

# Supplementary Materials: Boosting Speech Recognition Robustness to Modality-Distortion with Contrast-Augmented Prompts

Anonymous Authors

## 1 IMPLEMENTATION DETAILS

### 1.1 Baselines Training Setup

In this paper, we compare our experimental results with previous work [1, 5–8]. Source code, detailed experiment setup, and pre-trained checkpoints are available for reference on their official GitHub repository. We utilize models pre-trained on LRS3 [2] and VoxCeleb2 [3] datasets, finetuning on LRS2-DISTORTED to simulate real-world distortion scenarios. Despite variations in model scale constraints by the provided checkpoints, PCD method with AV-Hubert as the backbone achieves superior performance using only the Transformer-BASE scale model compared to other baselines.

### 1.2 Data Preprocessing

Prior to model input, we conduct preprocessing on the audio-visual source files to generate audio-video utterance following [8]. we utilize dlib [4] to detect 68 facial keypoints and align each frame with a reference face frame through affine transformation. Subsequently, We extract a  $96 \times 96$  region-of-interest (ROI) from each visual utterance, focusing on the talking head video centered on the mouth. To acquire knowledge from high-quality domains, we employ image augmentation during the pretraining process, including cropping  $88 \times 88$  from the entire ROI and randomly flipping them horizontally. For the raw audio processing, we extract the 26-dimensional log filterbank energy feature at a stride of 10 ms, to generate audio utterance served as model audio input. Given that the image frames are sampled at 25Hz, we stack the 4 neighboring acoustic frames to synchronize the two modalities. We also apply audio noise with 25% noise rate to each audio utterance.

### 1.3 PCD Training Setup

The pre-training process of PCD is built upon AV-Hubert, where we utilize the provided checkpoints incorporating pre-training of both the AV-Hubert structure and the decoder for subsequent fine-tuning. Two scales of models are provided, namely Transformer-BASE and Transformer-LARGE, with 12/24 Transformer layers, 768/1024 embedding dimension, 3072/4096 feed-forward dimension, and 12/16 attention heads respectively. We adopt the audio-visual alignment encoder pre-training on LRS3 and VoxCeleb2 and decoder structure pre-training on LRS3.

The fine-tuning process of PCD follows a two-stage approach. Initially, we fine-tune the backbone model on LRS2-DISTORTED, which can be regarded as our low-quality target domain, to facilitate model adaptation to the common domain shared by the diverse modality-distortion condition in LRS2. Subsequently, we freeze the backbone model and proceed to train the cluster-prompt under the same configuration, thereby enhancing the model’s robustness to diverse scenarios. To comprehensively validate the robustness of

the PCD method to modality-distortion, we conduct experiments on both the 29h and 224h subsets of the LRS2 dataset using both Transformer-BASE and Transformer-LARGE model configurations. Each tuning phase of PCD is trained utilizing the Adam optimizer, where the learning rate undergoes a warm-up process for the initial 50% of updates, reaching 0.0005. In the main experiment, we fine-tune PCD and baselines under distortion settings with  $\eta=70\%$  and  $\mu=60\%$ , and conducted inference under various configurations. For ablation experiments, we maintain identical training settings and default to setting  $\eta=70\%$  and  $\mu=80\%$  for inference. We primarily employ the Transformer-BASE model and 29h dataset to validate the effectiveness of PCD and its robustness across different tasks.

## 2 ADDITIONAL EXPERIMENT RESULTS

### 2.1 Robustness to specific distortion scenarios

The objective of PCD is to adapt the model to each scenario using fewer computational resources by employing commonly encountered real-world distortion data. The experimental results showing in previous sections are conducted in testing configurations encompassed various modality-distortion scenarios, validating the enhancement of the PCD method across multiple scenarios. In this section, we conduct a comparative analysis between PCD and AV-Hubert for each modality-distortion condition, reflecting more practical application scenarios. As depicted in Figure 1, the comparison includes results obtained from audio-distorted data, and video-distorted data. To demonstrate the effectiveness of PCD, we include comparisons with AV-Hubert trained with the random parameter same size as prompt for specific scenarios. The guidance provided by PCD significantly enhances the recognition performance across diverse scenario configurations, with particularly notable improvements observed in instances of audio distortion. The results demonstrate that PCD can be fine-tuned on imperfect data to achieve adaptation across multiple scenarios, thus offering promising practical applications in situations where real data quality may be inadequate.

### 2.2 Qualitative Results

Table 1 shows the example outputs from AV-Hubert and PCD<sub>pro</sub>. All models are based on Transformer-BASE with LRS2-DISTORTED 29h dataset under the settings of  $\eta=70\%$  and  $\mu=60\%$  during training and inference phases. Under the influence of distortion, the integrity of audio-visual data is compromised, rendering the original model incapable of fully leveraging audio-visual features for recognition, thereby resulting in the generation of incoherent sentences. Qualitatively, our approach yields transcriptions with higher accuracy, indicating the effective discrimination of distortions and the guidance provided by PCD to the model.

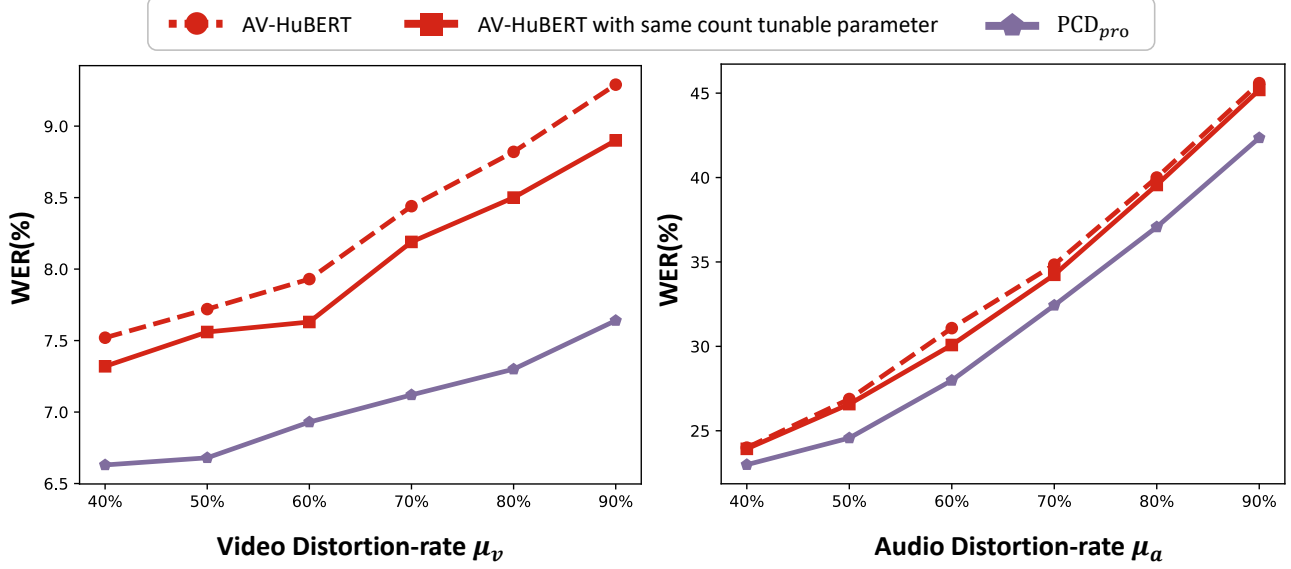





Figure 1: Performance of PCD in specific distortion scenarios where distortion occur in a single modality, including video and audio modalities.

Table 1: Qualitative comparison on various modality-distortion scenarios. Red words highlights misidentified words and the (red words) in parentheses highlight the absent words. The green boxes represent segments affected by distortion.

Distortion type	Qualitative examples
clean	 <p>Ground Truth: i shouldn't say those things            AV-Hubert: i shouldn't <b>take</b> those things            PCD<sub>pro</sub>: i shouldn't say those things</p>
visual-distortion	 <p>Ground Truth: buying a new build property should mean you don't have any work to do            AV-Hubert: buying a new <b>bill</b> property should mean you don't have (any) work to do            PCD<sub>pro</sub>: buying a new build property should mean you don't have any work to do</p>
audio-distortion	 <p>Ground Truth: four little turtles named after Italian renaissance artists            AV-Hubert: four little <b>title sniped</b> after <b>exactly release on</b> artists            PCD<sub>pro</sub>: four little turtles <b>framed</b> after <b>exactly</b> renaissance artists</p>

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence* 44, 12 (2018), 8717–8727.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496* (2018).

[3] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Interspeech 2018*. ISCA. <https://doi.org/10.21437/interspeech.2018-1929>

[4] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* 10 (dec 2009), 1755–1758.

[5] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), 1–5. <https://api.semanticscholar.org/CorpusID:257767381>

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

[7] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. 2022. Leveraging Unimodal Self-Supervised Learning for Multimodal Audio-Visual Speech Recognition. *arXiv:2203.07996* [cs.SD]

[8] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. *arXiv:2201.02184* [eess.AS]