When AVSR Meets Video Conferencing: Dataset, Degradation, and the Hidden Mechanism Behind Performance Collapse

Supplementary Material

8. Appendix

8.1. Detialed Transmission Process

We transmit the test sets of LRS3 (1321 videos), Lombard-Grid (585 videos), and Chinese-Lips (3908 videos) through three video conferencing platforms, including Tencent Meeting, Lark, and Zoom. Specifically, we first concatenate all videos in each test set into a single long video. Between two consecutive videos, we insert a 0.4 second pure black and pure white clip as a segmentation flag. We then start a video conferencing session. On the sender side, we use the virtual camera of OBS as the camera input of the session. On the receiver side, we use OBS screen recording to capture the transmitted video. Note that our capture pipeline and parameter settings on the receiver side are kept identical to those described in Section 4.2.4. For the transmitted videos, we cut them into individual clips that correspond to the original datasets according to the segmentation flags. We then re-encode these clips using the codec and resolution settings of the original datasets to ensure consistency.

8.2. Dataset Construction

8.2.1. Chinese Grid-style Corpus

We follow the design of DB-MMLC [25]. The Chinese corpus in MLD-VC consists of GRID-style Mandarin sentences composed of five components in a fixed order, namely name, verb, classifier, adjective, and noun. For each component, we provide 20 phonemically balanced candidate words and construct sentences by randomly sampling from these candidates. Tab. 6 lists all candidate words. Note that the randomly generated sentences do not carry any actual meaning.

8.2.2. Recoding Environment

The MLD-VC dataset was recorded in a classroom located next to a corridor without any soundproofing. The building that houses the classroom is inside the campus, where nearby buildings were under construction and airplanes frequently flew overhead. As a result, the microphones unavoidably captured various types of environmental noise, including footsteps, conversations, passing cars, aircraft rumble, wind, and construction noise. To preserve the authenticity of the meeting scenario, we deliberately avoided applying any noise reduction techniques to suppress these sounds in the recorded speech. The recording sessions were conducted from 9 a.m. to 9 p.m. over several consecutive days. With this setup, the data in MLD-VC depict realistic video conference conditions. Fig. 4 illustrates the specific classroom setting used for recording.

Table 6. The candidate phoneme-balanced word and the corresponding Pinyin.

Name	Verb	Classifier	Adjective	Noun
旭峰 (Xufeng)	买 (Mai)	零个 (Lingge)	大 (Da)	沙发 (Shafa)
青木 (Qingmu)	乘 (Cheng)	一架 (Yijia)	小 (Xiao)	飞机 (Feiji)
林俊 (Linjun)	坐 (Zuo)	两只 (Liangzhi)	旧 (Jiu)	火车 (Huoche)
建树 (Jianshu)	去 (Qu)	三条 (Santiao)	快 (Man)	菠萝 (Boluo)
郭浩 (Guohao)	拿 (Na)	四段 (Siduan)	慢 (Man)	枕头 (Zhentou)
南月 (Nanyue)	养 (Yang)	五棵 (Wuke)	新 (Xin)	床 (Chuang)
赵坤 (Zhaokun)	来 (Lai)	六艘 (Liusou)	硬 (Ying)	村 (Cun)
文路 (Wenlu)	给 (Gei)	七斤 (Qijin)	软 (Ruan)	蝴蝶 (Hudie)
范畴 (Fanchou)	挥 (Hui)	八两 (Baliang)	胖 (Pang)	鹅 (E)
宋坏 (Songhuai)	拔 (Ba)	九碗 (Jiuwan)	瘦 (Shou)	鸭 (Ya)
孙西 (Sunxi)	踢 (Ti)	零杯 (Lingbei)	长 (Chang)	平板 (Pingban)
弘扬 (Hongyang)	催 (Cui)	一瓶 (Yiping)	高 (Gao)	水杯 (Shuibei)
启辰 (Qichen)	蹲 (Dun)	二听 (Erting)	甜 (Tian)	西瓜 (Xigua)
加号 (Jiahao)	看 (Kan)	三根 (Sangen)	热 (Re)	台灯 (Taideng)
壮硕 (Zhuangshuo)	啃 (Ken)	四张 (Sizhang)	香 (Xiang)	表格 (Biaoge)
明惠 (Minghui)	抱 (Bao)	五枚 (Wumei)	乱 (Luan)	电池 (Dianchi)
凌翔 (Lingxiang)	喝 (He)	六则 (Liuze)	轻 (Qing)	大米 (Dami)
云妮 (Yunni)	试 (Shi)	七顿 (Qidun)	方 (Fang)	键盘 (Jianpan)
佳倩 (Jiaqian)	吃 (Chi)	八匹 (Bapi)	圆 (Yuan)	面条 (Miantiao)
耕田 (Gengtian)	种 (Zhong)	九位 (Jiuwei)	细 (Xi)	字帖 (Zitie)

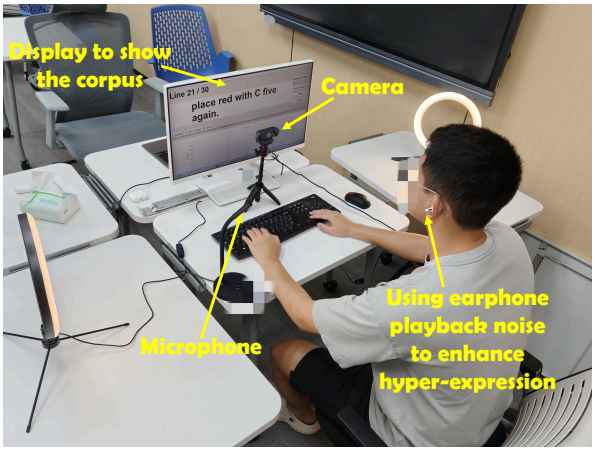

Figure 4. Picture of the recording environment.

8.2.3. Subdataset Duration

Fig. 5 presents the duration of each subset in the MLD-VC dataset. The “Offline” subset contains data that is not transmitted through any platform while still preserving the hyper-expression effect. Each of the remaining subsets corresponds to a specific video conferencing platform and therefore reflects both platform transmission and the presence of the hyper-expression effect.

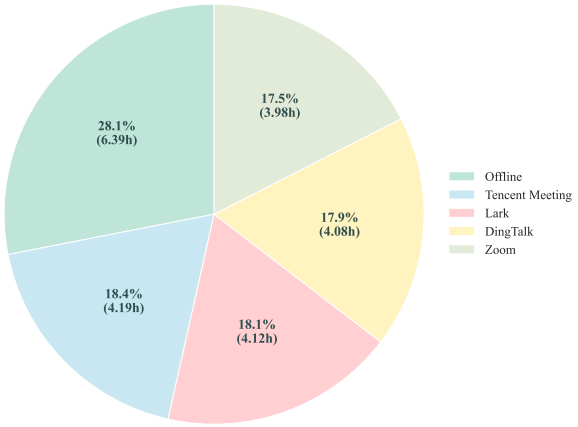

Figure 5. Duration distribution across subsets of the proposed MLD-VC dataset. "Offline" refers to the subset of recorded content that was captured before video conferencing.

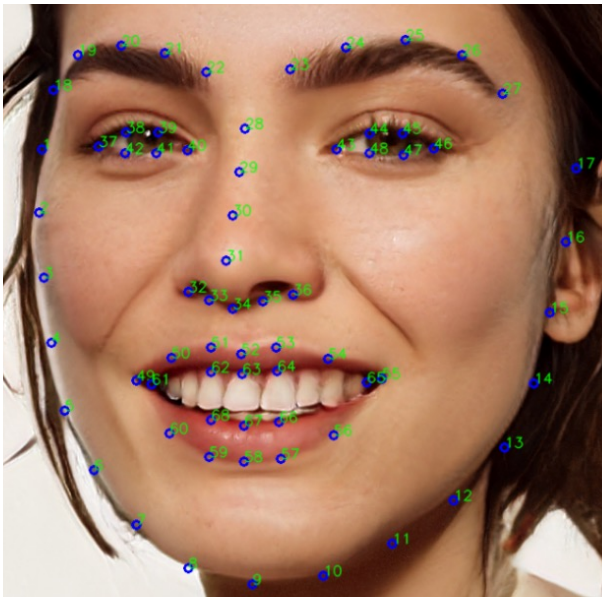

Figure 6. Duration distribution across subsets of the proposed MLD-VC dataset. "Offline" refers to the subset of recorded content that was captured before video conferencing.

8.3. Visiual Feature Analysis

8.3.1. Metric

For the analysis of visual features, we do not use traditional image quality metrics such as PSNR and SSIM. This is because compression and network latency in online transmission inevitably degrade image quality and can even cause blurring, so abnormal values of these metrics are unavoidable. For the visual input, the core objective of AVSR models is to recognize lip motion. Therefore, we select lip width, lip height, and lip roundness as visual metrics that are directly related to lip movement.

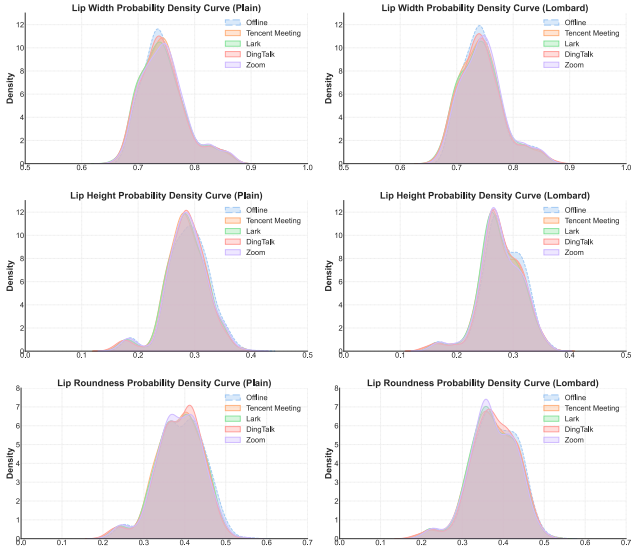

Figure 7. Probability density curves of three visual features (lip width, lip height, and lip roundness) across five subsets in the proposed MLD-VC under *Plain* (left column) and *Lombard* (right column).

Table 7. Peak abscissas of probability density curves for different visual features.

Visual Feature	Platform				
	Offline	Tencent Meeting	Lark	DingTalk	Zoom
Lip Width (*clean*)	0.73	0.74	0.74	0.74	0.75
Lip Width (*80 dB*)	0.74	0.74	0.74	0.74	0.75
Lip Height (*clean*)	0.29	0.28	0.28	0.29	0.28
Lip Height (*80 dB*)	0.27	0.26	0.26	0.27	0.27
Lip Roundness (*clean*)	0.41	0.40	0.40	0.41	0.37
Lip Roundness (*80 dB*)	0.36	0.36	0.36	0.36	0.36

Each metric is computed from facial landmark locations. Fig. 6 shows the facial landmarks and their indices. In implementation, we first detect 68 landmarks on each video frame and normalize the scale of all landmark coordinates using the interocular distance. Specifically, we compute the average of the landmarks of the left eye (points 37 to 42) and the right eye (points 43 to 48) to obtain the centers of the two eyes, take their Euclidean distance as the interocular distance, and use this distance as a normalization factor to rescale all facial landmark coordinates. This removes scale variations caused by different speakers and camera distances.

In the normalized coordinate system, we compute lip width, lip height, and lip roundness from the lip landmarks. First, we define lip width W as the Euclidean distance between the left and right mouth corners (points 49 and 55). Second, we compute the average vertical coordinates of the

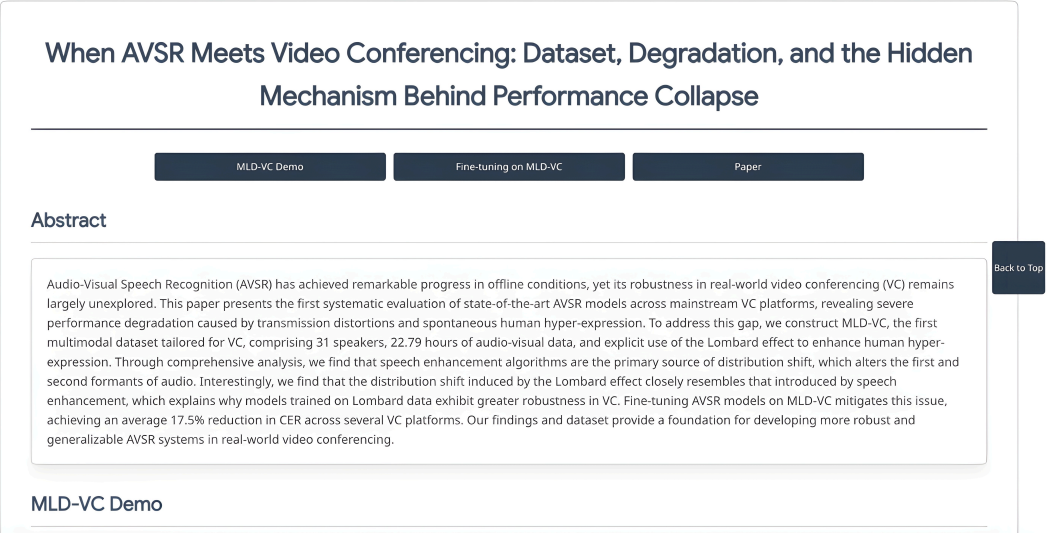

Figure 8. The static page MLD-VC demo in the supplementary material.

upper lip point set $\{51, 52, 53, 62, 63, 64\}$ and the lower lip point set $\{57, 58, 59, 66, 67, 68\}$, and define lip height H as the absolute difference between these two averages. On this basis, we define lip roundness C as the ratio between height and width, that is $C = \frac{H}{W}$. This metric reflects the relative proportion of the lip shape in the vertical and horizontal directions, where values of C closer to 1 indicate that the lip shape is closer to a circle.

8.3.2. Results

We plot the probability density curves of the visual features in Fig. 7. In addition, Tab. 7 lists the horizontal coordinate of the peak for each curve. The results show that the three selected visual features exhibit no obvious change between the offline condition and the various video conferencing platforms. This indicates that landmark-based visual geometric features are robust in video conferencing scenarios. Therefore, we recommend using these stable visual features as the input to the visual stream rather than relying solely on unstable lip images in the video conferencing scenarios.

8.4. Implement Details of Fine-tuning

To ensure a fair comparison, we maintained consistent hyperparameters between the fine-tuning and ablation experiments. The learning rate was set to 0.0001, the number of fine-tuning steps was set to 800, the weight decay was set to 0.01, and the dropout rate was set to 0.3. Besides, to prevent overfitting in the fine-tuned model, we added data from the original training set. Specifically, the ratio between fine-tuning data and original training data was 7 to 3.

9. Dataset Demo

We built a local static web page to present demos of our dataset and included it in the compressed supplementary material package, namely ”supplementary-material.zip”. In this web demo, we present recordings from two speakers of different genders, captured on different video conferencing platforms, in different languages, and under different noise conditions to enhance hyper-expression. Fig. 8 illustrates the static web page. We recommend that you view the file named “demo.html” in the supplementary material to gain a better understanding of MLD-VC.