from the resynthesised speech using SEGAN+ [49], SS [47], LMMSE [48], Audio-only (A) CochleaNet, Audio-Visual (AV) CochleaNet, and Oracle IBM. The reference SI-SDR for the unprocessed (Noisy) signal is included for relative comparison.

go through the material.

The times required to complete each screen were also recorded and used

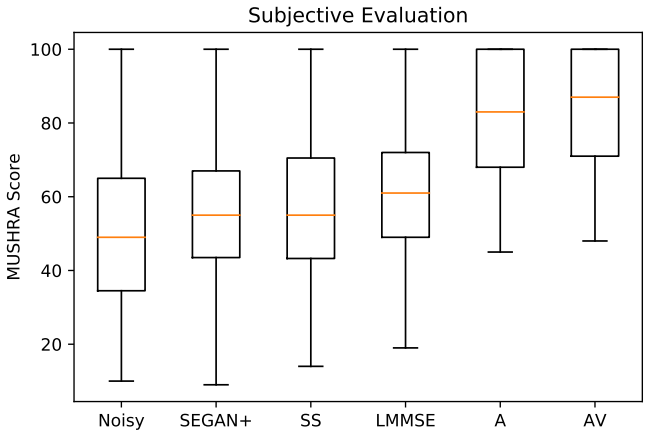


Figure 8: Result of MUSHRA listening test for ASPIRE corpus for the reconstructed speech signal using SS [47], LMMSE [48], SEGAN+ [49], A-only CochleaNet, AV CochleaNet. The reference MUSHRA score for the unprocessed (Noisy) signal is included for relative comparison.

for removing any outliers. We evaluated five SE models including SEAGN, SS, LMMSE, A-only CochleaNet and AV CochleaNet. Figure 8 shows the boxplot of listeners responses in terms of the rank order of systems for the ASPIRE corpus. The listening test results show that the superior performance of our AV CochleaNet, over A-only CochleaNet, SEGAN, spectral subtraction (SS), and log-minimum mean square error (LMMSE) based SE methods. The results demonstrate the capability of CochleaNet to deal with the reverberation caused by multiple competing background sources observed in a real-world noisy environment, by exploiting the audio and visual cues. In addition, the results show that an AV model trained on synthetic additive mixtures generalise well real noisy corpus.

5.6. Processing Latency

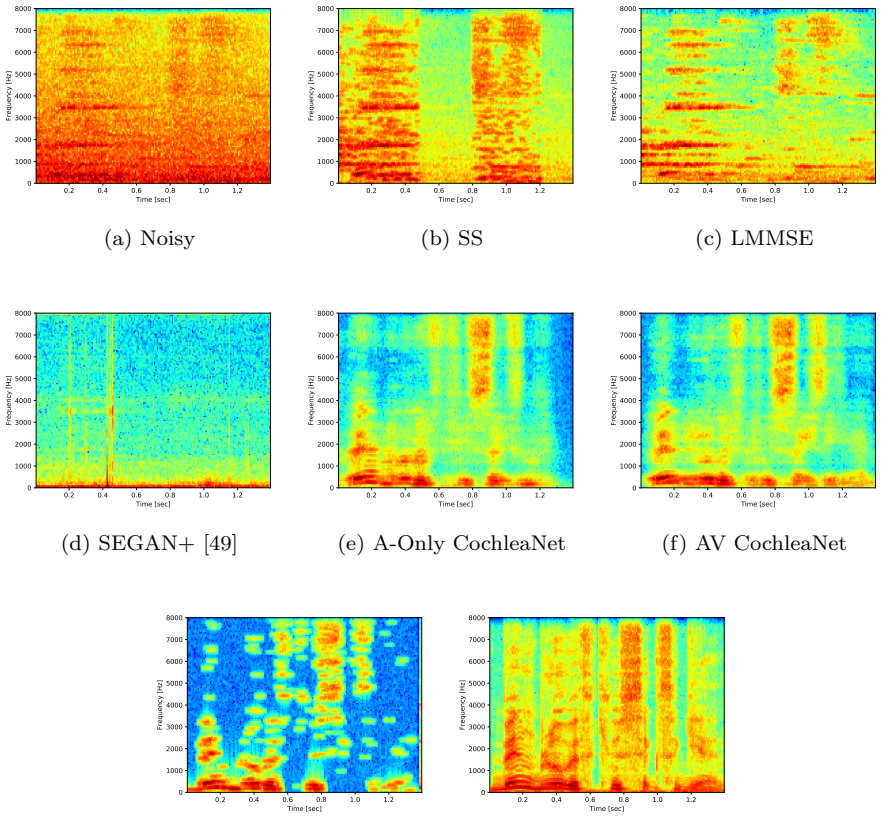
The processing latency for a listening device (generally measures in milliseconds) such as hearing aids is defined as the difference between the time of the original speech and the time when the enhanced speech leaves the device. If

the processing latency is more than 10ms, the current speech and the enhanced speech will result in an echo effect. In addition, such delay results in incorrect synchronisation of the auditory information with visual information and interfere in speech understanding.

The processing latency of the proposed model (25ms) is dependent upon the window shift of the Fourier transform (13ms), STFT latency (1ms), visual preprocessing latency for cropping lip images (1ms), CochleaNet model prediction time per window size (9ms), and ISTFT latency (1ms). The above values are calculated with 3.4 GHz Intel i7 processor and 16 GB RAM. It can be seen that, the main bottleneck for the deployment of the proposed model in real-time application such as hearing aids is the window size and window shift used for Fourier transform as well as processing delay of the model itself. The window size and model complexity can be further optimised to use the model in listening devices. In addition, the Fourier transform delays can be removed if the model directly ingests the time domain noisy speech signal. However, the model can be used without modification in applications such as telephone/video conferencing, noise-robust speech recognition systems etc.

5.7. Additional Analysis

Effect of occluded visual information. The model is trained and evaluated on a professionally recorded corpus that ensured none of the visual frames consists of occluded lip images (except a small number of Grid corpus utterances where visuals are absent). However, in real life scenarios specifically, when the source and the target is non-stationary the model needs to be robust against the missing visual information. Therefore, to experimentally evaluate the trained AV CochleaNet behaviour in such conditions we randomly replaced a percentage of lip images with a blank visual frame. The results for lip occlusion is depicted in Figure 10. It can be seen that, for both -9 dB and -12 dB, as the visual occlusion increases the PESQ score initially remains constant and after 20% occlusion linearly starts decreasing. It is worth mentioning that, AV model performs similar to the A-only model when visuals are completely absent even though the model



(a) Noisy (b) SS (c) LMMSE

(d) SEGAN+ [49] (e) A-Only CochleaNet (f) AV CochleaNet

(g) Oracle IBM (h) Clean

Figure 9: Spectrogram of a randomly enhanced -6 dB utterance from GRID + ChiME3 Speaker independent test set. It can be seen that A-only, and AV CochleaNet outperformed SS, LMMSE and SEGAN based enhancement. It is to be noted that, AV CochleaNet recovered some frequency components better than A-only CochleaNet.

has not encountered such situation during training.

Phoneme level comparison of audio-only and audio-visual CochleaNet. It is well known in the literature that, visual information help disambiguate the phonological ambiguity. In addition, some phonemes such as /p/ are visually distinguishable and phonemes such as /g/ cannot be visually distinguished. However, the relationship between the visually distinguishable phonemes and the AV SE

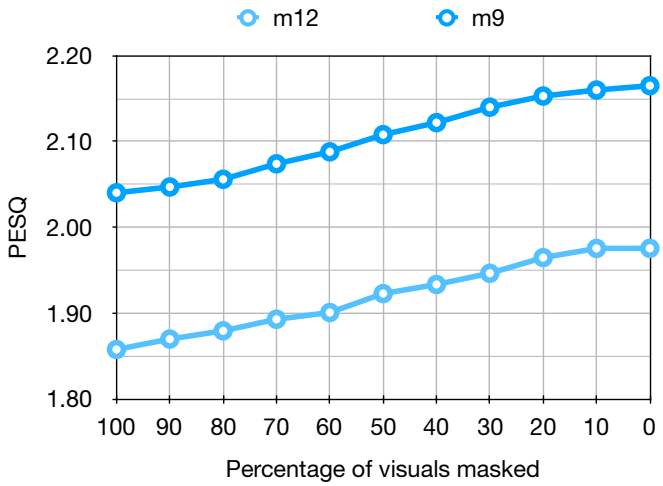


Figure 10: PESQ scores for different percentage of masked lip images

performance is not known. Therefore, we conducted comparative listening tests with 3 listeners and 1000 random enhanced utterances from Grid CHiME 3 speaker independent test set to empirically identify if there is a relation between the visually distinguishable phonemes and the phonemes that AV CochleaNet can enhance better than A-only CochleaNet. The listening tests revel that AV model enhanced the /r/, /p/, /l/, /w/, /EH1/, /AE1/, /IY1/, /EY1/, /AA1/ and /OW1/ phonemes better than A-only model and the AV performance on phoneme such as /h/, /g/ and /k/ was similar to A-only performance. This confirmed the hypothesis that there is a direct relation between visually distinguishable phonemes and the phonemes that AV model works better on.

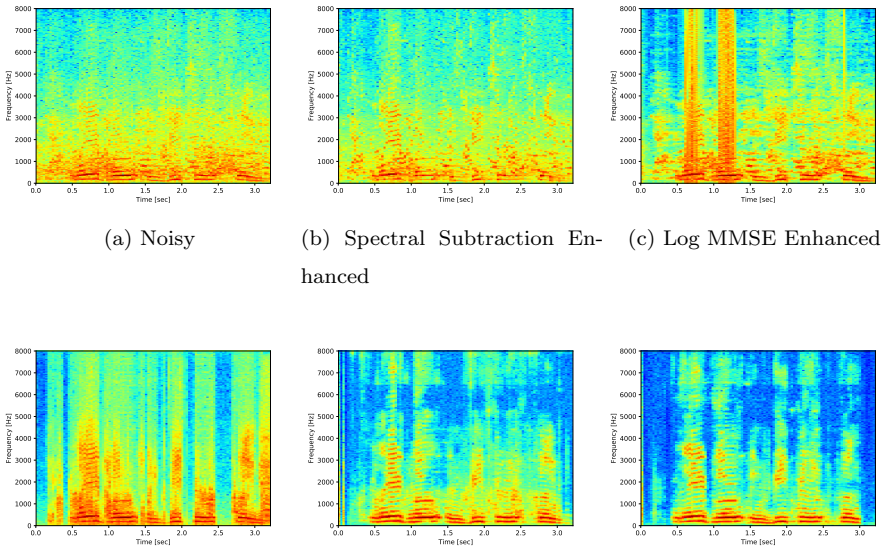
Comparison of audio-only and audio-visual CochleaNet in silent speech regions. The superior performance of AV CochleaNet as compared to A-only CochleaNet could be because of the visual cues, specifically, the closed lip, could give extra information to AV model in silent speech regions. In ordered to verify this hypothesis, we calculated the mean squared error (MSE) between the predicted masks and the IBM in the silent speech regions. The A-only model achieved MSE of 0.0123 as compared to the AV that achieved MSE of 0.0108. This

confirms the aforementioned hypothesis, however further analysis is needed to visualise the convolutional receptive fields and to check if a particular part of the model is active when the speaker is silent. Figure 11 presents the noisy spectrogram and spectrograms for the reconstructed speech signal of a random utterance from TCD-TIMIT corpus using SS, LMMSE, SEGAN+, A-only CochleaNet, AV CochleaNet. It can be seen that, the speech is completely swamped with background noise and the A-only and AV CochleaNet managed to suppress the noise dominant regions and speech dominant regions as compared to SS, LMMSE and SEGAN+. It can be seen that, in silent speech regions, AV CochleaNet outperformed A-only CochleaNet.

The main limitation with the proposed work is that: (1) the process of IBM based SE ignore the phase spectrum that leads to invalid STFT problem [18] (2) the model cannot separate the overlapping speech if more than one speaker is speaking simultaneously as the model is not trained with such mixed AV corpora (3) the ASPIRE corpus consists of only three speakers recorded in controlled real noisy environments with stationary speaker-listener setting and more challenging non-stationary real noisy corpora are required to assess the robustness of the model (4) the proposed model works only on a single channel audio and cannot exploit the binaural nature of speech we experience everyday (5) extracting the visual input (i.e. lip image) is still an open challenge for deployment of AV models as imperfections such as occlusion, poor lighting, head movements etc. need to be addressed. However, the aforementioned experiments on the effect of occluded visual information confirm that even if the visuals are absent the performance of the AV model is similar to the A-only model.

6. Conclusion

This paper presented a language, noise and speaker independent AV DNN model for causal SE that contextually exploits the audio and visual cues, independent of the SNR, to estimate the spectral IBM and enhance speech. In addi-



(a) Noisy (b) Spectral Subtraction Enhanced (c) Log MMSE Enhanced

(d) SEGAN+ [49] (e) A-only CochleaNet (f) AV CochleaNet

Figure 11: Spectrogram of a randomly enhanced utterance from ASPIRE corpus. It is to be noted that, AV CochleaNet outperforms A-only CochleaNet, specifically in silent speech regions where visual cues (lip position) help identify if the speaker is talking or not.

tion, we presented a novel AV corpus, ASPIRE[1], consisting of speech recorded in real noisy environments such as cafeteria and restaurant to evaluate the proposed model. The corpus can be used as a resource by speech community to evaluate AV SE models. We perform extensive experiments taking into consideration the noise, speaker and language-independent criteria. The performance evaluation in terms of objective metrics (PESQ, SI-SDR, and ESTOI) and subjective MUSHRA listening tests revealed significant improvement of our proposed AV CochleaNet as compared to the A-only CochleaNet, state-of-the-art SE (including SS, LMMSE) approaches as well as DNN based SE approaches (including SEGAN). The simulation results have validated the phenomena of

[1]ASPIRE Corpus, enhanced speech samples, and additional supplementary material is available on the project website: https://cochleanet.github.io

more effective visual cues at low SNRs, less effective visual cues at high SNRs. The visual occlusion study depicts that the model performance initially remains constant till 20% of the visuals are removed and after 20% occlusion the performance linearly decreases as the number of occluded frame increases. The empirical study to identify the role visual cues play in superior performance of AV model as compared to A-only model show that, there is a high correlation between visually distinguishable phonemes and the AV model performance. Moreover, the study shows that AV model significantly outperforms A-only in silent speech region because it is relatively easier to audio-visually distinguish if a speaker is speaking or not as compared to only using only audio input. In future, we intend to investigate the generalisation capability of our proposed DNN model with other more challenging conversational real noisy AV corpora as well as address issue of imperfect visual information. Ongoing and future work also addresses the real time implementation challenges and privacy concerns with multimodal AV hearing aids.

7. Acknowledgement

This work was supported by the Edinburgh Napier University Research Studentship and UK Engineering and Physical Sciences Research Council (EPSRC) Grant No. EP/M026981/1. The authors would also like to acknowledge Dr Ricard Marxer and Prof Jon Barker from the University of Sheffield. Finally, we would like to acknowledge all the participants and support staff involved in the collection of ASPIRE corpus.

References

References

[1] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (5588) (1976) 746.

[2] Q. Summerfield, Lipreading and audio-visual speech perception, Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 335 (1273) (1992) 71–78.

[3] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, D. Poeppel, Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party, Journal of Neuroscience 33 (4) (2013) 1417–1426.

[4] K. W. Grant, P.-F. Seitz, The use of visible speech cues for improving auditory detection of spoken sentences, The Journal of the Acoustical Society of America 108 (3) (2000) 1197–1208.

[5] K. W. Grant, S. Greenberg, Speech intelligibility derived from asynchronous processing of auditory-visual information, in: AVSP 2001-International Conference on Auditory-Visual Speech Processing, 2001.

[6] A. Narayanan, D. Wang, Investigation of speech separation as a front-end for noise robust speech recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (4) (2014) 826–835.

[7] H. Kayser, C. Spille, D. Marquardt, B. T. Meyer, Improving automatic speech recognition in spatially-aware hearing aids, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[8] D. Wang, G. J. Brown, Fundamentals of computational auditory scene analysis.

[9] J. Chen, D. Wang, Dnn based mask estimation for supervised speech separation, in: Audio source separation, Springer, 2018, pp. 207–235.

[10] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, M. Rubinstein, Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation, ACM Transactions on Graphics (TOG) 37 (4) (2018) 112.

[11] M. Gogate, A. Adeel, R. Marxer, J. Barker, A. Hussain, Dnn driven speaker independent audio-visual mask estimation for speech separation, Proc. Interspeech 2018 (2018) 2723–2727.

[12] A. Gabbay, A. Shamir, S. Peleg, Visual speech enhancement, in: Interspeech, ISCA, 2018, pp. 1170–1174.

[13] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, H.-M. Wang, Audio-visual speech enhancement using multimodal deep convolutional neural networks, IEEE Transactions on Emerging Topics in Computational Intelligence 2 (2) (2018) 117–128.

[14] A. Adeel, M. Gogate, A. Hussain, Towards next-generation lip-reading driven hearing-aids: A preliminary prototype demo, in: International Workshop on Challenges in Hearing Assistive Technology (CHAT-2017), Stockholm University, August 19th, Collocated with Interspeech 2017, 2017.

[15] A. Adeel, M. Gogate, A. Hussain, W. M. Whitmer, Lip-reading driven deep learning approach for speech enhancement, arXiv preprint arXiv:1808.00046.

[16] A. Adeel, M. Gogate, A. Hussain, Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments, Information Fusion.

[17] D. Rethage, J. Pons, X. Serra, A wavenet for speech denoising, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5069–5073.

[18] A. Pandey, D. Wang, A new framework for supervised speech enhancement in the time domain., in: Interspeech, 2018, pp. 1136–1140.

[19] Y. Luo, N. Mesgarani, Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (8) (2019) 1256–1266.

[20] J. Barker, R. Marxer, E. Vincent, S. Watanabe, 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.

[21] S. Pascual, M. Park, J. Serrà, A. Bonafonte, K.-H. Ahn, Language and noise transfer in speech enhancement generative adversarial network, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5019–5023.

[22] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, The Journal of the Acoustical Society of America 120 (5) (2006) 2421–2424.

[23] N. Harte, E. Gillen, Tcd-timit: An audio-visual corpus of continuous speech, IEEE Transactions on Multimedia 17 (5) (2015) 603–615. doi:10.1109/TMM.2015.2407694.

[24] A. Varga, H. J. 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, Speech communication 12 (3) (1993) 247–251.

[25] D. Snyder, G. Chen, D. Povey, Musan: A music, speech, and noise corpus, arXiv preprint arXiv:1510.08484.

[26] A. Owens, A. A. Efros, Audio-visual scene analysis with self-supervised multisensory features, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 631–648.

[27] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, Deep audio-visual speech recognition, IEEE transactions on pattern analysis and machine intelligence.

[28] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, A. Torralba, The sound of pixels, in: The European Conference on Computer Vision (ECCV), 2018.

[29] J. R. Hershey, Z. Chen, J. Le Roux, S. Watanabe, Deep clustering: Discriminative embeddings for segmentation and separation, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 31–35.

[30] D. Griffin, J. Lim, Signal estimation from modified short-time fourier transform, IEEE Transactions on Acoustics, Speech, and Signal Processing 32 (2) (1984) 236–243.

[31] D. Yu, M. Kolbæk, Z.-H. Tan, J. Jensen, Permutation invariant training of deep models for speaker-independent multi-talker speech separation, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 241–245.

[32] U. Kjems, M. S. Pedersen, J. B. Boldt, T. Lunner, D. Wang, Speech intelligibility of ideal binary masked mixtures, in: 2010 18th European Signal Processing Conference, IEEE, 2010, pp. 1909–1913.

[33] M. Ahmadi, V. L. Gross, D. G. Sinex, Perceptual learning for speech in noise after application of binary time-frequency masks, The Journal of the Acoustical Society of America 133 (3) (2013) 1687–1692.

[34] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, T. Lunner, Speech intelligibility in background noise with ideal binary time-frequency masking, The Journal of the Acoustical Society of America 125 (4) (2009) 2336–2347.

[35] Y. Wang, A. Narayanan, D. Wang, On training targets for supervised speech separation, IEEE/ACM transactions on audio, speech, and language processing 22 (12) (2014) 1849–1858.

[36] D. Wang, J. Chen, Supervised speech separation based on deep learning: An overview, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (10) (2018) 1702–1726.

[37] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122.

[38] Y. Li, X. Zhang, D. Chen, Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1091–1100.

[39] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[40] F. Developers, ffmpeg tool [software], `http://ffmpeg.org/` (2000–2019).

[41] D. E. King, Dlib-ml: A machine learning toolkit, Journal of Machine Learning Research 10 (2009) 1755–1758.

[42] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[43] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, 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.

[44] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time–frequency weighted noisy speech, IEEE Transactions on Audio, Speech, and Language Processing 19 (7) (2011) 2125–2136.

[45] J. Le Roux, S. Wisdom, H. Erdogan, J. R. Hershey, 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.

[46] Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Transactions on audio, speech, and language processing 16 (1) (2007) 229–238.

[47] 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.

[48] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE transactions on acoustics, speech, and signal processing 33 (2) (1985) 443–445.

[49] S. Pascual, A. Bonafonte, J. Serrà, Segan: Speech enhancement generative adversarial network, Proc. Interspeech 2017 (2017) 3642–3646.

[50] I. Recommendation, 1534-1,"method for the subjective assessment of intermediate sound quality (mushra)", International Telecommunications Union, Geneva, Switzerland.