CogBiD System: Robust Audio-Visual Speech Enhancement with Intelligibility-oriented Loss Function

Tassadaq Hussain[1], *Mandar Gogate*[1], *Kia Dashtipour*[1], *Amir Hussain*[1]

[1]School of Computing, Edinburgh Napier University

{t.hussain, m.gogate, k.dashtipour, a.hussain}@napier.ac.uk

Abstract

In this work, we present the preliminary results of our single channel audio-visual speech enhancement (AVSE) model for the first AVSE challenge (AVSEC). Specifically, we used a deep network-based model that considers both auditory and visual information. The visual cues are utilized to enhance the quality and intelligibility of SE system and focus on the auditory information of specific speakers in a scene. We use the AVSEC dataset to train our combined AVSE model. This dataset includes spoken English sentences from TED and TEDx lectures and is made up of hundreds of hours of video. We show how our AVSE system can be used to address both conventional SE issues and real-world situations with speech- and noise-interfering background scenarios, including environments with non-stationary noise.

System Description

We assess the efficiency of deep learning (DL)-based audio-visual speech enhancement (AVSE) framework using short-time objective intelligibility (STOI) [1] loss functions in order to enhance the generalisation performance. Our objective is to examine the effects of incorporating visual information into a SE model, perform optimization using human perceptually-inspired loss functions, and assess the effectiveness and measurability of the augmented speech signal for AVSE challenge (AVSEC).

Short time Objective Intelligibility (STOI)

Here, we employ a modified version of a well-known STOI intelligibility measure as an objective function to train our AVSE model. The STOI is an intrusive measure that requires both estimated speech and reference (clean) speech signals and ranges from 0 to 1, with 1 denoting the highest intelligibility of the speech signal. The STOI function takes the clean and estimated speech signals as input and computes the score in five steps: i) Removal of silent regions from clean and estimated speech signals, ii) Application of the short-time Fourier transform (STFT); iii) Estimation of the short-time envelope of clean and noisy speech using one-third octave-band analysis of the STFT frames; iv) Normalization and clipping to compensate for global level differences and stabilisation of the STOI evaluation; and v) Intelligibility measure computation: the correlation coefficient between the two spectral envelopes is estimated using the equations below. More information about STOI as a loss function can be found in [2].

Proposed Audio-visual SE System

For AVSE challenge, we employed a recently proposed deep fully convolutional network (FCN)-based UNet architecture with various loss functions to optimize the AVSE framework. Figure 1 depicts the block diagram and Figure 2 presents the proposed DL-based UNet framework of AVSE architecture. We begin by considering an MSE loss function to train an FCN-based UNet model for SE. The MSE loss function can be calculated as follows between the noisy and estimated magnitude spectrum:

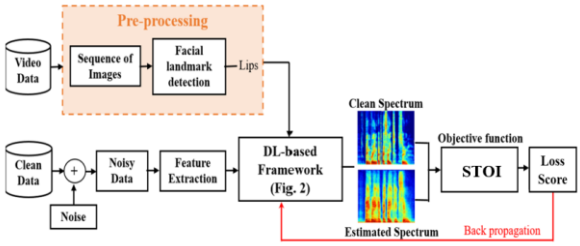

Fig. 1 : Block Diagram of AVSE Framework with STOI Loss Function.

$$L_{MSE} = min(\frac{1}{N}\sum_{n=1}^{N}\|\hat{Y}_n - Y_n\|_2)$$

where N signifies the number of speech frames, $\hat{Y}_n$ and Y_n are the estimated and clean magnitude spectra, and $\|.\|_2$ stands for L2-normalization.

Audio-visual Feature Extraction:

We modified an encoder-decoder style UNet architecture [3] to train AVSE models utilising three different loss functions. For audio processing, we first give the magnitude of $F \times T$ dimensional noisy STFT (F and T are the spectrogram's frequency and time dimensions) to the network as shown in Fig. 2.

The input is then sent through two convolutional layers with a filter size of 4 and a stride of 2, reducing the time-frequency dimensions to 64. Three convolutional blocks, each with three convolutional layers with a filter size of 3 and a stride of 1, process the downsampled features, followed by a frequency pooling layer to lower the frequency dimension by 2. It is worth mentioning that throughout convolutional block processing, the spatial dimension is preserved.

The visual processing pipeline comprised of a 3D convolutional layer for spatio-temporal convolution followed by 18-layer RESNET [4] (i.e., RESNET-18). The short-term dynamics of the lip articulations are extracted using the 3D layer. A convolutional layer with 64 3D kernels of size $5 \times 7 \times 7$ ($time \times width \times height$)) and a stride of $1 \times 2 \times 2$ makes up the 3D layer. The output of RESNET-18 is then input into a temporal convolutional network, as indicated in [4] (TCN). The network is given a temporal sequence of images of size $N \times 224 \times 224$, where N is the number of frames. The visual feature network outputs a 512-D vector for image

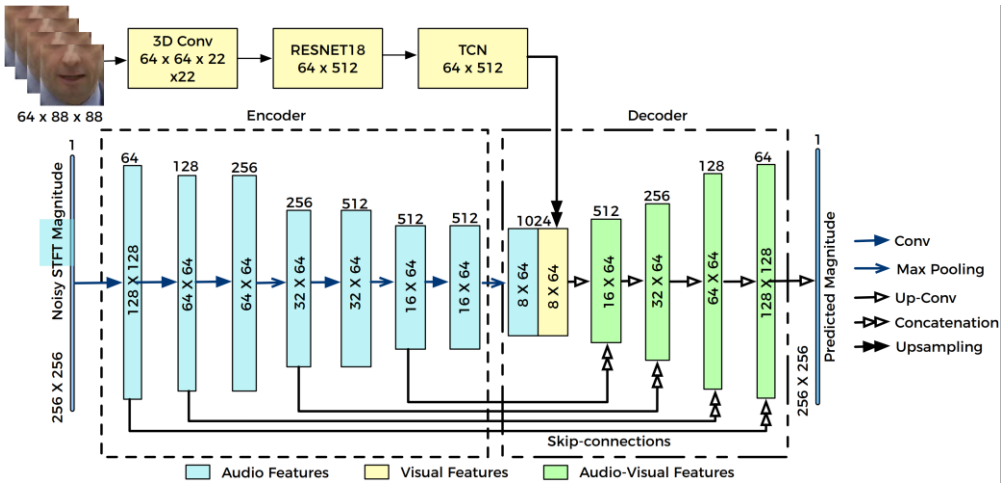

Fig. 2: FCN-based UNet for AVSEC

frames. The visual features are upsampled to match the audio features' sampling rate.

Multimodal fusion and Speech Resynthesis

As illustrated in Fig. 2, the upsampled visual and audio features are concatenated and sent to a UNet decoder. The decoder is made up of three up convolutional blocks, each of which is made up of two upsampling layers that upsample the time dimension by two, followed by convolutional layers with a filter size of 3 and a stride of 1. After that, the AV features are fed into two transposed convolutional layers with a filter size of 4 and a stride of 2 to upsample the time-frequency dimension until it equals the input. The output is then mapped in the range of 0 to 1 using a sigmoid layer. The anticipated mask is then multiplied by the input spectrogram to get the output masked spectrogram.

The noisy STFT features with clipped lip images are provided into the trained UNet framework to estimate the clean STFT features. The phase of the original noisy speech is used with inverse STFT to generate the enhanced speech signal.

References

[1] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. 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.

[2] T. Hussain, M. Gogate, K. Dashtipour, and A. Hussain, "Towards intelligibility-oriented audio-visual speech enhancement," arXiv preprint arXiv:2111.09642, 2021.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-NET: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, pp. 770–778, 2016.

[5] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in Proc. ICASSP, pp. 6319–6323, 2020.