ENHANCING SPEECH QUALITY AND INTELLIGIBILITY IN NOISY ENVIRONMENTS THROUGH EFFECTIVE FUSION OF AUDIO-VISUAL MODALITIES

Abstract:

This paper proposes an Audio-Visual Speech Enhancement (AVSE) model for enhancing speech signals in noisy audio-visual environments. The model utilizes dilated convolutions for audio feature extraction, capturing multi-scale contextual information and expanding the receptive field. The audio encoder generates a 1028-D vector for each Short-Time Fourier Transform (STFT) frame. For visual feature extraction, the model employs 3D convolutional layers with a filter size of 5 × 7 × 7 and a stride of 1 × 2 × 2, followed by a ResNet-18 backbone. The visual feature network outputs a 512-D vector for each face image. The final output is obtained by multiplying the masked magnitude with the noisy speech features, using the mean absolute error between the masked magnitude and clean magnitude IBM as the loss function during training.

1. **Proposed architecture:**

The proposed AVSE model aims to enhance speech signals in challenging audio-visual environments, an extension of AVSE challenge 1 [1]. It utilizes audio and visual cues to improve speech separation from background noise. The audio feature extraction module employs dilated convolutions for efficient and effective multi-scale information aggregation, and the visual feature extraction module employs 3D convolutional layers combined with a ResNet-18 backbone to capture meaningful visual features from face images.

Audio Feature Extraction: The audio feature extraction module comprises dilated convolutional layers, each followed by a ReLU activation function. The dilated convolutions help aggregate multi-scale contextual information, while maintaining coverage and resolution. The audio encoder generates a 1028-D vector for each STFT frame, which serves as input for the subsequent stages.

Visual Feature Extraction: The visual feature extraction module consists of 3D convolutional layers with a filter size of 5 × 7 × 7 and a stride of 1 × 2 × 2. These layers are followed by a ResNet-18 backbone, which further refines the visual features. The output is a 512-D vector for each face image, providing crucial visual information for the enhancement process.

Fusion and Separation: The AVSE model fuses the upsampled visual features with the audio features across the time dimension. The fused features are then processed by a LSTM layer with 257 units. Fully connected layers with 257 neurons and a sigmoid activation function are applied to the LSTM output. The weights of the fully connected layers are shared across the time dimension. The output is multiplied with noisy speech features to generate the masked magnitude.

Training and Optimization: The AVSE model is trained to minimize the mean absolute error using the Adam optimizer with a learning rate of 16e-3. The learning rate is reduced by a factor of 0.8 when the model validation loss stops decreasing for 2 consecutive epochs. The model with the best validation loss is selected for evaluation.

Results:

We have submitted five different results under a different name, ENU_JHU_1 to ENU_JHU_5, however, two of the models could not match the exact duration as the clear audio for testing purposes. Three model results are being displayed at the leaderboard; we have illustrated model descriptions of the model showing best results amongst all.

Table 1. Proposed models and baseline models

Model	STOI	PESQ	SISDR
Noisy	0.441041	1.136362	-5.07333
Base-model	0.556281	1.414132	3.667919
ResNet18	0.465583	1.199859	1.657748
ResNet50	0.402715	1.162721	-4.50824
VGG	0.396627	1.163339	-4.50174

Conclusion: The proposed AVSE model demonstrates promising results in enhancing speech signals in noisy audio-visual environments. The utilization of dilated convolutions and 3D convolutional layers with a ResNet-18 backbone enables effective multi-scale contextual information aggregation and meaningful visual feature extraction, respectively. The AVSE model showcases its potential in improving speech separation and enhancing speech quality.

Reference:

1. Blanco, A. L. A., Valentini-Botinhao, C., Klejch, O., Gogate, M., Dashtipour, K., Hussain, A., & Bell, P. (2023, January). AVSE Challenge: Audio-Visual Speech Enhancement Challenge. In 2022 IEEE Spoken Language Technology Workshop (SLT) (pp. 465-471). IEEE.