AN END-TO-END SYSTEM FOR AUDIO-VISUAL SPEECH ENHANCEMENT FOR 1ST COG-MHEAR AVSE CHALLENGE

Mandar Gogate, Kia Dashtipour, Amir Hussain

Edinburgh Napier University, Edinburgh, UK

1. PROPOSED APPROACH

We proposes a novel end-to-end system for audio-visual speech enhancement (SE) based on deep neural networks as depicted in Fig. 1 for the 1st COG-MHEAR audio-visual speech enhancement challenge (AVSEC). The model exploits noisy speech, target speakers face and pose-invariant landmark flow features to estimate an ideal binary mask (IBM) [1] that selectively enhance target speech dominant regions and suppresses interfering background noises.

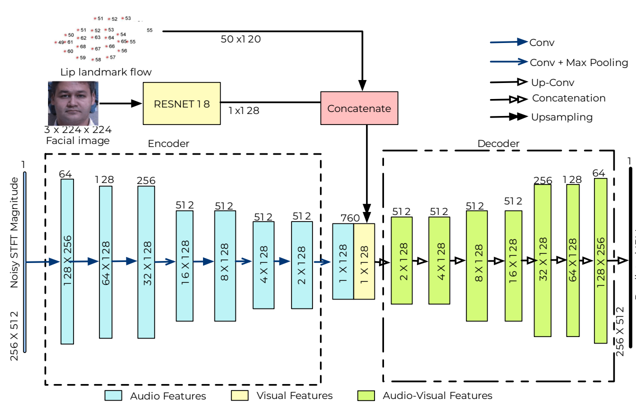

Fig. 1. Proposed Framework

Audio feature extraction: The audio feature extraction consists of a U-net encoder block [2]. The magnitude of noisy speech spectogram is fed as input to the network. The input is then fed to two convolutional layers with filter size of 4 and stride of 2 to downsample the time-frequency dimension until the time dimension is equal to 128. The downsampled features are passed through three convolutional blocks each consisting of two convolutional layers with filter size of 3 and stride of 1, followed by a frequency pooling layer that reduces the frequency dimension by 2.

Visual feature extraction:- The visual feature extraction consists of RESNET-18 to extract facial attribute features given a cropped face region. The extracted facial features are upsampled to match the video sampling rate. The upsampled facial attribute feature is combined with pose-invariant landmark flow features to generate final visual features.

Multimodal fusion: The upsampled visual features and audio features are concatenated and fed to a U-net decoder. The decoder consists of 3 up convolutional blocks each consisting of two upsampling layers that upsample the time dimension by 2, followed by convolutional layers with a filter size of 3 and stride of 1. The AV features are then fed to two transposed convolutional layers with filter size of 4 and stride of 2 to upsample the time-frequency dimension, until the time-frequency dimension is equal to the input. Next we use a sigmoid layer to map the output in the range of 0 to 1.

2. RESULTS

2.1. Objective evaluation

Table 1 demonstrated the overview results for objective evaluation on the dev set. It can be seen that for all objective measures the proposed AV outperforms baseline.

Table 1. Objective evaluation on dev set

	PESQ	STOI	SI-SDR
Noisy	1.154	0.639	-4.736
Baseline	1.271	0.678	0.577
Proposed AV	1.437	0.783	3.491
Oracle IBM	1.974	0.907	12.539

3. REFERENCES

[1] 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.

[2] Tassadaq Hussain, Mandar Gogate, Kia Dashtipour, and Amir Hussain, "Towards intelligibility-oriented audio-visual speech enhancement," *arXiv preprint arXiv:2111.09642*, 2021.