A Study of Dropout-Induced Modality Bias on Robustness to Missing Video Frames for Audio-Visual Speech Recognition

Supplementary Material

10. Additional Experiments

Analysis of Latent Space Distribution Samples We further analyze the latent space distribution samples of the proposed robust AVSR model and achieve following two conclusions: (1) In Figure 7, we observe that MDA-KD effectively avoids dropout-induced bias and make sure the model to employ a collaborative decision strategy, even with video frames missing input. (2) In Figure 8, we demonstrate that the model decision-making pattern indeed dynamically switches to an audio-dominant one by activating the MS-Adapter when facing complete video missing input.

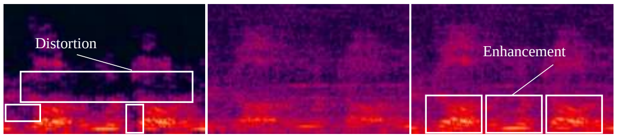
Analysis of Zero-shot Noise Robustness We further evaluate the system performance with zero-shot noise. Specifically, we simulate the noise speech with unseen Babble noise from NOISEX [56] and the near-field audio captured by a head-worn microphone at 0 dB, -2.5 dB, and -5.0 dB SNR levels. In Table 3, we reuse the symbols from Table 3. The results demonstrate that the proposed modality-unbiased model, AV6, outperforms both the modality-biased model AV1 and the unimodal model A0 in both Near Field and Far Field settings with in-set noise. More importantly, we highlight the advantage of zero-shot noise robustness of the proposed method across all SNR levels, aligning with the target of AVSR as a robust system for real-world applications.

Analysis of Computational Consumption We analyze the computational efficiency in FLOPS with audio-only input to demonstrate the effectiveness of reducing computation by activating the MS-Adapter and interrupting the data flow in the video branch. Upon activating the MS-Adapter, data solely flows through the audio branch, requiring only 3.89 GFLOPS with 94.21 M parameters for computation. This contrasts favorably with conventional methods that necessitate padding video tensor inputs, consuming 12.64 GFLOPS with totaling 144.78 M parameters.

Experiment Details on Different Test Dropout Methods In Figure 9, we provide more comprehensive experimental results and present performance degradation curves across all three test suites (Segment Dropout, Utterance Dropout, and Interval Dropout) to facilitate further research.

Models	Near Field	Far Field	Zero-shot Babble Noise		
			0dB	-2.5dB	-5dB
A0	18.10	25.13	33.52	62.17	75.76
AV1	17.71	23.26	29.40	51.63	63.80
AV6	**16.86**	**21.11**	**26.67**	**44.97**	**55.65**

Table 5. CER comparison of zero-shot noise roubustness.



CPU-GSS Enhanced Original GPU-GSS Enhanced

Figure 6. Spectral analysis of GSS-enhanced signals

11. Distinctions between MBVD and MVD

There are three key distinctions between the Modality Bias Venn Diagram (MBVD) and the Modality Venn Diagram (MVD) [28]. Firstly, MBVD focuses on the latent space to describe the decision pattern of a multimodal model, while MVD space is essentially another form of the original feature space. Secondly, for the generation order, MBVD maps from the original feature space $\mathcal{X}$ to the decisive feature space $\mathcal{Z}$, while MVD follows the opposite direction. Lastly, MBVD is employed to describe modality bias in decision-making processes, whereas MVD is utilized for knowledge distillation.

12. Limitations of the work

Modality dropout presents two facets. On one hand, it could address the out-of-distribution (OOD) issue by missing modalities. On the other hand, if applied on supplementary modalities, it can induce dropout-induced modality bias in modality-biased systems. For our further exploration, we realize the manifestation of these characteristics is related to input quality. In this work, we focus on real-world TV room scenarios with relatively low-resolution video and noisy speech. Under such conditions, dropout-induced modality bias is observed prominently. While for high-quality datasets, such as LRS2 and LRS3, dropout serves more as a form of data augmentation, and the dropout-induced modality bias are mitigated by high-quality input. Nevertheless, in all conditions, the proposed MDA-KD and MS-Adapter consistently lead to relative improvements to original dropout method.

13. Implement Details

Data Processing Details We apply conventional signal processing algorithms, such as Weighted Prediction Error (WPE) [57] and Guided Source Separation (GSS) [58], to

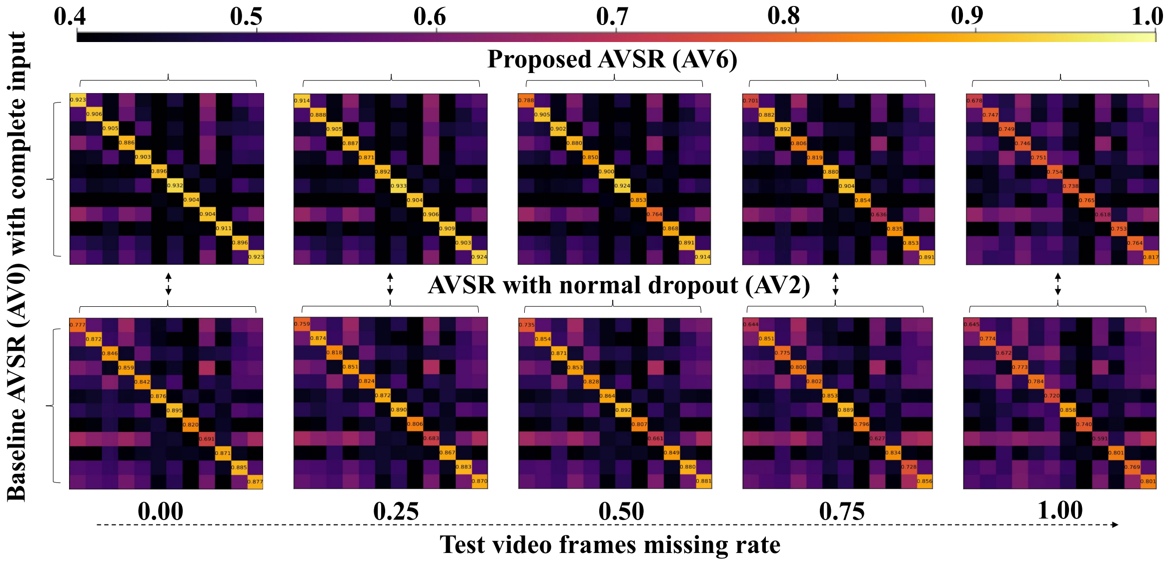


Figure 7. We investigate the decision discrepancies between the proposed robust AVSR (AV6) and the AVSR trained using the normal dropout technique (AV2) across different test video frames missing rates. Similar to Figure 3, we quantify the divergence by calculating the cosine distance similarity of latent decision distribution samples from both models with missing video frames input and those of AV0 with complete data input. The latter samples represent an ideal collaborative decision strategy. Each diagonal element in the cosine distance-based similarity matrix represents the similarity between intermediate representations with the same sample index but may have different missing rates. As a result, two prominent phenomena emerge. (1) In vertical comparison between AV6 and AV2, the sample similarities of AV6 consistently surpass those of AV2 along the diagonal line, indicating a closer approximation to the ideal collaborative decision distribution in latent space. These results suggest that MDA-KD enables AV6 to adopt a decision strategy similar to AV0, whether facing complete input or missing video frames, effectively utilizing content information and modality general information audio modality. (2) In horizontal comparison, in the first row, the diagonal elements in each subplot consistently darken as the missing rate increases, and the last subplot darkens sharply with the shift of decisive bias on audio modality upon activating the MS-Adapter. This trend is less pronounced in the second row, as AV2 exhibits an excessive modality bias on audio modality, deviating from the collaborative decision strategy.

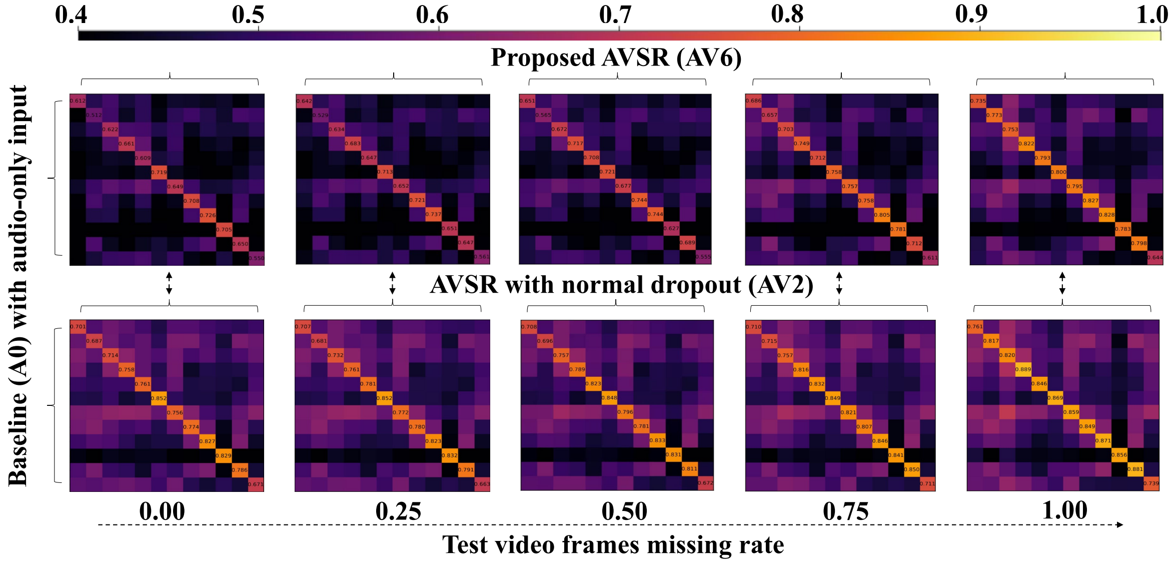


Figure 8. We compare the decision discrepancies between AV6 and AV2 with A0, revealing two distinct phenomena. (1) In the first row, the diagonal line of the last subplot sharply brightens compared to the former four subplots, indicating the effectiveness of the MS-Adapter in dynamically switching the decisive pattern towards the audio-dominant one. (2) In comparison to the first row, the diagonal line of the second row remains consistently bright across various missing video frame rate inputs. This further confirms that AV2 is a modality-biased model that consistently relies on the audio modality.

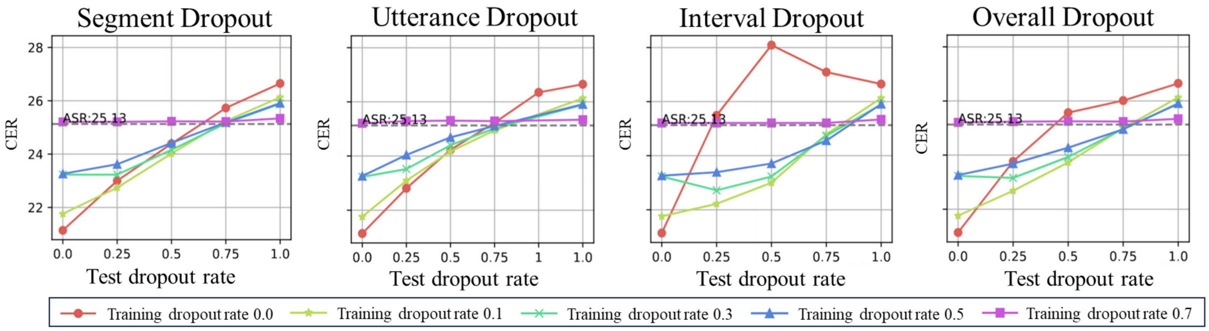


Figure 9. Performance degradation curves of AVSR systems with different training dropout rate test in different test dropout methods.

multichannel far/middle-field audio for dereverberation and source signal separation in both the training and test sets. Specifically, we utilize a GPU-accelerated version of GSS [59]. As shown in Figure 6, it effectively enhances the spectral speech components for the target speaker while minimizing speech distortion compared to the CPU version [58]. Then We apply a short fourier transform and mel filter to obtain 80-dimensional Fbank frames in the frequency domain, with a 0.25s window length and a 0.01s frame shift, using a 16k sample rate. For video, following [3], we acquire grayscale lip ROI with 88×88 pixels before inputting it into the network. In ASR training, all enhanced far/near/middle-field audio is used, employing various data augmentation techniques, such as adding noise, Room Impulse Response (RIR) convolution, speed perturbation, and concatenating nearby segments to create a 10-fold training set. The technique of concatenating nearby segments effectively generates a longer segment, providing additional content information. This technique can used in both training and decoding phrases. For VSR, we pre-train the visual frontend on far/middle-filed video following [18] by correlating lip shapes with syllabic HMM states (3168 Senone units). In AVSR training, the audio and visual branches are initialized with pre-trained ASR and VSR representations. We create an 8-fold training set, incorporating two effective data augmentation techniques: (1) matching synchronous audio and video segments recorded in different fields and (2) concatenating nearby segments in both video and audio.

Training Implementation Details All conformers in our network use the same set of hyperparameters ($n_{\text{head}} = 8$, $d_{\text{model}} = 512$, $d_{\text{ffn}} = 2048$, $CNN_{kernel} = 5$). The decoder consists of six transformer blocks ($n_{head} = 8$, $d_{\text{model}} = 512$, $d_{\text{ffn}} = 2048$). For unimodal model training, we strictly adhere to [18]. In robustness training for this work, all models are optimized using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 0.0012. For MDA-KD implementation in Section 7, we utilize the intermediate representation samples from the output of ResNet-18 and the first layer of Conformer in the video branch in practice. For further exploration, we successfully validate that the output of the multimodal encoder exhibits similar effectiveness in achieve both missing robustness and accuracy with complete input. The learning rate undergoes a linear warm-up during the first 3000 steps and subsequently decreases proportionally to the inverse square root of the step number. We train for 12 epochs with a training batch size of 128, utilizing 4 NVIDIA Tesla A100 48GB GPUs. For MS-Adapter adaptation, we train 5 epochs with a batch size of 144 and a learning rate of 0.0002. During decoding, the beam size is set to 10 in beam search. Additionally, a 6-layer transformer-based language model trained on the transcription of the training set is employed in decoding, with a weight of 0.2, although it brings negligible performance improvement.

Dropout Setting Details Segment Dropout, Utterance Dropout, and Interval Dropout are employed to simulate missing video modality in different scenarios. Segment dropout occurs when contiguous segments of video frames are dropped, which often occurs when the lips are covered or when the person is in a side-face pose. Utterance Dropout refers to dropping the entire video, which represents situations where the camera is turned off. Interval Dropout means dropping ($dropout\ rate < 0.5$) or preserving ($dropout\ rate > 0.5$) video frames at a fixed interval, indicating missing due to network latency or hardware computation bottleneck. Unlike previous work [19], we have simplified the test suites by removing frame-level random dropout to ensure experimental reproducibility. Furthermore, the starting position for segment dropout is randomly determined. Considering our study on modality bias and robustness, the focus lies more on the dropout rate than the dropout method.

14. More Discussions on Related Works

Missing Modality in Multimodal Learning The missing modalities problem is common in multimodal applications, whether in the training or testing stage, and has attracted a lot of research interest. For modality-balanced models like Multimodal Emotion Recognition (MER) and multimodal sensor fusion in autonomous driving, the mainstream approach is to learn joint multimodal representations to capture intra- or inter-modal features cross modalities [53, 54]. For modality-biased models, data augmentation methods such as modality dropout effectively address out-of-distribution issues [19–21]. In cases of severe modality absence, generative models [51, 52] and meta-learning based methods [60] are used to directly predict the missing modalities based on available modalities or a few-shot paired samples. For AVSR, we prioritize efficiency and opt for dropout due to its plug-and-play nature and lightweight implementation.

Modality Bias in Multimodal Learning The modality bias is observed in many multimodal applications, since there is a direct correlation between a specific modality and the target task, leading to one modality dominating the decision-making process [61]. In the VQA, several de-bias methods have been proposed. New datasets following the answer distribution balancing rule have been constructed to address the language prior problem [62]. Guo et al. [63] develop plug-and-play loss function methods that can adaptively learn the feature space for each label. Gat et al. [61] have proposed a method based on the log-Sobolev inequality. Although many studies have been conducted on removing bias, there is a lack of conception or mathematical models to describe model bias and limited research on the impact of bias on the modality missing problem.

Dropout-Induced Modality Bias on Mulitmodal Tasks For AVSR, this excessive modality bias towards audio is a double-edged sword, as it brings robustness to missing video data while degrading the performance of a multimodal model on complete multimodal data. It causes the model to tend towards trivial solutions and ignore optimal ones. As a result, the model neglects visual cues, making it sensitive to perturbations in speech. This contradicts the intention of AVSR as a multimodal robust speech recognition application in noisy environments.

For other multimodal applications, Hazarika et al. investigate the robustness of Multimodal Sentiment Analysis (MSA), which is a multimodal classifier with text, visual, and audio as input [22]. By applying dropout on the training text, the robustness against missing text can be achieved without compromising the original performance. These findings seem to be inconsistent with the degradation results observed in AVSR. While the truth is that the common MSA system exhibits a severe modality bias dominated by text, and it is sensitive to perturbations in text but robust to other modalities. Applying dropout on text helps to mitigate over-reliance and encourages the model to leverage supplementary information across modalities. A similar phenomenon has been observed in AVSR when applying dropout on the audio modality [20, 21]. Interestingly, in our research on video robustness in AVSR, video is a supplementary modality within the system rather than the dominant one. As a result, we emphasize that it is important to first determine whether the system has a dominant or supplementary modality when studying the robustness of a specific modality within a multimodal bias system.