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
{joanna2587, 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.[1]

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

[1] https://github.com/joannahong/AV-RelScore

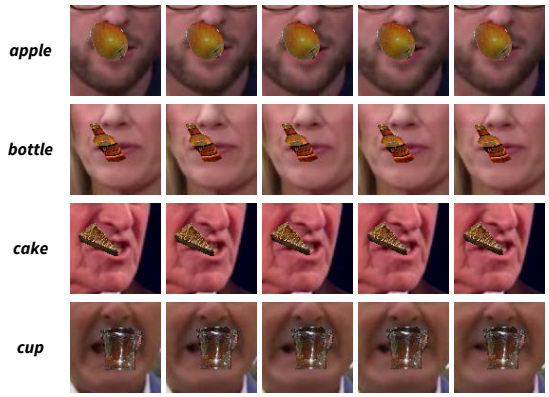

Figure 1. Examples of visual occlusion with NatOcc patches.

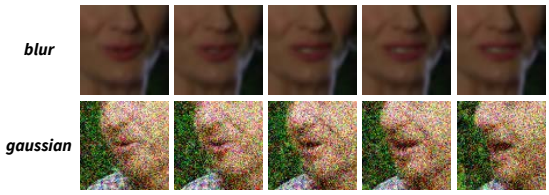

Figure 2. Examples of visual corruption with noises.

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

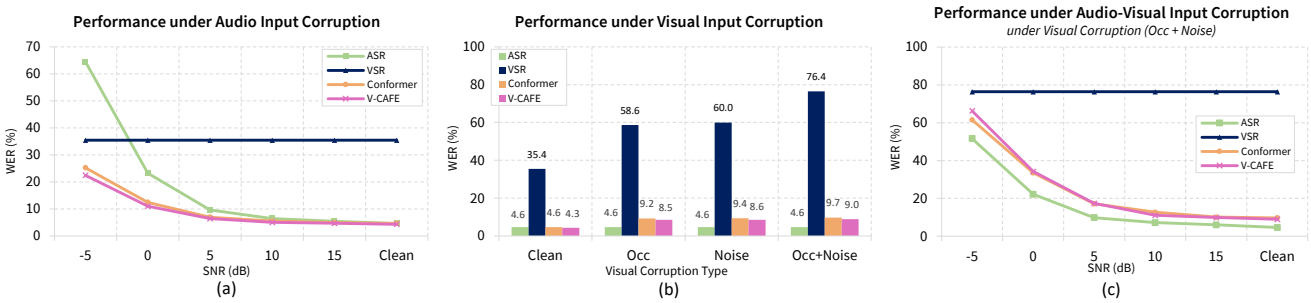

Figure 3. Speech recognition performances of ASR, VSR, and AVSR models on LRS2 dataset under different input corruption types: (a) Audio input corruption with babble noise. (b) Visual input corruption with occlusion and noise. (c) Audio-visual input corruption.

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 $\boldsymbol{x}_v \in \mathbb{R}^{T\times H\times W\times C}$ be a lip-centered talking face video with frame lengths of T and frame size of $H\times W\times C$, and $\boldsymbol{x}_a \in \mathbb{R}^{S}$ 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 $\boldsymbol{x}_v$ 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 $\boldsymbol{x}_v$ are evenly divided by occurrence number N, called segment lip videos $\boldsymbol{x}_v = \{x_{v,n}\}_{n=1}^{N}$. 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 $\boldsymbol{x}_v$ 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

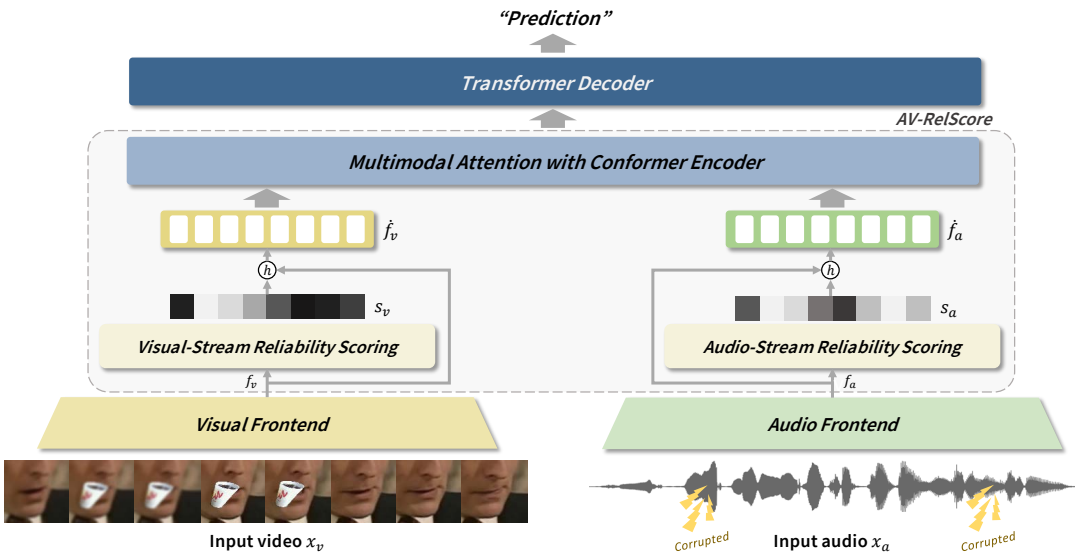

Figure 4. Overall architecture of the proposed AVSR framework.

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:

$$\dot{f}_a = h(f_a, s_a), \tag{1}$$

where $\dot{f}_a$ 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,

$$\dot{f}_v = h(f_v, s_v). \tag{2}$$

Therefore, the emphasized modality features, $\dot{f}_a$ and $\dot{f}_v$, 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, $\dot{f}_a$ and $\dot{f}_v$, 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 $\dot{f}_{av} = [\dot{f}_a, \dot{f}_v] \in \mathbb{R}^{2T\times D}$. Then, $\dot{f}_{av}$ 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

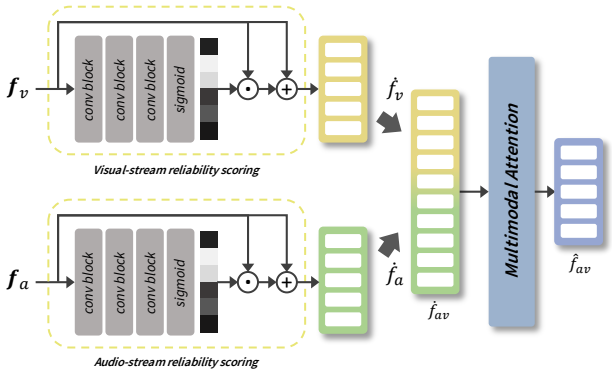

Figure 5. Detailed architecture of AV-RelScore.

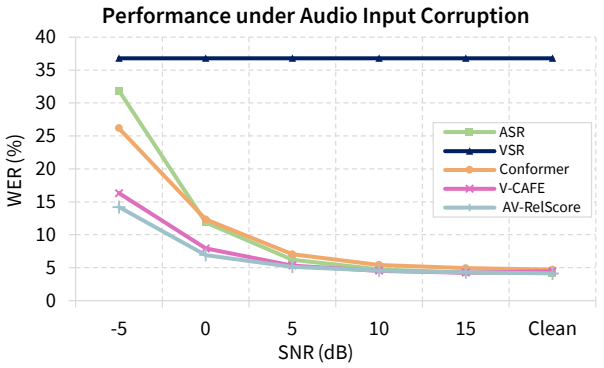

Figure 6. Speech recognition performances of ASR, VSR, and the proposed model under audio input-only corruption using LRS2 dataset. Note the models are trained with audio-visual corruption modeling.

[37] loss is defined as $p_c(y|x) \approx \Pi_{t=1}^{T} p(y_t|x)$ with an independent assumption of each output, and attention-based loss is defined as $p_a(y|x) = \Pi_{j=1}^{J} p(y_j|y_{<j}, x)$ 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, $\mathcal{L} = \lambda \log p_a(y|x) + (1-\lambda) \log p_c(y|x)$, where λ 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.

For the Audio-Stream Reliability Scoring module and the Visual-Stream Reliability Scoring module, we exploit three 1D convolution layers where each layer is followed by batch normalization and ReLU activation function. The output features are taken into sigmoid activation for obtaining scores in the range of $[0, 1]$.

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],

Dataset	Input Modal	Method	Occlusion						Noise						Occlusion+Noise					
			clean	15	10	5	0	-5	clean	15	10	5	0	-5	clean	15	10	5	0	-5
LRS2	A	ASR [24]	4.17	4.37	4.76	5.57	8.26	17.58	**4.17**	**4.37**	4.76	5.57	8.26	17.58	**4.17**	4.37	4.76	5.57	8.26	17.58
	V	VSR [24]	48.11	48.11	48.11	48.11	48.11	48.11	60.11	60.11	60.11	60.11	60.11	60.11	69.61	69.61	69.61	69.61	69.61	69.61
	A + V	Conformer [24]	4.91	5.17	5.36	6.51	9.85	17.30	4.84	5.06	5.33	6.40	9.67	17.66	5.03	5.32	5.63	6.56	10.41	20.15
	A + V	V-CAFE [12]	4.57	4.44	4.84	5.50	7.34	12.15	4.87	4.64	4.93	5.44	7.12	**11.62**	4.69	4.55	5.06	5.66	7.78	14.07
	A + V	**AV-RelScore**	**4.16**	**4.34**	**4.37**	**5.21**	**6.38**	**11.32**	4.54	4.42	**4.45**	**5.24**	**6.31**	11.79	4.25	**4.35**	**4.49**	**5.45**	**6.95**	**13.36**
LRS3	A	ASR [24]	**2.53**	**2.68**	2.97	3.53	5.64	12.95	**2.53**	**2.68**	2.97	3.53	5.64	12.95	**2.53**	**2.68**	**2.97**	3.53	5.64	12.95
	V	VSR [24]	56.45	56.45	56.45	56.45	56.45	56.45	61.45	61.45	61.45	61.45	61.45	61.45	71.52	71.52	71.52	71.52	71.52	71.52
	A + V	Conformer [24]	2.93	3.11	3.32	3.79	5.61	10.98	3.00	3.00	3.32	3.79	5.62	10.62	3.03	3.03	3.33	3.85	5.64	11.82
	A + V	V-CAFE [12]	3.39	3.38	3.46	3.84	5.34	9.00	3.49	3.48	3.63	3.83	5.31	8.69	3.67	3.37	3.69	4.17	5.70	10.04
	A + V	**AV-RelScore**	2.91	2.83	**2.89**	**3.25**	**4.81**	**8.70**	3.05	2.89	**2.92**	**3.31**	**4.61**	**8.51**	2.95	2.91	3.10	**3.34**	**5.11**	**9.41**

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 both, and the second row indicates audio noise with different levels, SNR(dB).

LRS3 [26], LibriSpeech [73], Voxforge, TED-LIUM 3 [74], and Common Voice [75], following [46]. Therefore, the decoding procedure can be written as follows,

$$p(y|x) = \alpha \log p_a(y|x) + (1-\alpha) \log p_c(y|x) + \beta \log p_{lm}(y), \quad (3)$$

where p_{lm} is the decoding score from the external language model, and α and β are the weight parameters. We use 0.9 for λ in the training objective function, 0.9 and 0.5 for α and β in LRS2, respectively, and 0.9 and 0.6 for α and β in LRS3, respectively. Note that we re-implement all the previous methods and reproduce the results with the same experimental settings as ours for fair comparisons.

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.

Proposed Method			WER(%)
Baseline	Multimodal attention	Reliability scoring	
✓	✗	✗	20.15
✓	✓	✗	13.70
✓	✓	✓	**13.36**

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 3^{rd} 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 1^{st} 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

Figure 7. Visualization of visual reliability scores and audio reliability scores from AV-RelScore module of LRS2 dataset.

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

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

1^{st} video segment of the second example. The scores are gradually increasing when the visual corruption is getting eliminated, indicated in the 2^{nd} 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, 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

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016. 1, 2

[2] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016. 1

[3] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International conference on latent variable analysis and signal separation*, pages 91–99. Springer, 2015. 1

[4] Ke Tan and DeLiang Wang. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:380–390, 2019. 1

[5] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. Complex spectral mapping for single-and multi-channel speech enhancement and robust asr. *IEEE/ACM transactions on audio, speech, and language processing*, 28:1778–1787, 2020. 1

[6] Sebastian Braun, Hannes Gamper, Chandan KA Reddy, and Ivan Tashev. Towards efficient models for real-time deep noise suppression. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 656–660. IEEE, 2021. 1

[7] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42(4):722–737, 2015. 1, 2

[8] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1, 2, 8

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2, 5, 6

[10] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6548–6552. IEEE, 2018. 1, 2

[11] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020. 1, 2, 5, 6

[12] Joanna Hong, Minsu Kim, Daehun Yoo, and Yong Man Ro. Visual context-driven audio feature enhancement for robust end-to-end audio-visual speech recognition. *arXiv preprint arXiv:2207.06020*, 2022. 1, 2, 3, 7, 8

[13] Chenhao Wang. Multi-grained spatio-temporal modeling for lip-reading. *arXiv preprint arXiv:1908.11618*, 2019. 1

[14] Jingyun Xiao, Shuang Yang, Yuanhang Zhang, Shiguang Shan, and Xilin Chen. Deformation flow based two-stream network for lip reading. *arXiv preprint arXiv:2003.05709*, 2020. 1

[15] X. Zhao, S. Yang, S. Shan, and X. Chen. Mutual information maximization for effective lip reading. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 843–850, Los Alamitos, CA, USA, may 2020. IEEE Computer Society. 1

[16] Yuanhang Zhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen. Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. *arXiv preprint arXiv:2003.03206*, 2020. 1

[17] Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. Multi-modality associative bridging through memory: Speech sound recollected from face video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 296–306, October 2021. 1

[18] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip to speech synthesis with visual context attentional gan. *Advances in Neural Information Processing Systems*, 34:2758–2770, 2021. 1

[19] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip-to-speech synthesis in the wild with multi-task learning. *arXiv preprint arXiv:2302.08841*, 2023. 1

[20] Joanna Hong, Minsu Kim, Se Jin Park, and Yong Man Ro. Speech reconstruction with reminiscent sound via visual voice memory. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3654–3667, 2021. 1

[21] Joanna Hong, Minsu Kim, and Yong Man Ro. Visagesyntalk: Unseen speaker video-to-speech synthesis via speech-visage feature selection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 452–468. Springer, 2022. 1

[22] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020. 1, 2

[23] Minsu Kim, Jeong Hun Yeo, and Yong Man Ro. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada*, volume 22, 2022. 1, 2

[24] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021. 1, 2, 3, 6, 7, 8

[25] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE, 2017. 2, 6

[26] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 2, 6, 7

[27] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20(1):14–22, 2011. 2

[28] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. 2

[29] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2011. 2

[30] Frank Seide, Gang Li, and Dong Yu. Conversational speech transcription using context-dependent deep neural networks. In *Twelfth annual conference of the international speech communication association*, 2011. 2

[31] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*, pages 4277–4280. IEEE, 2012. 2

[32] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014. 2

[33] Tara N Sainath, Brian Kingsbury, Abdel-rahman Mohamed, George E Dahl, George Saon, Hagen Soltau, Tomas Beran, Aleksandr Y Aravkin, and Bhuvana Ramabhadran. Improvements to deep convolutional neural networks for lvcsr. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 315–320. IEEE, 2013. 2

[34] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013. 2

[35] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR, 2014. 2

[36] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014. 2

[37] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006. 2, 6

[38] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 2

[39] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 2

[40] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019. 2

[41] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019. 2

[42] Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. Cromm-vsr: Cross-modal memory augmented visual speech recognition. *IEEE Transactions on Multimedia*, pages 1–1, 2021. 2

[43] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE, 2020. 2

[44] Xingxuan Zhang, Feng Cheng, and Shilin Wang. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 713–722, 2019. 2

[45] Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. Hearing lips: Improving lip reading by distilling speech recognizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6917–6924, 2020. 2

[46] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *arXiv preprint arXiv:2202.13084*, 2022. 2, 6, 7

[47] Minsu Kim, Hyunjun Kim, and Yong Man Ro. Speaker-adaptive lip reading with user-dependent padding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 576–593. Springer, 2022. 2

[48] Minsu Kim, Hyung-Il Kim, and Yong Man Ro. Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition. *arXiv preprint arXiv:2302.08102*, 2023. 2

[49] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5162–5172, 2022. 2

[50] Jing Huang and Brian Kingsbury. Audio-visual deep learning for noise robust speech recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 7596–7599. IEEE, 2013. 2

[51] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE, 2015. 2

[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

[53] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6984–6988. IEEE, 2020. 2

[54] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE, 2018. 2

[55] Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang. Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14433–14442, 2020. 2, 8

[56] Feng Cen, Xiaoyu Zhao, Wuzhuang Li, and Guanghui Wang. Deep feature augmentation for occluded image classification. *Pattern Recognition*, 111:107737, 2021. 3

[57] Yongjun Li, Shasha Li, Haohao Du, Lijia Chen, Dongming Zhang, and Yao Li. Yolo-acn: Focusing on small target and occluded object detection. *IEEE Access*, 8:227288–227303, 2020. 3

[58] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018. 3

[59] Feng Cen, Xiaoyu Zhao, Wuzhuang Li, and Fanglai Zhu. Classification of occluded images for large-scale datasets with numerous occlusion patterns. *IEEE Access*, 8:170883–170897, 2020. 3

[60] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Patch-gated cnn for occlusion-aware facial expression recognition. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2209–2214. IEEE, 2018. 3

[61] Kenny TR Voo, Liming Jiang, and Chen Change Loy. Delving into high-quality synthetic face occlusion segmentation datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4711–4720, 2022. 3, 4

[62] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12645–12654, 2020. 3

[63] Andrew Varga and Herman JM Steeneken. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*, 12(3):247–251, 1993. 3, 4

[64] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 035081. Acoustical Society of America, 2013. 4

[65] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Sixteenth annual conference of the international speech communication association*, 2015. 4

[66] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 5

[67] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE, 2017. 5

[68] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[69] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016. 6

[70] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE, 2020. 6

[71] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 6

[72] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[73] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 7

[74] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International conference on speech and computer*, pages 198–208. Springer, 2018. 7

[75] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019. 7

[76] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE, 2018. 8

[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-*

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6984–6988. IEEE, 2020. 8

[78] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912. IEEE, 2019. 8