TOWARDS POSE-INVARIANT AUDIO-VISUAL SPEECH ENHANCEMENT IN THE WILD FOR NEXT-GENERATION MULTI-MODAL HEARING AIDS

Mandar Gogate, Kia Dashtipour, Amir Hussain

School of Computing, Edinburgh Napier University, Scotland, UK

ABSTRACT

Classical audio-visual (AV) speech enhancement (SE) and separation methods have been successful at operating under constrained environments; however, the speech quality and intelligibility improvement is significantly reduced in unconstrained real-world environments where variation in pose and illumination are encountered. In this paper, we present a novel privacy-preserving approach for real world unconstrained pose-invariant AV SE and separation that contextually exploits pose-invariant 3D landmark flow features and noisy speech features to selectively suppress unwanted background speech and non-speech noises. In addition, we present a unified architecture that integrates state-of-the-art transformers with temporal convolution neural networks for effective pose-invariant AV SE. The preliminary systematic experimentation on benchmark multi-pose OuluVS2 and LRS3-TED corpora demonstrate that the privacy preserving 3D landmark flow features are effective for pose-invariant SE and separation. In addition, the proposed AV SE model significantly outperforms state-of-the-art audio-only SE model, oracle ideal binary mask, and A-only variant of the proposed model in speaker and noise independent settings.

Index Terms— Audio-visual speech enhancement, pose-invariant, multimodal hearing aids

1. INTRODUCTION

Speech enhancement (SE) is used to improve the speech intelligibility in the presence of background interfering noises. SE has been used in diverse real-world applications, including hearing aids, smart human-computer interaction systems, teleconferencing, and automatic speech recognition [1]. Despite considerable research efforts in the area of SE, understanding speech in the presence of multiple competing background sources, commonly encountered in cocktail party scenarios, has been a major challenge for several decades [2].

Although the incorporation of visual modality in audio-only SE models [3, 4, 5, 6, 7] demonstrated a notable improvement in constraints cocktail party environments (with near frontal speaker poses), the more general issue of audio-visual (AV) SE in the context of pose variations remains largely unexplored in real-world applications. It is to be noted that, the majority of state-of-the-art AV SE models employ a corpus recorded in studio environments [1] with near frontal face views used for both training and evaluation. As a result, the performance of such models degrade in real world cocktail party like environments. However, in future AV hearing aids, the speaker may not always face the hearing-impaired listener, especially when addressing a group of more than two individuals.

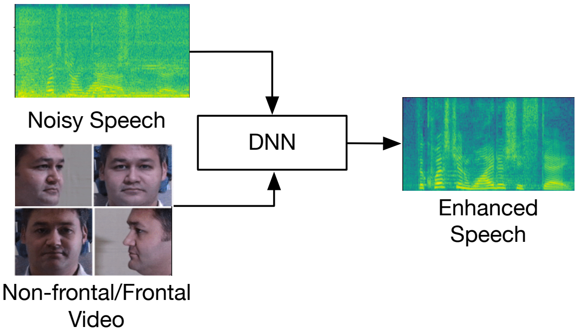

Fig. 1. System overview: Pose-invariant Audio-Visual Speech Separation

The main objective of this study is to address the performance disparity shown by AV SE models when confronted with pose variations. To this end, we propose a novel transformer based architecture that leverages visual features extracted from the target speaker's face to isolate their voice from competing background speech and noise sources. Specifically, our framework generates an enhanced audio signal that exclusively contains the target speaker's voice when supplied with a noisy audio signal and frontal/non-frontal target speaker video as inputs. Simultaneously, our framework effectively suppresses background noise (speech and/or noise) regardless of the target speaker's pose. An overview of the new framework is depicted in Figure 1.

The proposed model leverages the complementary strengths of multi-headed attention, transformer and temporal convolutional networks for optimal AV SE in real-world unconstrained environments. Specifically, a unified deep neural network model effectively learns correlations between noisy speech and optical flow based 3D landmark flow [8] features to generate spectral mask irrespective of pose variations. The application of spectral mask to noisy speech features retains the target speech dominant regions and suppresses

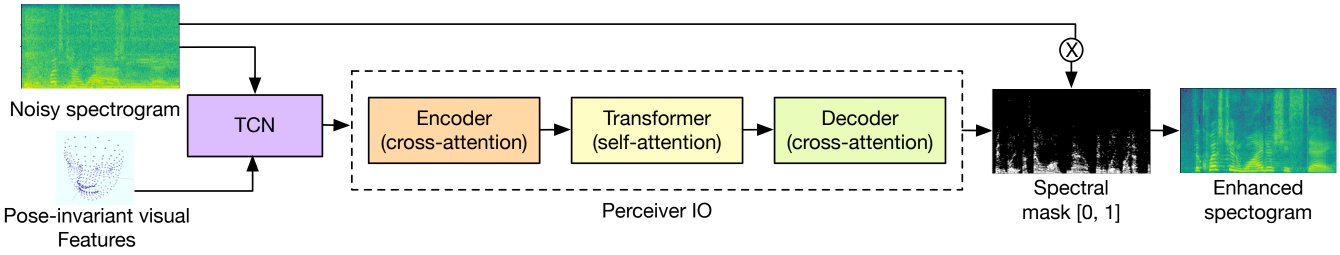
Noisy spectrogram
TCN
Pose-invariant visual Features
Encoder (cross-attention)
Transformer (self-attention)
Decoder (cross-attention)
Perceiver IO
Spectral mask [0, 1]
Enhanced spectogram

Fig. 2. Proposed Speech Separation Model

background noise dominant regions. The enhanced speech is resynthesised by combining the masked spectogram with noisy phase. To the best of our knowledge, our study is the first to propose a privacy-preserving pose-invariant AV SE framework that is speaker and noise independent.

The rest of the paper is organised as follows: section 2 presents the proposed pose-invariant framework; section 3 provides synthetic AV corpora and data preprocessing; section 4 presents the experimental results; and finally section 5 concludes the work and proposes some future directions.

2. POSE-INVARIANT AUDIO-VISUAL SPEECH ENHANCEMENT

This section describes our proposed deep neural network (DNN) architecture depicted in Fig. 2. The DNN ingests optical flow of 3D landmark flow features and a noisy magnitude spectogram to generate an enhanced spectogram as output. The enhanced spectogram is combined with the noisy phase to resynthesise enhanced speech.

2.1. Model architecture

The proposed model consists of three parts: a temporal convolutional network [9], Perceiver IO [10] and a fully connected layer.
1. Temporal convolutional network (TCN): The TCN ingests the concatenation of noisy spectogram and upsampled pose invariant Face Mesh features as input. Specifically, the TCN consists of multiple temporal blocks with exponential increase (2^{i-1} where i varies from 1 to number of temporal blocks). Each temporal block consists of 4 convolution layers where the first/third are depth-wise convolutional layers with a dilation factor 2^{i-1} and the second/fourth are pointwise convolutional layers with a dilation factor of 1. Each convolution layer is followed by batch normalisation, a parametric rectified linear unit (PReLU) and a dropout layer. The TCN part of the proposed model contains 4 temporal blocks with dilation 1, 2, 4, and 8 respectively. The kernel size and dropout were set to 3 and 0.05 respectively. The extracted features are fed to Perceiver IO for further processing.
2. Perceiver IO: The Perceiver IO is a generic DNN architecture that has been shown to achieve state-of-the-art results in a wide variety of applications including language modelling, multimodal auto encoding and optical flow prediction. The model scales linearly in terms of processing time and model complexity as the input dimension is increased, making it attractive for practical applications. The Perceiver IO mainly consists of 3 layers: (1) encoder with cross-attention (2) a series of transformer modules with self attention and (3) decoder with self attention. The encoder module maps the inputs to a latent space by applying a cross-attention module. The encoded latent space is then processed using a series of transformer modules. The processed latent space is then fed to a decoder module that maps the latent space to the output dimension after applying cross attention. The input, latent and output dimensions of the Perceiver IO block are 377, 257 and 257 respectively. The cross attention modules present in the encoder and decoder layers comprise 4 heads and 16 dimensions each. In addition, the model consists of 3 transformer modules stacked on top of each other. The self attention head present in each transformer module consists of 4 heads and 16 dimension per head. The aforementioned parameters leads to a Perceiver IO module with 6M parameters.
3. Fully connected layer: The fully connected layer consists of 257 neurons with sigmoid activation to predict a time-frequency mask. The predicted spectral mask is multiplied with the noisy speech magnitude spectogram to generate a masked spectogram as the network output. The enhanced speech is resynthesised by combining the masked magnitude with the noisy phase.

3. AV DATASET AND PRE-PROCESSING

3.1. Datasets

The models are trained and evaluated using synthetic benchmark corpora generated using OuluVS2 [11] and LRS3-TED [12]. Specifically, two different scenarios were considered for each corpus: (1) Two speaker mixture (2Mix) - target speaker mixed with background speaker at a randomly selected SNR ranging from a uniform distribution between 0 dB and +10 dB (2) Two speaker mixture with real ambient background noise (2Mix + Noise) - target speaker mixed with background speaker (at SNR ranging from 0 dB to +10 dB) and background noise (at SNR ranging from -6 dB to +6 dB).

OuluVS2: The OuluVS2 corpus was recorded for non-rigid mouth motion analysis with 53 speakers (40 males and 13 females) and simultaneous recording of five different views: 0° (frontal), 30° , 45°, 60° and 90° (profile). The availability of multi-view recordings makes the corpus suitable for training pose-invariant SE model. For training, validation and testing, the data was split into 37, 6, and 10 speakers respectively.

LRS3-TED: The LRS3-TED corpus consists of videos of around 4500 speakers collected from TED and TEDx. As compared to OuluVS2 the data do not consist of explicit fixed view of the speaker. However, as the speaker naturally moves across the stage the videos consist of wide variety of facial postures. For training, validation and testing, the data was split into 3500, 500, and 412 speakers respectively.

WHAM! Noises: The WHAM corpus [13] was developed to benchmark SE methods with a more realistic cocktail party scenario where a two speaker mixture is combined with real-world ambient noises recorded in coffee shops, restaurants, and bars in the San Francisco Bay Area. The dataset is used for generating 2Mix + Noise scenario.

3.2. Data preprocessing

Audio: The audio signal is resampled to 16000 Hz sampling frequency. The signals are then segmented into 32 ms frames (512 samples per frame) with a 8 ms frame increment (128 sample per increment). A short time Fourier transform (STFT) with hanning window is applied to produce a 257 bin spectrum. The magnitude of the spectrum is fed as input to the model.

Video: The videos are resampled at 25 frames per second. BlazeFace [14] is used to extract the face region from the video. The region is resized to a square of size 224×224. The cropped region is then fed to Face Mesh [8] model to extract 486 dimensional 3D facial landmarks per frame. Since there is high correlation between the audio and lip movements, only the lip part of Face Mesh features is considered resulting in 40 dimensional 3D lip landmark features per frame (3×40). The landmark of each current frame is subtracted from the previous frame to generate an optical flow of landmarks as visual features. The visual features are upsampled to match the audio feature sampling rate.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

The pose-invariant framework is developed in Pytorch and trained on NVIDIA A100 GPUs. The speakers present in OuluVS2 and LRS3-TED corpora are split into train (70%), validation (10%) and test sets (20%) for speaker independent evaluation. Note that, the speakers are divided to ensure similar gender representation across all sets. The WHAM! noises are also separated into train, validation and test set to ensure noise independent settings. The model is trained with an Adam optimiser (lr=9e-4) for 50 epochs. The learning rate is multiplied by 0.8 when the model validation accuracy stops decreasing for 4 consecutive epochs. The model with best validation accuracy is used for evaluation.

Table 1. OuluVS2 2Mix Results

	PESQ	STOI	CSIG	CBAK	COVL
Noisy	2.191	0.537	2.951	2.062	2.302
ConvTasNet [15]	2.406	0.541	2.633	1.893	2.051
Proposed A-only	2.712	0.546	3.083	2.095	2.464
Proposed AV	3.082	0.578	3.675	2.220	2.964
Oracle IBM	2.690	0.620	2.373	1.747	1.929
Oracle IRM	3.308	0.640	4.241	2.238	3.446

Table 2. OuluVS2 2Mix+Noise Results

	PESQ	STOI	CSIG	CBAK	COVL
Noisy	2.191	0.537	2.951	2.062	2.302
ConvTasNet [15]	2.312	0.538	2.995	1.871	2.136
Proposed A-only	2.712	0.546	3.083	2.095	2.464
Proposed AV	3.082	0.578	3.675	2.220	2.964
Oracle IBM	2.690	0.620	2.373	1.747	1.929
Oracle IRM	3.308	0.640	4.241	2.238	3.446

Table 3. LRS3-TED 2Mix Results

	PESQ	STOI	CSIG	CBAK	COVL
Noisy	1.778	0.606	2.209	1.459	1.733
ConvTasNet [15]	2.472	0.628	2.615	1.966	2.235
Proposed A-only	2.669	0.651	3.099	1.916	2.417
Proposed AV	2.890	0.680	3.632	2.042	2.802
Oracle IBM	2.458	0.695	2.053	1.822	1.922
Oracle IRM	3.076	0.701	4.034	2.389	3.302

Table 4. LRS3-TED 2Mix+Noise Results

	PESQ	STOI	CSIG	CBAK	COVL
Noisy	1.401	0.479	1.601	1.739	1.292
ConvTasNet [15]	2.432	0.558	1.503	1.871	1.430
Proposed A-only	2.695	0.579	2.521	1.955	2.119
Proposed AV	2.871	0.589	3.438	1.949	2.643
Oracle IBM	2.238	0.590	1.616	1.506	1.383
Oracle IRM	2.822	0.601	3.402	1.840	2.650

4.2. Results

In order to evaluate the quality and intelligibility of the enhanced speech, five widely used objective evaluation metrics are used that aim to approximate the speech quality without conducting subjective listening tests, specifically: (1) PESQ [16]: one of the most widely used metrics to approx-

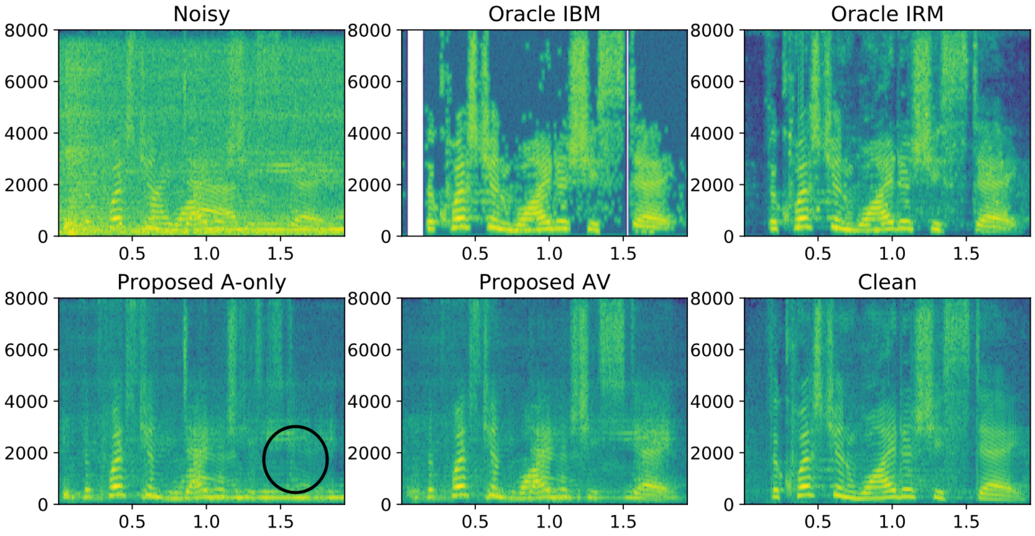

Fig. 3. Spectogram comparison for LRS3-TED : 2 Female + Noise mixture

imate subjective listening test score and ranges from [-0.5, 4.5] (2) STOI [17]: used to approximate speech intelligibility and ranges from [0,1] (3) CSIG [18]: used to predict speech distortion and ranges from [1, 5] (4) CBAK [18]: used to predict background distortion and ranges from [1, 5] (5) COVL [18]: predicts the overall quality and ranges from [1, 5].

The proposed AV model is compared with an A-only variant, Conv-TasNet [15] and ideal time frequency magnitude masks including ideal binary mask (IBM) and ideal ratio mask (IRM). The objective evaluation comparison of ConvTasNet, our proposed AV, A-only, IBM and IRM for OuluVS2 (2Mix), OuluVS2 (2Mix+Noise), LRS3-TED (2Mix) and LRS3-TED (2Mix+noise) are presented in Table 1, 2, 3, and 4 respectively. It can be seen that for all objective measures the proposed pose-invariant AV model significantly outperforms ConvTasNet, A-only model, and IBM across both corpora and experimental scenarios (i.e. 2Mix and 2Mix + Noise). In addition, the AV model achieves performance similar to oracle IRM.

Figure 3 depicts a spectogram of resynthesised speech generated using IBM, IRM, A-only model, and AV model for a randomly selected utterance from LRS3 (2Mix + Noise) test set. The spectogram is compared with both reference (clean) speech and degraded (noisy) speech. It can be seen that, both our proposed A-only and AV models achieve near ideal reconstruction. However, the A-only model failed to reconstruct the highlighted speech region.

4.3. Limitations

The main limitations of our proposed framework are outlined as follows (1) the 3D facial mesh features cannot be extracted for profile posture (i.e. ±90°) (2) the visual feature extraction model extrapolates the unseen part of the face to predict 3D landmark points and the accuracy of extrapolation decreases as the angle increases from 0° to 90° (a decreasing angle leads to enhanced performance). (3) The OuluVS2 corpus comprises fixed angle (0°, 30°, 45°, 60°, and 90°) facial postures.

5. CONCLUSIONS

This paper presented a novel pose invariant audio-visual (AV) speech enhancement (SE) and separation framework to address the challenging issue of pose variation in real-world unconstrained scenarios by exploiting privacy-preserving optical flow of 3D landmark features. The proposed framework is based on an innovative end-to-end deep neural architecture that unifies state-of-the-art transformers and temporal convolutional networks for pose-invariant SE. Comparative simulation results in terms of objective evaluation metrics (PESQ, STOI, CSIG, CBAK, COVL) revealed significant improvement of our proposed AV model compared to Conv-TasNet, oracle ideal binary mask, and A-only variant of the proposed model. Ongoing work includes subjective intelligibility evaluation of proposed framework with state-of-the-art AV SE models and more challenging real-world corpora along with a detailed theoretical, complexity and latency analysis. In future, we intend to explore sparse and generative adversarial networks for more robust and real-time pose-invariant AV speech separation.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[2] R Plomp, "Acoustical aspects of cocktail parties," *Acta Acustica united with Acustica*, vol. 38, no. 3, pp. 186–191, 1977.

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," *Proc. Interspeech 2019*, pp. 4295–4299, 2019.

[4] Mandar Gogate, Kia Dashtipour, Ahsan Adeel, and Amir Hussain, "Cochleanet: A robust language-independent audio-visual model for real-time speech enhancement," *Information Fusion*, vol. 63, pp. 273–285, 2020.

[5] Hyung Yong Kim, Ji Won Yoon, Sung Jun Cheon, Woo Hyun Kang, and Nam Soo Kim, "A multi-resolution approach to gan-based speech enhancement," *Applied Sciences*, vol. 11, no. 2, pp. 721, 2021.

[6] Ander Arriandiaga, Giovanni Morrone, Luca Pasa, Leonardo Badino, and Chiara Bartolozzi, "Audio-visual target speaker enhancement on multi-talker environment using event-driven cameras," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

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

[8] Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann, "Attention mesh: High-fidelity face mesh prediction in real-time," *arXiv preprint arXiv:2006.10962*, 2020.

[9] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[10] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira, "Perceiver IO: A general architecture for structured inputs & outputs," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[11] Iryna Anina, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen, "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2015, vol. 1, pp. 1–5.

[12] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[13] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. Interspeech*, Sept. 2019.

[14] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile gpus," *arXiv preprint arXiv:1907.05047*, 2019.

[15] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[16] 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.

[17] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[18] Yi Hu and Philipos C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.