COGBID: AN ATTENTION BASED AUDIO-VISUAL SPEECH ENHANCEMENT SYSTEM FOR AVSEC2

Tassadaq Hussain, Kia Dashtipour, Mandar Gogate, Amir Hussain
Edinburgh Napier University

1. ABSTRACT

We propose an attention-based audio-visual speech enhancement system for the second Audio-Visual Speech Enhancement Challenge (AVSEC). The system comprises two parallel feature processing blocks: a UNet-based encoder with a parallel attention LSTM block and a UNet encoder. These components are designed to effectively capture spatial and temporal cues present in speech signals, aiming to enhance the intelligibility and quality of the provided audio-visual data by the AVSEC organizer. By integrating both audio and visual cues, our system gains a comprehensive understanding of the context and nuances in speech, leading to notable improvements in speech quality and intelligibility across diverse scenarios in the AVSEC challenge.

2. PROPOSED APPROACH

The proposed system is designed for the second Audio-Visual Speech Enhancement Challenge (AVSEC) with the aim to enhance intelligibility and quality of the provided audio-visual data. It introduces a parallel feature processing block: a UNet-based encoder with a parallel attention LSTM block (termed LCNN). This block effectively capture spatial and temporal cues in speech signals, improving audio-visual speech enhancement performance in various scenarios. The system utilizes both audio and visual cues to better understand the context and nuances of speech, leading to improved speech quality and intelligibility in AVSEC's provided audio-visual data. The UNet encoder processes audio spectrograms, while visual frames are processed separately using a ResNet-18 architecture for extracting facial expressions and other visual cues. Training the AVSE models involves employing an encoder-decoder style UNet architecture with different loss functions. The visual processing pipeline involves a 3D convolutional layer for spatio-temporal analysis and short-term dynamics extraction, followed by RESNET-18 for further visual feature extraction. The proposed system aims to provide robust and effective audio-visual speech enhancement for diverse scenarios in the AVSEC challenge.

1.1. Attention LSTM and Transformer Encoder

In LCNN, we apply attention to LSTM in parallel to a 2D-CNN, similar to the previous CNN framework. The attention LSTM processes input features through max-pooling and LSTM layers to extract relevant information. Attention weights are computed from the LSTM embedding and multiplied with them to focus on significant aspects of the data. The batch of data has a shape of (B, C, F, T), where B is the batch size, C represents the number of channels, F is the number of frequency bins, and T is the number of time steps. The data is passed through two separate frameworks: the first being a UNet encoder with 2D convolutional layers followed by batch normalization and max-pooling layers. The output of the UNet encoder is (B, 256, 1, 1). We flatten it to reshape it to (B, 256). On the other hand, the Attention LSTM takes an input of shape (B, C, F, T), applies max-pooling to obtain (B, 1, F/2, T/4), and further reshapes it to (B, F/2, T/4). This reshaped data is then passed to the LSTM layer, giving an output of shape (B, T/4, 256). Attention weights of shape (B, 1, T/4) are computed, flattened to (B, T/4), and multiplied with the LSTM output (B, T/4, 256) to get an output of shape (B, 256).

The outputs of LCNN (B, 256) and the UNet encoder (B, 256) are combined into an embedding of shape (B, 512). This concatenated embedding passes through a linear layer and softmax for leveraging both spatial and temporal features. Additionally, the transformer embedding and attention LSTM embedding are combined with the 2D-CNN's output to jointly learn from spatial and temporal information. By integrating the output of UNet encoder and LCNN, our proposed framework effectively allows the model to adapt to varying noisy environments. This robust architecture is poised to significantly improve speech intelligibility and quality in unseen noisy scenarios.

2.2. Multimodal fusion and speech resynthesis

To achieve multimodal fusion and speech resynthesis, the upsampled visual and joint audio features are concatenated and sent to the UNet decoder. The decoder component comprises 3 up convolutional blocks, each containing two upsampling layers that increase the time dimension by a factor of 2. Following this, convolutional layers with a filter size of 3 and a stride of 1 are applied. The AV features are then passed through two transposed convolutional layers with a filter size of 4 and a stride of 2, effectively upsampling the time-frequency dimension until it matches the input size. A sigmoid layer is used to map the output within the range of 0 to 1. The predicted mask is multiplied with the input spectrogram, resulting in the generation of the masked spectrogram as the output.