Watch or Listen: Robust Audio-Visual Speech Recognition with Visual Corruption Modeling and Reliability Scoring

Joanna Hong* Minsu Kim* Jeongsoo Choi Yong Man Ro†
Image and Video Systems Lab, KAIST
{jonna2587, ms.k, jeongsoo.choi, ymro}@kaist.ac.kr

Abstract

This paper deals with Audio-Visual Speech Recognition (AVSR) under multimodal input corruption situations where audio inputs and visual inputs are both corrupted, which is not well addressed in previous research directions. Previous studies have focused on how to complement the corrupted audio inputs with the clean visual inputs with the assumption of the availability of clean visual inputs. However, in real life, clean visual inputs are not always accessible and can even be corrupted by occluded lip regions or noises. Thus, we firstly analyze that the previous AVSR models are not indeed robust to the corruption of multimodal input streams, the audio and the visual inputs, compared to uni-modal models. Then, we design multimodal input corruption modeling to develop robust AVSR models. Lastly, we propose a novel AVSR framework, namely Audio-Visual Reliability Scoring module (AV-RelScore), that is robust to the corrupted multimodal inputs. The AV-RelScore can determine which input modal stream is reliable or not for the prediction and also can exploit the more reliable streams in prediction. The effectiveness of the proposed method is evaluated with comprehensive experiments on popular benchmark databases, LRS2 and LRS3. We also show that the reliability scores obtained by AV-RelScore well reflect the degree of corruption and make the proposed model focus on the reliable multimodal representations.

1. Introduction

Imagine you are watching the news on Youtube. Whether the recording microphone is a problem or the video encoding is wrong, the anchor’s voice keeps breaking off, so you cannot hear well. You try to understand her by her lip motions, but making matters worse, the microphone keeps covering her mouth, so the news is hardly recognizable. These days, people often face these kinds of situations, even in video conferences or interviews where the internet situation cuts in and out.

As understanding speech is the core part of human communication, there have been a number of works on speech recognition [1, 2], especially based on deep learning. These works have tried to enhance audio representation for recognizing speech in a noisy situation [3–6] or to utilize additional visual information for obtaining complementary effects [7–12]. Recently, even technologies that comprehend speech from only visual information have been developed [13–21].

With the research efforts, automatic speech recognition technologies including Audio Speech Recognition (ASR), Visual Speech Recognition (VSR), and Audio-Visual Speech Recognition (AVSR) are achieving great developments with outstanding performances [22–24]. With the advantages of utilizing multimodal inputs, audio and visual, AVSR that can robustly recognize speech even in a noisy environment, such as in a crowded restaurant, is rising for the future speech recognition technology. However, the previous studies have mostly considered the case where the audio inputs are corrupted and utilizing the additional clean visual inputs for complementing the corrupted audio information. Looking at the case, we come up with an important question, what if both visual and audio information are corrupted, even simultaneously? In real life, like the aforementioned news situation, cases where both visual and audio inputs are corrupted alternatively or even simultaneously, are frequently happening. To deal with the question, we firstly analyze the robustness of the previous ASR, VSR, and AVSR models on three different input corruption situations, 1) audio input corruption, 2) visual input corruption, and 3) audio-visual input corruption. Then, we show that the previous AVSR models are not indeed robust to audio-visual input corruption and show even worse performances than uni-modal models, which is eventually losing the benefit of utilizing multimodal inputs.

To maximize the superiority of using multimodal systems over the uni-modal system, in this paper, we propose a novel multimodal corruption modeling method and show its importance in developing robust AVSR technologies for

*Both authors have contributed equally to this work.
†Corresponding author
diverse input corruption situations including audio-visual corruption. To this end, we model the visual corruption with lip occlusion and noises that are composed of blurry frames and additive noise perturbation, along with the audio corruption modeling. Then, we propose a novel AVSR framework, namely Audio-Visual Reliability Scoring module (AV-RelScore), that can evaluate which modal of the current input representations is more reliable than others. The proposed AV-RelScore produces the reliability scores for each time step, which represent how much the current audio features and the visual features are helpful for recognizing speech. With the reliability scores, meaningful speech representations can be emphasized at each modal stream. Then, through multimodal attentive encoder, the emphasized multimodal representations are fused by considering inter-modal relationships. Therefore, with the AV-RelScore, the AVSR model can refer to the audio stream when the given visual stream is determined as less reliable (i.e., corrupted), and vice versa. We provide the audio-visual corruption modeling for the reproducibility and the future research.

Our key contributions are as follows:

• To the best of our knowledge, this is the first attempt to analyze the robustness of deep learning-based AVSR under the corruption of multimodal inputs including lip occlusions.
• We propose an audio-visual corruption modeling method and show that it is key for developing robust AVSR technologies under diverse environments.
• We propose Audio-Visual Reliability Scoring module (AV-RelScore) to figure out whether the current input modal is reliable or not, so that to robustly recognize the input speech even if one modality is corrupted, or even both.
• We conduct comprehensive experiments with ASR, VSR, and AVSR models to validate the effectiveness of the proposed audio-visual corruption modeling and AV-RelScore on LRS2 [25] and LRS3 [26], the largest audio-visual datasets obtained in the wild.

2. Related Works

2.1. Audio-visual speech recognition

There has been a great development in automatic speech recognition in both audio-based (ASR) and visual-based (VSR), along with the progress of deep learning.

Deep Neural Networks (DNNs) were embedded in the standard ASR pipeline [27–30]. Convolutional neural networks [31–33] and recurrent neural networks [34–36] brought large improvement in ASR performances. Connectionist Temporal Classification (CTC) [37] and Sequence-to-Sequence [38] learning methods were developed for end-to-end speech recognition. With advanced DNN architectures such as Transformer [9] and self-supervised/unsupervised learning methods, recent ASR methods [1, 11, 22, 39–41] achieved significant performances sufficient to be used in the real world. With the demands on recognizing speech even if there is no speech audio available, VSR has been developed mainly based on the developed technology in ASR. [10,23,25,42–49] narrowed the performance gap between ASR and VSR by proposing network architecture and learning methods specialized for VSR.

By combining the two technologies, Audio-Visual Speech Recognition (AVSR) is developed. Since AVSR is robust to acoustic noises by complementing the corrupted information with visual inputs, it is preferred under wild environments and real-world applications. With the pioneers of AVSR models [7,50–52], the research of AVSR has been placed in the mainstream of the deep multimodal field. Recent works have shown much greater improvements in AVSR [53, 54]: deep AVSR with large-scale datasets [8], a two-stage speech recognition model [55] that firstly enhances the speech with visual information then performs recognition, AVSR using Conformer architecture [24], and speech enhanced AVSR [12] were proposed. While there have been great developments in AVSR, there was a lack of consideration for visual impairment in AVSR. In this paper, we analyze that the previous deep learning-based AVSR models are weak for audio-visual input corruption and even show lower performance than ASR models. To
maximize the advantages of using multimodal inputs, we propose 1) audio-visual corruption modeling to train robust AVSR models and 2) Audio-Visual Reliability Scoring module (AV-RelScore) to effectively utilize a more reliable input stream while suppressing unreliable representations.

2.2. Visual occlusion modeling

In the real world, we can occasionally meet some objects that are occluded by other objects. This makes the trained DNNs perform worse when they were trained without consideration of the occluded situation. For example, there are image classification [56, 57] where the target object is overlapped by other objects, and facial expression recognition [58] where some facial parts are occluded by an object. To overcome this, many works tried to model the occlusion during training so that the trained network can robustly perform on its given task [59–62]. We try to model the lip occluded situation that frequently occurs when the speaker uses a mic or eats some food by using Naturalistic Occlusion Generation (NatOcc) of [61]. We show that visual corruption modeling is as important as audio corruption modeling, and is important to build a robust AVSR model.

3. Methodology

Let \( x_a \in \mathbb{R}^{T \times H \times W \times C} \) be a lip-centered talking face video with frame lengths of \( T \) and frame size of \( H \times W \times C \), and \( x_v \in \mathbb{R}^S \) be a paired speech audio with the input video, where \( S \) represents the length of audio. The objective of AVSR model is translating audio-visual inputs into ground-truth text, \( y \). Ideally, since the model utilizes two input modalities, it also can robustly perform the recognition when one modal input is corrupted with environmental noise by leaning on the other modality. However, previous AVSR models failed to consider the corruption of visual inputs and only considered acoustic corruption. In the following subsections, we firstly analyze the robustness of the previous AVSR models on diverse environments, and present a robust AVSR method for both acoustic and visual corruption.

3.1. Robustness of AVSR to acoustic and visual noise

In this subsection, we analyze the robustness of the previous AVSR models [12, 24], an ASR model [12], and a VSR model [24] on three different corrupted input situations, 1) audio input corruption, 2) visual input corruption, and 3) audio-visual input corruption, using LRS2 dataset. We directly utilize pre-trained models that do not consider the visual input corruption during training, in order to analyze their performances in different situations. For the audio input corruption, we injected a babble noise of [63] to the entire audio with different Signal-to-Noise Ratio (SNR) levels, -5 to 15dB, following [12, 24]. For the visual input corruption, there can be various noise types, additive noise, blur, color distortion, occlusion, etc. We investigate two different types of visual corruption, occlusion and noise (blur + additive noise), that we can frequently face in practice. For the audio-visual input corruption, both audio corruption and visual corruption are injected for random chunks of each stream so that both streams can be corrupted simultaneously or alternatively.

Audio input corruption. Since audio and visual inputs are highly correlated for the speech content instead of the background noise, AVSR models are expected to robustly recognize speech compared to the ASR model. As shown in Figure 3(a), the performance of ASR, Word Error Rate (WER), is steeply degraded when the noise level becomes higher (i.e., lower SNR). On the other hand, AVSR models, Conformer [24] and V-CAFE [12], show the robustness against acoustic noise by complementing the noisy audio signals with the visual inputs. Since the VSR model is not affected by the acoustic noise, it shows consistent performance for all SNR ranges. This result is in line with our expectations that AVSR models would be robust against acoustic noises.

Visual input corruption. Similar to audio input corruption, we perturb the visual inputs and examine performances of each model. Figure 3(b) shows the results under three different types of visual corruption, occlusion, noise, and both occlusion and noise. The performance of VSR model is largely affected by the visual corruption and nearly crushed when both occlusion and noises are applied to the input. The AVSR models also show degraded performances so
that Conformer loses 5.1% WER and V-CAFE loses 4.7% WER, from their original performances (i.e., clean situation), when both visual corruptions are applied. We found that the AVSR models tend to depend on the audio stream and are somewhat robust to visual input-only corruptions compared to VSR. Since ASR model is not affected by visual corruption and shows the best performance, the results indicate that it would be better to use the ASR model instead of the AVSR model when we know the inferring environment has clean audio and perturbed visual inputs.

**Audio-visual input corruption.** Finally, we model audio-visual input corruption where both audio and visual sequences are randomly corrupted, to confirm the robustness of previous models against multimodal corruption. The cases of alternative corruption of audio and visual sequences and simultaneous corruption are included. Ideally, when audio is corrupted, the model is expected to utilize the visual stream to produce appropriate results and vice versa. The results on different audio SNR levels with visual corruption using both occlusion and noise are illustrated in Figure 3(c). In this situation, the previous AVSR models, Conformer and V-CAFE, show even worse performances than the audio-only model (i.e., ASR) on all SNR ranges. As the previous AVSR models are largely depending on the audio stream and they tend to refer to the visual stream when the audio input is corrupted, if both audio and visual inputs are corrupted, they fail in complementing from the multimodal information. Especially, V-CAFE that explicitly enhances the corrupted audio with visual inputs shows the worst performance than Conformer that doesn’t contain the audio enhancement part, under strong noise situations (-5 to 5dB SNR). This shows the risk when only considering the audio corruption during training and model designing; when there comes a visually perturbed input, the trained model’s performance can be degraded. Therefore, with the currently developed AVSR models, when there is audio-visual corruption, it would be better to use the audio-only model (i.e., ASR) instead of the multimodal model (i.e., AVSR).

From the examples of three input corruption cases, we show that even if we use two modal inputs, audio and video, in AVSR, AVSR model is not always robust to different input corruption cases. Therefore, to develop a robust AVSR method for maximizing the advantages of using multimodal inputs, we should model the visual input corruption case as well as the audio input corruption during training, and design the appropriate model architecture. In the following subsection, we introduce visual input corruption modeling for developing a robust AVSR model.

### 3.2. Visual corruption modeling

Acoustic noise modeling is well-known and generally utilized in both ASR and AVSR; it can be modeled by injecting various environmental noise data [63, 64] into clean audio inputs, rejecting some frequency ranges, and distorting the signals [65]. However, previous works have missed considering visual noise modeling, which is important for building robust AVSR models. We proposed to model the visual input corruption with occlusion patches and noises.

#### 3.2.1 Visual corruption with occlusion patch

The occlusion is frequently induced when a speaker gives a speech with a microphone or a script. The lips of the speaker can be repeatedly occluded by such objects, thus hard to recognize speech by solely watching the lip movements. To simulate the occlusion, we introduce attaching patches to the region of lips. We utilize Naturalistic Occlusion Generation (NatOcc) patches from high-quality synthetic face occlusion datasets of [61] that are originally designed for occlusion-aware face segmentation tasks. The NatOcc patches consist of various objects generally seen in our everyday lives, such as fruits, desserts, cups, and so on. Since it is designed for producing naturalistic occluded faces, it is appropriate for our lip occlusion modeling.

Given the input lip-centered talking face video \( x_v \), with length \( T \), we randomly choose, \( N \), representing how many times the occluded chunk occurs in whole sequences. We design that the patches are not overlapped throughout the entire input video. To do so, the input video frames \( x_v \) are evenly divided by occurrence number \( N \), called segment lip videos \( x_{v,n} = \{x_{v,n}\}_{n=1}^{N} \). Then, we select a random ratio \( t \) with the range in 0.3 to 0.5 of the segment of video frames, \( x_{v,n} \), where we are going to attach the patch. Then the patch is located on the random position of lip landmarks. We insert the occlusion patch with a probability of 0.8 and otherwise we use clean visual inputs. The examples of visual corruption with occlusion are shown in Figure 1.

#### 3.2.2 Visual corruption with noise

Furthermore, visual corruption can occur when there are actual noises in the input video sequences. These kinds of problems usually occur when the encoding process goes wrong, the camera is out of focus, or there are communication issues. To implement these noise-corrupted video situations, we adopt two types of noises: blur and additive noise. We randomly insert blur or Gaussian noise to input face video \( x_v \), with a probability of 0.3, respectively; otherwise, we utilize the clean sequence. For visual corruption with noise, we also follow the same scheme of selecting the random ratio \( t \) and occurrence number \( N \) applied in the visual corruption with occlusion patches. The examples of resulting noise corruption are indicated in Figure 2. By combining the two visual corruption, occlusion and noises, we can train an AVSR model with visual input corruption along with audio input corruption.

### 3.3. Audio-visual reliability scoring

In addition to the audio-visual corruption modeling, we propose a novel AVSR model that can robustly recognize
speech under diverse noise situations.

Humans can easily be aware of whether the given speech audio is noisy or clean just by hearing. Similarly, corrupted video such as occlusion or blur in the lip region is also easily detectable just by watching the video. In analogy to the human input systems, we propose Audio-Visual Reliability Scoring module (AV-RelScore) which can determine whether the input audio or video is corrupted or not, and minimize the effect of the corrupted stream in prediction. Therefore, the model can focus more on the other modal stream for speech modeling when one modal is determined as corrupted. Also, if it is determined as both modalities are corrupted, the model can pay attention more to capture context to infer the corrupted speech. The overall architecture of the proposed AVSR framework is illustrated in Figure 4.

Each modal feature, audio feature $f_a \in \mathbb{R}^{T \times D}$ and visual feature $f_v \in \mathbb{R}^{T \times D}$, is embedded through modality-specific front-ends, respectively. The audio front-end downsamples the time length of the input audio to have the same length as that of the visual feature (i.e., $T$). Then, to determine the corrupted frames, AV-RelScore is designed. Firstly, Audio-Stream Reliability Scoring module inspects the audio features by modeling its temporal information with temporal convolutions, and outputs audio reliability scores $s_a \in \mathbb{R}^{T \times D}$ with the value range of $[0, 1]$, where 0 indicates less reliability of audio representation in speech modeling while 1 indicates full reliability. With the obtained reliability scores that inform which audio features are less knowledgeable for the speech modeling, we can emphasize the more reliable representations through the emphasis function, $h(a, b) = a + a \odot b$, where $\odot$ represents Hadamard product:

$$\hat{f}_a = h(f_a, s_a), \quad (1)$$

where $\hat{f}_a$ represents the emphasized audio features using the reliability score. Similar to the audio stream, we can obtain emphasized visual features from Visual-Stream Reliability Scoring module as follows,

$$\hat{f}_v = h(f_v, s_v). \quad (2)$$

Therefore, the emphasized modality features, $\hat{f}_a$ and $\hat{f}_v$, respectively contain speech information of reliable frames while the corrupted representations are suppressed.

When frames of one modal features are determined as corrupted, it is important to refer to other modal features which possibly not corrupted. To this end, we encode the emphasized multimodal features, $\hat{f}_a$ and $\hat{f}_v$, using multimodal attentive encoder, so the inter-modal relationships can be considered. Inspired by [66], we concatenate multimodal features into the temporal dimension to create the combined emphasized modality features $f_{av} = [\hat{f}_a, \hat{f}_v] \in \mathbb{R}^{2T \times D}$. Then, $f_{av}$ is fed into attention-based network, Conformer [11], to produce $\hat{f}_{av} \in \mathbb{R}^{T \times D}$.

By using Conformer for the multimodal attentive encoder, the network can attend across modalities so that the reliable modality can be utilized for modeling speech information. Moreover, both intra-modal relationships and inter-modal relationships can be captured through local convolution and global attention. Finally, the first $T$ output features from the multimodal attentive encoder are utilized to predict the sentence via Transformer decoder [9]. The visualization of the proposed AV-RelScore is shown in Figure 5.

### 3.3.1 Objective functions

The proposed AVSR framework is trained in an end-to-end manner with audio-visual input corruption. For the objective function, we utilize joint CTC/attention [67]. CTC
[37] loss is defined as $p_c(y|x) \approx \prod_{t=1}^{T} p(y_t|x)$ with an independent assumption of each output, and attention-based loss is defined as $p_a(y|x) = \prod_{j=1}^{J} p(y_j|y_{<j},x)$ that the current prediction is determined by previous predictions and inputs, thus including the learning of internal language model, where $J$ represents the total length of ground-truth text. Then, the total objective can be written as follows, \[ \mathcal{L} = \lambda \log p_c(y|x) + (1-\lambda) \log p_a(y|x), \] where $\lambda$ is a weight parameter for balancing two loss terms.

4. Experimental setup

4.1. Dataset

LRS2 [25] is an English sentence-level audio-visual dataset collected from BBC television shows. It has about 142,000 utterances including pre-train and train sets, about 1,000 utterances for validation set, and about 1,200 utterances for test set. We utilize both sets for training, and test the model on the test set.

LRS3 [26] is another large-scale English sentence-level audio-visual dataset. It consists of about 150,000 videos which are a total of about 439 hours long and collected from TED. About 131,000 utterances are utilized for training, and about 1,300 utterances are used for testing.

4.2. Architecture details

We adopt the visual front-end and the audio front-end from [24]. The visual front-end module is comprised of a 3D convolutional layer with a kernel size of $5 \times 7 \times 7$ followed by a ResNet18 [68]. Then the output features are squeezed along the spatial dimension by a global average pooling layer. The audio front-end module consists of a 1D convolutional layer with blocks of ResNet18. Both visual and audio front-ends are initialized using a pre-trained model on LRW [69].

For the multimodal attention with Conformer encoder [11], we use hidden dimensions of 256, feed-forward dimensions of 2048, 12 layers, 8 attention heads, and a convolution kernel size of 31. We utilize Transformer decoder [9] for prediction, and the decoder is composed of hidden dimensions of 256, feed-forward dimensions of 2048, 6 layers, and 8 attention heads.

4.3. Implementation details

For training and testing, every input frame is converted into grayscale. For data augmentation purposes, random cropping and horizontal flipping are applied to the visual inputs during training. For visual corruption modeling, occlusion is modeled with maximum occurrence number ($i.e.$, $N$) as 3, Gaussian blur with a kernel size of 7 and random sigma range of 0.1 to 2.0 is utilized for corruption using blur, and Gaussian noise with a maximum variance of 0.2 is utilized for additive noise corruption. For audio corruption modeling, we set the same setting as [24] that utilizes a babble noise of [70] with an SNR level from -5dB to 20dB. We adopt curriculum learning [71]. We initially train the model with the input videos that have lengths within 100 frames. Then, the model is again trained with those with lengths within 150 frames, 300 frames, and finally 600 frames length. We train 50 epochs for 100 and 150 frames, and 20 epochs for 300 and 600 frames. The whole network is trained with Adam optimizer [72] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. We utilize a learning rate scheduler and follow the same scheme as [46], where it increases linearly in the first 25,000 steps, yielding a peak learning rate of 0.0004 and thereafter decreasing in proportional to the inverse square root of the step number. We utilize 4 Nvidia RTX 3090 GPUs for training.

During decoding, we use beam search with a beam width of 40 and an external language model trained on LRS2 [25].
performs all other previous methods including ASR model. 

are also corrupted, the proposed model, A V-RelScore out-
importantly, when the SNR is lower (-5 to 5dB), meaning that 
is important in building robust AVSR models. More im-
confirm that the proposed audio-visual corruption modeling 
under strong acoustic noise situations of -5, 0, and 5dB 
models achieve better performances than the ASR model 
compared to the results with the same experimental set-
trials, which are presented in Table 1. Table 1. WER (%) comparisons with the state-of-the-art methods on audio-visual corrupted environment. The first row represents the types of visual corruption: patch occlusion, noise, and occlusion+noise, and the second row indicates audio noise with different levels, SNR(dB).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Input Modal</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRS2</td>
<td>V</td>
<td>A + V</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AVSR</td>
</tr>
<tr>
<td>LRS3</td>
<td>V</td>
<td>A + V</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AVSR</td>
</tr>
</tbody>
</table>

5. Experimental results

5.1. Robustness to audio-visual corruption

To begin with, we compare the proposed AVSR framework with the state-of-the-art methods including ASR and VSR on LRS2 and LRS3 datasets. We report performances under different corruption environments with three types of visual corruption modeling: occlusion, noise, and occlusion+noise, and audio corruption modeling: an SNR range from -5 to 15dB and a clean-audio environment. Table 1 shows the comparison results. The results emphasize the importance of audio-visual corruption modeling. All AVSR models achieve better performances than the ASR model under strong acoustic noise situations of -5, 0, and 5dB SNR even if there is strong visual corruption, the combination of occlusion and noises. By comparing the results with that shown in Figure 3(c) that the ASR model always yields the best performance under audio-visual corruption, we can confirm that the proposed audio-visual corruption modeling is important in building robust AVSR models. More importantly, when the SNR is lower (-5 to 5dB), meaning that the audio corruption is much applied, and the visual inputs are also corrupted, the proposed model, AV-RelScore outperforms all other previous methods including ASR model.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Input Modal</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRS2</td>
<td>V</td>
<td>A + V</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AVSR</td>
</tr>
<tr>
<td>LRS3</td>
<td>V</td>
<td>A + V</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AVSR</td>
</tr>
</tbody>
</table>

Table 2. Ablation study on LRS2 dataset.

This clearly verifies that our proposed AV-RelScore module effectively finds and utilizes the more reliable modal for recognizing speech.

In addition, we also compare the performances of each method in an audio-only corrupted environment which is the standard setting of previous methods [12, 24] so that all audio sequences are corrupted with babble noise, instead of corrupting random chunks. Figure 6 shows the comparison results using LRS2 dataset. For this case, AVSR models show the best robustness compared to ASR model. Moreover, the proposed AV-RelScore outperforms the other methods in severe noise conditions.

5.2. Visualization of reliability score

In this section, we analyze the effectiveness of our proposed AV-RelScore module by visualizing the reliability scores of audio and visual modality, shown in Figure 7. The visual reliability scores are represented with red bars, and the audio reliability scores are indicated as green bars. From the visual and audio reliability scores, we can clearly observe that the more highly corrupted visual inputs are, the fewer reliability scores are obtained. In addition, if the occlusion patch directly covers the lip movements, shown in the 3rd video segment of the first example, the visual reliability score is much less than that with the occlusion patch located slightly left to the lip, shown in the 1st video segment of the first example. If the occlusion patch is not covering the lip movements, the visual reliability scores are high enough to recognize the speech properly, shown in the 1st video segment of the first example.
Figure 7. Visualization of visual reliability scores and audio reliability scores from AV-RelScore module of LRS2 dataset.

Table 3. WER (%) comparisons with state-of-the-art methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>LRS2</th>
<th>LRS3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM-Seq2Seq [8]</td>
<td>8.5</td>
<td>7.2</td>
</tr>
<tr>
<td>CTC/Attention [76]</td>
<td>7.0</td>
<td>-</td>
</tr>
<tr>
<td>LF-MMI TDNN [77]</td>
<td>5.9</td>
<td>-</td>
</tr>
<tr>
<td>EG-Seq2Seq [55]</td>
<td>-</td>
<td>6.8</td>
</tr>
<tr>
<td>RNN-T [78]</td>
<td>-</td>
<td>4.5</td>
</tr>
<tr>
<td>Conformer [24]</td>
<td>4.7</td>
<td>3.2</td>
</tr>
<tr>
<td>V-CAFE [12]</td>
<td>4.5</td>
<td>3.4</td>
</tr>
<tr>
<td>AV-RelScore</td>
<td>4.1</td>
<td>2.8</td>
</tr>
</tbody>
</table>

1st video segment of the second example. The scores are gradually increasing when the visual corruption is getting eliminated, indicated in the 2nd video segment of the second example. We also notice that each visual reliability and audio reliability score properly supplement each other when there is one modal stream corruption, so when the visual reliability scores are low, the audio reliability scores are high enough to help recognize the speech.

5.3. Ablation study

We conduct an ablation study to confirm the effect of the proposed architecture on LRS2. We evaluate the performances in audio-visual corrupted dataset (audio: -5dB SNR, visual: occlusion+noise). To this end, we examine the performances of different variants of the proposed model. By setting the baseline network as [24] which does not contain both reliability scoring and multimodal attention, we firstly add multimodal attention from the baseline architecture, and the performance improves to 13.70% WER from 20.15% WER, indicated in the second row of Table 2. The results show that it is beneficial to use cross-modal attention so that the model can refer to other modalities when the given modality is less informative, compared to the method of independent modeling and fusing the two representations with a linear layer. Then, we add the reliability scoring which is the final proposed model. The performance attains 13.36% WER. Since it is important to evaluate the significance of each visual frame and audio sequence, especially when both are corrupted, the AV-RelScore is necessary for achieving better performance. From the ablation study, it is clearly implied that it is important to not only refer to other modal features when one modal feature is corrupted but also consider essential parts of the corrupted audio-visual inputs, through both Reliability scoring.

5.4. Comparisons with audio-visual corruption free

We also verify the proposed AV-RelScore by comparing with the previous state-of-the-art methods [8, 12, 24, 55, 76–78] in audio-visual clean environment, shown in Table 3. The results indicate that the proposed AVSR framework outperforms the state-of-the-art methods even in a clean environment, by achieving 4.1% WER and 2.8% WER on the LRS2 and LRS3 datasets, respectively.

6. Conclusion

In this paper, we propose audio-visual corruption modeling for AVSR and show its importance in boosting the robustness of AVSR systems. Moreover, we propose AV-RelScore that determines which modal input stream is more meaningful than others and emphasizes the meaningful representations. The effectiveness of the proposed audio-visual corruption modeling and AV-RelScore is verified through comprehensive analysis and experiments on two large-scale audio-visual databases, LRS2 and LRS3.

Acknowledgement This work was partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.NRF-2022R1A2C2005529), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities).
References


[52] Darryl Stewart, Rowan Seymour, Adrian Pass, and Ji Ming. Robust audio-visual speech recognition under noisy
audio-video conditions. *IEEE transactions on cybernetics*, 44(2):175–184, 2013. 2


[77] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE In-