Deep Neural Network Driven Binaural Audio Visual Speech Separation

1st Mandar Gogate
School of Computing
Edinburgh Napier University
Edinburgh, UK
m.gogate@napier.ac.uk

2nd Kia Dashtipour
School of Computing,
Edinburgh Napier University
Edinburgh, UK
k.dashtipour@napier.ac.uk

3rd Peter Bell
School of Informatics
University of Edinburgh
Edinburgh, UK
peter.bell@ed.ac.uk

4th Amir Hussain
School of Computing
Edinburgh Napier University
Edinburgh, UK
a.hussain@napier.ac.uk

***Abstract*—The central auditory pathway exploits the auditory signals and visual information sent by both ears and eyes to segregate speech from multiple competing noise sources and help disambiguate phonological ambiguity. In this study, inspired from this unique human ability, we present a deep neural network (DNN) that ingest the binaural sounds received at the two ears as well as the visual frames to selectively suppress the competing noise sources individually at both ears. The model exploits the noisy binaural cues and noise robust visual cues to improve speech intelligibility. The comparative simulation results in terms of objective metrics such as PESQ, STOI, SI-SDR and DBSTOI demonstrate significant performance improvement of the proposed audio-visual (AV) DNN as compared to the audio-only (A-only) variant of the proposed model. Finally, subjective listening tests with the real noisy AV ASPIRE corpus shows the superiority of the proposed AV DNN as compared to state-of-the-art approaches.**

***Index Terms*—Binaural Speech Separation, Audio-Visual, Deep Learning, Mask Estimation**

I. Introduction

The human brain has a remarkable ability to track and segregate speech in everyday listening environments with multiple competing noise sources, including reflections from physical surfaces. The separation of noise from the target sound is important in many applications including hearing aids, cochlear implants, speech recognition and mobile telecommunication [1]. However, despite the extensive research in the area of signal processing, speech separation remains a technical challenge as the current systems often make the speech more audible but do not always restore intelligibility in busy social situations [2]. This often result in isolation of hearing impaired listener leading to a range of negative consequences including depression [2]. Despite the extensive research advancements over the past decades in the area of speech separation, the listening scenarios have become more complex with wide range of non-stationary noises and reverberation in physical space.

Speech separation techniques can be mainly divided into two categories: single channel and multichannel [3]. The main limitation with single channel separation is that the process of noise separation often distorts the speech signal, especially when the target speech and background noise are overlapped. However, generally the speech and noises are located at different spatial positions in the physical space and can be exploited for enhanced speech separation. Therefore, beamforming and multichannel wiener filter based approaches have shown to achieve better performance as compared to single channel approaches. Recently, deep neural networks (DNNs) have been widely used in both single channel and multichannel speech separation (including binaural) due to its ability to construct statistical models from large supervised corpus. In addition, it has been observed that DNN based approaches often perform better than traditional multichannel speech separation approaches [1].

In the literature, extensive research has been carried out to develop binaural audio-only (A-only) speech separation approaches based on beamforming, multichannel wiener filter as well as with DNN that exploits a range of spatial and spectral features [4]–[7]. In addition, it has been shown that, the visual cues help improve performance of single channel audio-visual (AV) speech separation as compared to A-only speech separation especially at low signal-to-noise ratios (SNRs) [8]–[11]. However, to the best of our knowledge, the integration of binaural cues (providing spatial information) and visual lip images (helping disambiguate phonetic ambiguities) have not yet been explored for more robust speech separation. We hypothesise that the exploitation of spatial and visual information will improve noise reduction while preserving the binaural cues leading to enhanced speech understanding in noise (i.e. the cocktail party problem).

In this study, we propose a multimodal DNN that exploits the binaural cues received at two ears and the visual input to enhance speech of unknown speaker from unseen binaural noise. The developed hybrid DNN model integrates a multi-stream convolutional neural network for optimised binaural ideal binary mask (IBM) estimation, taking into account the temporal and spatial dynamics of binaural audio as well as temporal visual lip images. The developed model learns a joint representation of features extracted from left channel, right channel and visual data to estimate a spectral mask per channel. An overview of the proposed binaural AV framework is depicted in Fig 1. The multi-stream DNN architecture, presented in Fig 2 effectively exploits the correlation between binaural noisy features, noise robust visual features and IBM to estimate speech and noise dominant regions for each channel. The estimated IBM is then applied to the time-frequency (T-

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

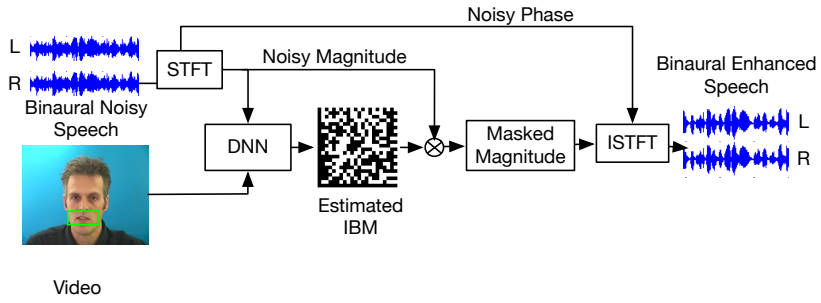

Fig. 1. Framework Overview

F) representation of noisy speech to produce the enhanced binaural speech. We perform an extensive evaluation of our proposed AV model, using real noisy ASPIRE corpus [12], with state-of-the-art single-channel A-only SE approaches (including spectral subtraction (SS), and log-minimum mean square error (LMMSE)) using objective measures (PESQ, SI-SDR, STOI and DBSTOI) and subjective MOS listening tests.

The rest of the paper is organised as follows: Section 2 briefly reviews the related work. Section 3 presents the AV binaural dataset and preprocessing. Section 4 introduce the proposed framework used for binaural AV speech separation. Section 5 explains the experimental results and finally, Section 6 concludes this work and propose future directions.

II. Related Work

This section briefly reviews the related works in the area of A-only binaural speech separation and single channel AV speech separation.

A. Binaural Audio-only Speech Separation

In this subsection, we present an overview of the state-of-the-art binaural speech separation approach that have shown to outperform single channel approaches.

Wang et al. [4] proposed a binaural speech enhancement in virtual reality scenes consisting of multiple noise sources, range of SNRs and sound directions. Two DNN models were proposed: one that maps the binaural noisy features to clean speech features, the other DNN handled the two channels separately. On the other hand, Wood et al. [13] proposed the atomic speech presence probability (ASPP) speech enhancement framework that exploits interaural level difference (ILD), interaural phase difference (IPD), interaural coherence magnitude (ICM), and their combinations found in the interaural transfer function (ITF). Comparative simulation results reveal the usefulness of the proposed approach as compared to state-of-the-art approaches for binaural noise reduction and binaural cue preservation.

B. Single channel Audio-Visual Speech Separation

In this subsection, we present an overview of state-of-the-art AV speech separation approach that have shown to perform better than A-only approaches.

Ephrat et al. [14] proposed a speaker independent AV DNN for complex ratio mask estimation to separate speech from overlapping speech and background noises. The main limitation with the study is that the model is trained and evaluated on a fixed SNR that do not reflect real world scenarios as in later SNR generally varies over a wide range (e.g. from -12dB to 18dB). Similarly, Gogate et al. [9] presented a single channel AV framework for mask estimation to separate speech from background noises. It is shown that, the model separates the speech of an unknown speaker from unseen noises.

In addition, Hou et al. [8] proposed an AV SE model for Mandarin that is trained and evaluated on a single speaker. The model predicts the enhanced spectrogram from the noisy spectrogram using AV deep convolutional network. On the other hand, Gabbay et al. [15] trained a convolutional autoencoder architecture to map the noisy to enhanced speech using speech spectrogram and cropped lip regions. However, the model fails to work when the visuals are occluded. Finally, Adeel et al. [10], [16] proposed an AV SE models by exploiting an enhanced visually-derived wiener filter (EVWF) and lip reading regression model. The preliminary results demonstrated the effectiveness to deal with spectro-temporal variations in wide variety of noisy environments.

It can be seen that, none of the aforementioned studies have explored the integration of noisy binaural acoustic cues and noise robust visual features for more robust speech separation. We hypothesise that joint exploitation of multimodal spatial and temporal cues will lead to improve noise reduction while preserving the binaural cues.

III. AV Binaural Dataset and Preprocessing

A. Binaural Noises

In order to build a synthetic mixture of binaural AV dataset for training we collected approximately 36 hours of binaural noises in real noisy environments including social gatherings, train stations, cafeteria, restaurants, and streets. The listener was wearing a binaural microphone connected to the Zoom H4n pro recorder, recording at a sampling rate of 44.1 kHz. It is to be noted that, most of the aforementioned studies used a simulation environment to generate the synthetic noisy dataset with noise sources manually placed at a specific position in the 3D space.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE I
GRID/ASPIRE SENTENCE GRAMMAR

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

B. Grid Corpus + Binaural Noises

The benchmark Grid corpus [17] is utilised for the training and evaluation of the proposed DNN. The Grid sentence structure is shown in Table 2. The Grid corpus is randomly combined with the recorded binaural noises for SNR ranging from -12 to 6 dB with 6dB increment. It is worth mentioning that, for training, 4000 utterances from 4 speakers (2 male and 2 female) were used. The model was validated and tested on 2000 utterance from 2 speakers (1 male and 1 female).

C. ASPIRE Corpus

The ASPIRE corpus [12] is an AV binaural speech corpus recorded in real noisy environments such as cafeteria and restaurant. The corpus consists of 3000 sentences recorded from three speakers and is used for the subjective evaluation of the proposed binaural AV DNN based speech separation.

D. Pre-processing

1) Audio: The audio signals are resampled at 16 kHz. The resampled audio signals are segmented into 65 milliseconds (ms) frames and 20% increment rate. A hanning window and STFT is applied to produce 526-bin magnitude spectrogram.

2) Video: The speakers lip images are extracted the 25 frames-per-second (fps) Grid corpus video using a minified dlib [18] model for extracting the lip landmarks. The extracted lip regions are converted to grey scale and resized to 40 pixels x 80 pixels. It is worth mentioning that, the lip sequence is extracted at 25 fps while the audio features are extracted at 75 vectors-per-second (VPS).

IV. DNN DRIVEN BINAURAL AV MASK ESTIMATION

This section describes the DNN architecture, depicted in Fig. 2, that ingests the binaural audio and visual cues, to estimate an IBM for each channel. Finally, speech is reconstructed by applying the estimated mask to the T-F representation of the noisy magnitude spectrum.

A. Data Representation

a) Input features: The DNN ingests the raw magnitude spectrum of both left and right channel, as well as the cropped lip regions 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 for 3 second clip recorded at 25 fps). For audio input, we compute STFT of audio segments and a magnitude spectrogram is used. It is to be noted that, the trained model can be applied to sequences of arbitrary lengths during inference.

b) Output: The output of our network is IBM for left and right channel. IBM is 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:

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

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.

1) Audio Feature Extraction: The audio feature extraction part of the network consists of 5 dilated convolutional layers. The first four convolutional layers consist of 96 filters and the last convolutional layer consists of 4 filters. Each filter has size 5 x 5. The dilation across time dimension is increased by a factor of 2 after each layer: the first layer has dilation of 1 x 1, the second layer has 2 x 1 and so on. The final convolutional layer has dilation of 1 x 1. After each convolutional layer, ReLU activation is applied. The output of the last convolutional layer is fed into the AV fusion part of the framework as shown in Fig. 2. It is to be noted that, no pooling is applied after convolutional layers and the dilated convolutional weights are shared across the left and right channel.

2) Visual Feature Extraction: The cropped temporal lip images are fed into visual feature extraction part of the framework with four convolutional layers with 32, 64, 96 and 96 filters respectively. Each convolutional filter has size 3 x 3 x 3. ReLU activation is used after each convolutional layer. The output of the last convolutional layer is fed into a long short-term memory (LSTM) layer with 526 cells. Max pooling of size (1 x 2 x 2) is applied after each convolutional layer. The visual features extracted at 25 frames-per-second were upsampled by a factor of 3 to match the audio feature sampling rate (i.e. 75 vectors-per-second). The output of the LSTM layer is then fed into the AV fusion part of the framework. It is to be noted that, the convolutional weights are shared across the temporal dimension.

B. Audio-Visual Fusion

The feature extracted from binaural audio and video were concatenated across the time dimension and were fed to a fully connected layer with 526 units and sigmoid activation. It is to be noted that, the A-only baseline model is constructed by removing the video feature extraction part of the network.

C. Mathematical Representation

The framework, shown in Fig. 1, ingests binaural noisy speech (X) and video (V) to output binaural enhanced speech ($\hat{X}$). Let $A_L n, A_L n-1, ..., A_L 1$ and $A_R n, A_R n-1, ..., A_R 1$

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

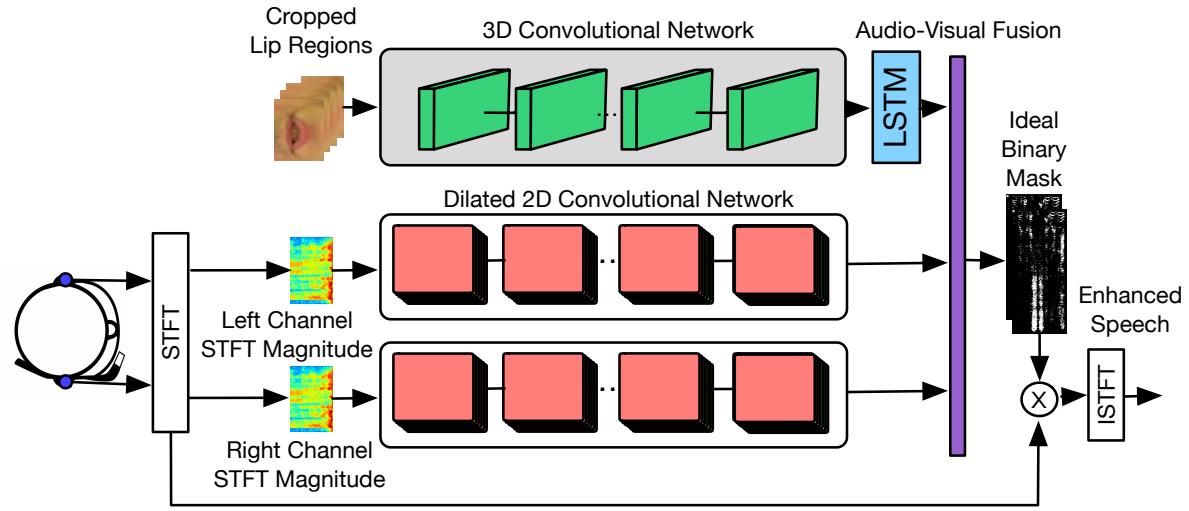

Fig. 2. Proposed DNN based speech separation Baseline Model

be the noisy STFT features obtained from left and right channel of x. In addition, let $\hat{A}_L n, \hat{A}_L n-1, ..., \hat{A}_L 1$, $\hat{A}_R n, \hat{A}_R n-1, ..., \hat{A}_R 1$ be the enhanced STFT features of left and right channel. Finally, let $F_t, F_{t-1}, ..., F_1$ 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 the current window frame. Let M_L and M_R be the IBM and $\hat{M}_L$ and $\hat{M}_R$ be the estimated IBM. The framework can be represented as follows:

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

$$\hat{A}_L n = \hat{M}_L n \bigodot A_L n \tag{3}$$

$$\hat{A}_R n = \hat{M}_R n \bigodot A_R n \tag{4}$$

where $\bigodot$ represents the element wise multiplication.

The DNN, shown in Fig. 2, ingest the noisy STFT features $(A_L 1, A_L 2, ..., A_L n, A_R 1, A_R 2, ..., A_R n)$ and cropped images of speakers lips $(F_1, F_2, ..., F_t)$ as an input to output a multiplicative T-F masks ($\hat{M}_L n$ and $\hat{M}_R n$) for the left and right channel. The DNN can be represented as follows:

$$\hat{M}_L n, \hat{M}_R n = g(A_L 1, .., A_L n, A_R 1, .., A_R n, F_1, .., F_t) \tag{5}$$

i.e.

$$\hat{M}_L 1, \hat{M}_R 1 = g(A_L 1, A_R 1, F_1) \tag{6}$$

$$\hat{M}_L 2, \hat{M}_R 2 = g(A_L 1, A_L 2, A_R 1, A_R 2, F_1, F_2) \tag{7}$$

It can be seen that, the model can be used for streaming prediction as the estimated IBM 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 sum of binary cross entropy between the $M_L n$, $M_R n$ and $\hat{M}_L n$, $\hat{M}_R n$. The loss function can be represented as follows:

$$Loss = BCE(\hat{M}_L t, M_L t) + BCE(\hat{M}_R t, M_R t) \tag{8}$$

where

$$BCE(\hat{M}_t, M_t) = -\frac{1}{N}\sum_{i=1}^{N} M_t \cdot log(\hat{M}_t) + (1 - M_t) \cdot log(1 - \hat{M}_t) \tag{9}$$

D. Speech Resynthesis

The model estimates a T-F IBM for each binaural channel. The estimated multiplicative spectral mask is applied to the respective noisy magnitude spectogram. The masked magnitude is then combined with the noisy phase to get the enhanced speech using ISTFT.

V. Experiments

A. Experimental Setup

In order to evaluate the performance of the approach, the network is trained using Pytorch library and NVIDIA Titan Xp GPUs. The subset of speakers from Grid binaural corpus as explained in section III-B are used for training and evaluation of the proposed DNN, and the ASPIRE corpus to test the performance of the approach. It should be pointed out that, there is no overlap of speakers and noises present in the train, validation and test set to ensure the speaker and noises independent criteria. The network is trained for 50 epochs using backpropagation with Adam optimiser [19] 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.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

B. Objective Testing

In order to measure the objective quality of the resynthesized speech PESQ, STOI, SI-SDR and DBSTOI are employed. The comparison of noisy speech, and enhanced speech of SS, LMMSE, A-only, AV and Oracle IBM using the aforementioned metrics is presented in Table II. PESQ [20] is one of the most commonly used objective assessment metric to predict the subjective listening test scores in the speech enhancement literature and has shown to correlate well with the subjective listening tests. The PESQ score ranging from -0.5 to 4.5, indicating the minimum and maximum possible reconstructed speech quality, is measured as linear combination of the average disturbance value and the average asymmetrical disturbance values. In addition, STOI [21] is another benchmark objective evaluation metric used for speech intelligibility that shows a high correlation with subjective listening test scores. The correlation of short-time temporal envelopes between the clean and modified speech is calculated in STOI with values ranging from $[0, 1]$, where higher value indicates better intelligibility. DBSTOI [22] is a binaural extension of the traditional STOI that ingests the noisy signals as provided in the left and right ears of the listeners, and clean signal in both ears. The DBSTOI ranges from 0 to 1. Finally, Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [23] is used to calculate the amount of distortion introduced by the process of speech separation. 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 objective evaluation metrics for Noisy, SS, LMMSE, A-only DNN, AV DNN and Oracle IBM are presented in Table. II and Fig. 3. It can be seen that, at low SNRs, AV and A-only model outperformed SS [24], LMMSE [25] based speech enhancement methods. However, for higher SNRs the performance of SS, LMMSE, A-only and AV models have relatively similar performance. In addition, AV performs better than A-only model especially for low SNR ranges (i.e. $SNR <= 0$ dB), where AV model achieved the DBSTOI scores of 0.607, 0.693, and 0.780 at SNR levels of -12 dB, -6 dB, and 0 dB respectively, as compared to 0.557, 0.655, and 0.768 achieved by A-only model.

C. Subjective Listening tests

In the literature, a significant number of objective metrics [20], [21], [23] 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. Therefore, to explore the generalisation of the proposed binaural speech separation DNN in real noisy settings, subjective listening tests were carried out with 15 self reported normal hearing listeners. Each listener was presented with 25 randomly selected utterances from real noisy ASPIRE corpus. The listeners were presented enhanced speech and were asked to rate the resynthesized speech quality on a scale of 0 to 5. The rating score were: 0) - Incomprehensible, 1) - Very Annoying (Bad), (2) - Annoying (Poor), (3) - Slightly Annoying (Fair) (4) Perceptible but annoying (Good), (5) - Perceptible (Excellent). The noisy speech was included in the test for participants to have a reference for the degraded speech. The proposed AV model is compared with SS, LMMSE, A-only model. Fig. 4 indicates the box plot of subjective ratings from the MOS test. It can be seen that, the binaural AV model significantly outperforms the SS, LMMSE and A-only model. Furthermore, the results show the ability of the proposed DNN trained using synthetic AV dataset to generalise to real noisy settings, including the reverberation caused by multiple competing background sources. The results demonstrate the capability of proposed DNN to deal with the reverberation caused by multiple competing background sources observed in a real-world noisy environment, by exploiting the binaural audio and visual cues. In addition, the results show that an AV model trained on synthetic additive mixtures generalise well on real noisy corpus.

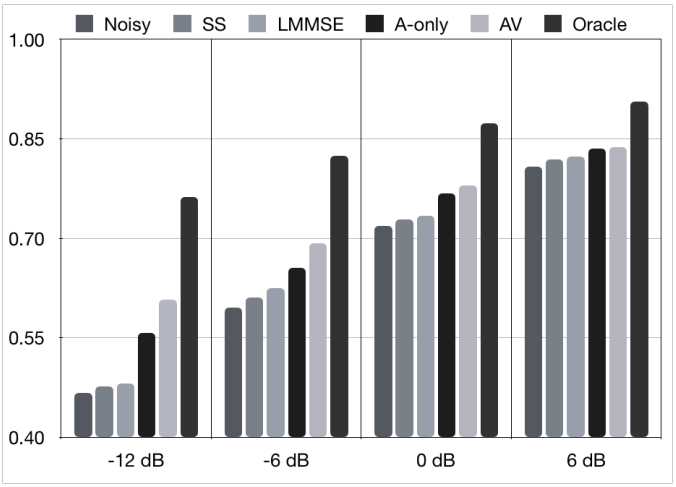

Fig. 3. Comparison of DBSTOI for Noisy, SS, LMMSE, A-only DNN, AV DNN and Oracle IBM for the synthetic AV test set

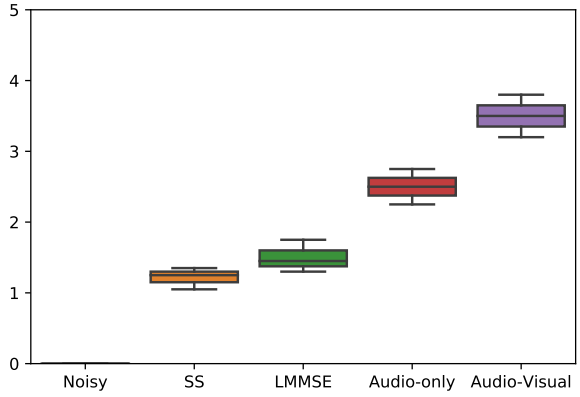

Fig. 4. Result of MOS Subjective listening test

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE II
OBJECTIVE EVALUATION METRICS FOR UNPROCESSED NOISY, SS, LMMSE, A-ONLY DNN, AV DNN AND ORACLE IBM FOR THE SYNTHETIC AV TEST SET

SNR	DBSTOI						Average PESQ					
	Noisy	SS	LMMSE	A-only	AV	Oracle	Noisy	SS	LMMSE	A-only	AV	Oracle
-12 dB	0.467	0.477	0.481	0.557	0.607	0.762	1.348	1.401	1.424	1.718	1.885	2.346
-6 dB	0.596	0.611	0.625	0.655	0.693	0.825	1.605	1.712	1.731	2.168	2.271	2.687
0 dB	0.719	0.728	0.734	0.768	0.780	0.874	2.035	2.211	2.256	2.589	2.647	2.959
6 dB	0.808	0.819	0.823	0.835	0.837	0.906	2.435	2.701	2.754	2.909	2.924	3.238

SNR	Average STOI						Average SI-SDR					
	Noisy	SS	LMMSE	A-only	AV	Oracle	Noisy	SS	LMMSE	A-only	AV	Oracle
-12 dB	0.520	0.521	0.523	0.532	0.598	0.717	-11.167	-9.245	-8.441	0.066	1.029	9.348
-6 dB	0.599	0.603	0.611	0.637	0.682	0.772	-5.724	-3.271	-2.154	5.833	6.356	12.055
0 dB	0.691	0.702	0.711	0.730	0.729	0.819	0.124	3.112	3.451	10.999	11.269	14.802
6 dB	0.768	0.770	0.775	0.788	0.799	0.851	6.133	9.137	10.145	14.179	14.522	17.294

VI. CONCLUSION

This paper introduced a novel DNN based speaker independent binaural AV speech separation model that contextually integrates and exploits features received at both ears and lip images, independent of the SNR, for robust speech separation. The performance evaluation in terms of objective metrics including PESQ, STOI, SI-SDR and DBSTOI revealed significant performance improvement of our proposed AV model as compared to SS, LMMSE and A-only binaural speech separation model. In addition, we show that the proposed AV DNN trained on a synthetic mixture of GRID and binaural noises generalises well on the real noisy ASPIRE corpus. It is worth mentioning that, both GRID and ASPIRE follows the same sentence structure/vocabulary and could help achieve the superior performance. Ongoing work includes the exploitation of spatial features such as phase or level differences (ITDs, IPDs or ILDs) for more robust speech separation. In addition, we intend to compare the performance of the proposed AV binaural framework with other state-of-the-art binaural A-only approaches. In future, we intend to investigate the generalisation capability of our proposed DNN model with other more challenging real noisy AV corpora.

VII. 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 Ahsan Adeel, Dr Ricard Marxer and Prof Jon Barker.

REFERENCES

[1] Xueliang Zhang and DeLiang Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[2] Nicholas A Lesica, "Hearing aids: Limitations and opportunities," *The Hearing Journal*, vol. 71, no. 5, pp. 43–46, 2018.

[3] Naohiro Tawara, Tetsunori Kobayashi, and Tetsuji Ogawa, "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder," *Proc. Interspeech 2019*, pp. 86–90, 2019.

[4] Jin Wang, Jing Wang, Ming Liu, and Zhaoyu Yan, "Binaural speech enhancement based on dnn for the application of virtual reality," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*. IEEE, 2018, pp. 629–633.

[5] Mathew Shaji Kavalekalam, Jesper K Nielsen, Mads G Christensen, and Jesper B Boldt, "Hearing aid-controlled beamformer for binaural speech enhancement using a model-based approach," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 321–325.

[6] Tobias May, Steven Van de Par, and Armin Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2016–2030, 2012.

[7] Qingju Liu, Wenwu Wang, Philip Jackson, and Mark Barnard, "Reverberant speech separation based on audio-visual dictionary learning and binaural cues," in *2012 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2012, pp. 664–667.

[8] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[9] Mandar Gogate, Ahsan Adeel, Ricard Marxer, Jon Barker, and Amir Hussain, "Dnn driven speaker independent audio-visual mask estimation for speech separation," *Proc. Interspeech 2018*, pp. 2723–2727, 2018.

[10] Ahsan Adeel, Mandar Gogate, Amir Hussain, and William M Whitmer, "Lip-reading driven deep learning approach for speech enhancement," *arXiv preprint arXiv:1808.00046*, 2018.

[11] Mandar Gogate, Ahsan Adeel, Kia Dashtipour, Peter Derleth, and Amir Hussain, "Av speech enhancement challenge using a real noisy corpus," *arXiv preprint arXiv:1910.00424*, 2019.

[12] Mandar Gogate, Kia Dashtipour, Ahsan Adeel, and Amir Hussain, "Cochleanet: A robust language-independent audio-visual model for speech enhancement," *Information Fusion*, 2020.

[13] Sean UN Wood, Johannes KW Stahl, and Pejman Mowlaee, "Binaural codebook-based speech enhancement with atomic speech presence probability," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2150–2161, 2019.

[14] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 112, 2018.

[15] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg, "Visual speech enhancement," in *Interspeech*. 2018, pp. 1170–1174, ISCA.

[16] Ahsan Adeel, Mandar Gogate, and Amir Hussain, "Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments," *Information Fusion*, 2019.

[17] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[18] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[20] Antony W Rix, John G Beerends, Michael P Hollier, and Andries 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)*. IEEE, 2001, vol. 2, pp. 749–752.
[21] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
[22] Asger Heidemann Andersen, Jan Mark de Haan, Zheng-Hua Tan, and Jesper Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
[23] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John 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.
[24] 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.* IEEE, 1979, vol. 4, pp. 200–203.
[25] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE