Candor-LR: A Dyadic Conversational Dataset for Audio-Visual Speech Recognition

Rishabh Jain Aristeidis Papadopoulos Zhaofeng Lin Naomi Harte
Sigmedia Group, School of Engineering, Trinity College Dublin, Ireland
{rijain, papadoar, linzh, nharte}@tcd.ie

***Abstract*—Current audio-visual speech recognition (AVSR) benchmarks, like LRS3, rely heavily on clean, scripted and rehearsed speech. They fail to reflect the complexity of natural conversation, which involves overlapping speech, spontaneous turn-taking, unscripted vocabulary and variable acoustic conditions. To shift the field toward realistic dialogue, we introduce Candor-LR, a conversational benchmark derived from the CANDOR corpus of 1,656 natural dyadic videoconferences. Our custom data preparation pipeline yields 713.5, 10.1, and 60.1 hours of training, validation, and test data, respectively. Evaluating pretrained AVSR models on Candor-LR reveals that audio-only accuracy drops sharply compared to LRS3, but visual cues compensate effectively, driving much larger performance gains on Candor-LR than on LRS3. Furthermore, training on this corpus significantly improves cross-domain robustness under both clean and noisy conditions, as its realistic conversational data captures broader audio-video features. We open-source our pipeline to ensure reproducibility, establishing Candor-LR as a challenging benchmark for conversational AVSR.**

***Index Terms*—Audio-visual speech recognition, CANDOR dataset, benchmark, noise robustness, video conferencing data**

I. Introduction

Audio-visual speech recognition has made remarkable progress on benchmark datasets [1]–[7], with AV-HuBERT [8] reaching 1.47% WER on LRS3 [9] and Auto-AVSR [10] reaching 0.90%. However, these gains are largely measured on scripted datasets such as LRS2 [11] and LRS3 [9]. These datasets feature professional speakers delivering well-articulated speech under relatively consistent recording conditions and with high-quality microphones, which differ substantially from everyday conversational settings. They contain no overlapping speech, no spontaneous interruptions, and little acoustic variability. As a result, strong benchmark performance does not necessarily guarantee robustness in real conversational settings.

Real conversation is fundamentally different. People interrupt each other, turn away mid-sentence, speak over one another, and deal with variable room acoustics [12], [13]. These conversational behaviors are largely absent from current AVSR benchmarks, which means strong performance on scripted speech does not necessarily transfer to everyday dialogue. Recent work has also questioned what current systems actually learn from visual speech. State-of-the-art VSR and AVSR models often rely heavily on dataset patterns and language modeling rather than genuinely stronger visual representations [14], [15]. Other analyses further show that low WER does not necessarily imply strong use of visual information, suggesting that current benchmarks can overestimate true visual speech understanding [16].

Recent AVSR datasets have started to shift towards more realistic conditions [17]–[20]. WildVSR shows that models tuned to LRS3 degrade sharply on a harder in-the-wild test set [21]. LRS-VoxMM [22] further shows that visual cues become more important as audio quality degrades, while the addition of noise and reverberation creates substantially more challenging AVSR evaluation conditions. Cocktail-party AVSR [23], [24] adds overlapping speakers, interfering talkers, and silent-face segments, showing that current models can deteriorate sharply in realistic multi-speaker settings. In-the-wild conversational benchmarks [21], [25] further show that AVSR becomes more difficult once audio quality degrades, and recent work on video conferencing [26], [27] shows that platform-induced distortions and speech enhancement artifacts can also cause severe performance collapse. Together, these datasets are valuable, but they still do not provide the scale and realism needed to move AVSR beyond broadcast-style evaluation and toward a more realistic conversational style paradigm [28]. There remains a clear need for a benchmark that can support both training and testing in realistic dyadic settings and to drive progress towards conversational AVSR.

To address this gap, we present Candor-LR, a benchmark derived from the CANDOR [29] corpus of 1,656 natural dyadic videoconference conversations. We develop a custom pipeline to convert CANDOR into AVSR-ready data using Speechmatics [30] word-level time-aligned transcripts. Candor-LR contains 713.5 hours of AVSR-formatted data for training, 10.1 hours for validation, and 60.1 hours for testing. Unlike scripted benchmarks, Candor-LR is built from spontaneous dyadic conversation, including turn-taking, unscripted vocabulary, and realistic acoustics, with final segments extracted as single-speaker phrases from per-speaker tracks. Our experiments show that AVSR systems trained on scripted data perform poorly on real conversation. Visual cues matter more in these realistic settings, and training on Candor-LR improves robustness on both clean and noisy speech. Our results highlight a key limitation in AVSR: models trained on scripted datasets do not generalize well to real dialogue. Rather than offering only an incremental step in AVSR research, our work aims to reposition AVSR around natural conversational

This publication emanates from research supported by Taighde Éireann – Research Ireland, Grant number 22/FFP-A/11059.

```
{session_id}/
├ processed/
│ ├ {session_id}.mp4          # Combined video (640x240)
│ ├ {session_id}.mp3          # Combined stereo audio
│ ├ {user_id_1}.mp4           # Spkr 1 video (320x240, 30fps)
│ ├ {user_id_2}.mp4           # Spkr 2 video (320x240, 30fps)
│ ├ channel_map.json          # Maps L/R audio to user IDs
│ └ metadata.json             # Session info and sync details
├ transcription/
│ ├ transcribe_out.json       # Raw AWS ASR output
│ ├ transcript_aud.csv        # Turn-level transcripts
│ ├ transcript_back.csv       # Backchannel-aware model
│ └ transcript_cliff.csv      # Sentence-boundary model
└ raw/                        # Original recordings
```

Fig. 1: Directory structure of CANDOR dataset.

settings that better reflect real-world use. We also open-source the pipeline on GitHub for converting the original CANDOR corpus[1] into AVSR-ready format.

II. CANDOR Cleaning for AVSR

The CANDOR dataset [29] has 1,656 natural dyadic conversations. It contains over 850 hours of video and 7 million words. Most AVSR datasets, like LRS2 or LRS3, only show one person talking for a few seconds. In contrast, CANDOR contains two-speaker sessions of approximately 26-30 minutes. It was originally developed for research on natural conversation and social interaction, not audio-visual speech recognition. As the conversations are unscripted, they include overlapping speech, a wide range of vocabulary, and frequent head movements. Off-the-shelf AVSR preprocessing tools are designed for broadcast, single-speaker recordings and are not suited for conversational data of this kind. We therefore developed a custom pipeline to convert CANDOR into a format suitable for AVSR training and evaluation. The original directory structure of the CANDOR dataset is shown in Figure 1.

A. Transcript Source: Speechmatics over Originals

The CANDOR corpus includes automatic transcripts created with the AWS Transcribe API. Based on these, Reece et al. [29] provided three ways to split the dialogue into speaker turns: Audiophile (simple speaker switches), Cliffhanger (sentence boundaries), and Backbiter (which accounts for listener feedback). While these are useful for understanding the conversation, they do not include word-level timestamps. For AVSR training, we need millisecond-level precision to match the audio to the video, which these original transcripts do not provide. To address this limitation, we use word-level transcripts generated by the Speechmatics ASR system, as produced by Russell et al. [31]. Speechmatics was selected for its strong performance on conversational speech and its ability to accurately capture non-lexical tokens (like "uh" or "um"). Most importantly, it provides per-word confidence scores, punctuation, and millisecond-precision start and end timestamps for every word. These timestamps are used to drive all timing and segmentation in our pipeline. The organization of these Speechmatics outputs is shown in Figure 2.

[1]Users must obtain their own license from the CANDOR authors [29]; we do not distribute the dataset and are not responsible for licensing.

```
candor_speechmatics/

 {session_id}_0.json    # speaker 0 word-level JSON
 {session_id}_1.json    # speaker 1 word-level JSON
 {session_id}.TextGrid    # combined Praat alignment
```

Fig. 2: Directory structure of Speechmatics-ASR outputs.

B. Speaker-Video Mapping

CANDOR's dyadic nature introduces an alignment challenge: speaker videos are identified by unique User IDs (e.g., `{user_id}.mp4`), whereas the Speechmatics transcripts are indexed by audio channel number (e.g., `_0.json` and `_1.json`). Without a direct link, the transcript for a specific speaker cannot be automatically identified. To resolve this, we utilize the `channel_map.json` file (see Figure 1). This metadata file maps the Left (L) and Right (R) audio channels to their corresponding speaker User IDs. Specifically, we map channel `0` to the `L` video and channel `1` to the `R` video. This ensures that each word-level transcript is correctly synchronized with the appropriate speaker's video stream.

C. Phrase-Level Temporal Segmentation

CANDOR sessions average 26 minutes, necessitating segmentation into 2–5 second clips (similar to LRS3) suitable for AVSR training. We group words from the Speechmatics output into phrases, breaking at punctuation or inter-word gaps over 0.5 seconds. This method avoids the context loss of word-level segments (0.2-1.0s) and the silence of turn-level segments (0.5-84s), aligning our data with LRS2/LRS3 standards. On the full corpus, this process yields clips ranging from 0.8 to 5.0 seconds, with a mean duration of 3.6 seconds.

D. Face Detection and Mouth ROI Extraction

We applied RetinaFace [32] frame-by-frame to each video clip with detection threshold of 0.8. Following successful face detection, 68 facial landmarks were extracted, and the mouth region (landmarks 48-68) was cropped and resized to 96×96 pixels at 25 fps (downsampled from the original 30 fps for consistency with LRS2/LRS3 preprocessing). To reduce inter-frame jitter caused by spontaneous head movement, a 3-frame moving average is applied to the landmark positions before cropping.

E. Audio Processing and Synchronization

Audio segments are extracted directly from the individual per-speaker audio tracks using the same phrase-level timestamps derived from the Speechmatics word alignments. Extracting from dedicated per-speaker tracks, rather than the combined stereo recording, eliminates cross-talk from the other conversation participant. Each segment is converted from stereo to mono, resampled to 16 kHz, and saved as a lossless WAV file. Because audio and video are extracted using the same timestamps, frame-perfect synchronization is guaranteed without requiring any additional alignment step.

F. Text Normalization and Quality Filtering

To ensure compatibility with standard AVSR frameworks like AV-HuBERT [8] and Auto-AVSR [10], transcripts are normalized by removing punctuation, collapsing redundant whitespace, and converting text to lowercase. Remaining disfluency markers from the Speechmatics output are removed to maintain consistent training labels. We further apply three sequential filters at the phrase level to remove artifacts unsuitable for AVSR. First, a duration filter discards clips shorter than 800 ms, as these lack sufficient visual context. Second, a word count filter removes utterances with fewer than two words. Finally, a filler word filter removes phrases consisting entirely of non-lexical tokens (e.g., "uh," "um," "mhm"); phrases containing at least one lexical word are retained. These filters are applied to the full dataset prior to any splitting, ensuring consistent quality across training, validation, and test sets, and any clip failing them is discarded alongside its corresponding audio and transcript.

G. Data Splitting and Output Format

The dataset is split into training, validation, and testing sets using a speaker-aware partitioning approach. To construct a truly unseen test set, we identify speakers who appear in exactly one session across the entire corpus: a total of 701 such speakers accounting for approximately 164.9 hours of audio. These single-occurrence speakers are reserved exclusively for the test split, ensuring that no speaker in the test set is encountered during training or validation. The remaining multi-session speakers are partitioned into training and validation sets using a fixed random seed, with conversational contexts never overlapping across splits. These splits are saved as canonical ID files (`candor-train.id`, `candor-valid.id`, `candor-test.id`) to ensure reproducibility. The complete pipeline explanation, hyperparameters used and source code are available on our GitHub.[2] The prepared AVSR dataset is referred to as **Candor-LR** throughout the rest of this paper.

III. Candor-LR

Candor-LR contains **783.7** hours (787,670 utterances) of AVSR-formatted data. This comprises approximately **713.5 hours** (718,648 utterances) for training, **10.1** hours (10,304 utterances) for validation, and **60.1** hours (58,718 utterances) for testing, corresponding to 91.2%, 1.3%, and 7.5% of the total data respectively. Across all splits, Candor-LR contains 1,554 speakers with valid demographic metadata (1,350 train, 44 validation, 160 test), of which 1,521 are unique after accounting for 33 speakers overlapping between training and validation. Speakers in the test set are disjoint from those in the training and validation sets. For comparison, LRS3 consists of 9,506 videos (5,090 pre-train, 4,004 trainval, 412 test) totaling 151,819 utterances (118,516 pre-train, 31,982 trainval, 1,321 test) [9]. The LRS3 test set contains only 1,321 utterances

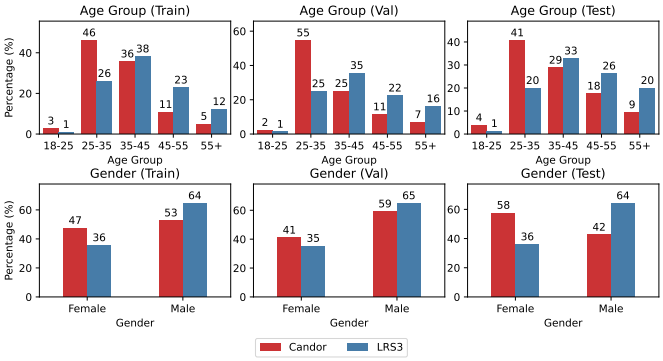


Fig. 3: Age and gender distributions for Candor-LR and LRS3 estimated using UniFace. Candor-LR shows improved demographic balance across both dimensions.

(≈ 1 hour), compared to Candor-LR's substantially larger test set of 58,718 utterances (60.1 hours).

The authors of CANDOR release no speaker-level demographic metadata alongside the corpus. Therefore to assess the demographic diversity of Candor-LR, we compare its speaker-level age and gender distribution against LRS3. We extract metadata using the open-source UniFace toolkit[3], which estimates age and gender from video frames following the same methodology as Lin et al. [33]. Figure 3 shows the percentage-wise distributions.

Demographics: Using UniFace metadata for age and gender estimation, Candor-LR shows a larger proportion of speakers in the 25-35 age bracket, 46% of training speakers are aged 25–35, compared to 26% in LRS3. LRS3 has higher concentration in older groups: 23% aged 45–55 and 12% aged 55+, compared to Candor-LR's 11% and 5%, respectively. For gender, Candor-LR is more balanced in training (47% female vs. 36% in LRS3), narrowing the gap by nearly 10 percentage points. This gender balance is even more pronounced in the test split, where Candor-LR has a slight female majority (57% female) while LRS3 remains male-skewed (64% male). Overall, Candor-LR offers balance across both age and gender.

IV. Experiment and Results

A. Benchmarking Candor-LR with pretrained AVSR Models

To assess zero-shot cross-domain generalization, we evaluate three pretrained AVSR models on Candor-LR without downstream finetuning. **AV-HuBERT** [8] is a state-of-the-art self-supervised representation learning framework that learns by capturing the correlation between audio and visual streams. The model encodes masked audio and image sequences via a hybrid ResNet-Transformer architecture to predict multimodal hidden units. For our work, we use the checkpoint trained on 1,759-hour LRS3 and VoxCeleb2 [34] datasets. **Llama-AVSR** [35] is a multimodal framework that utilizes a frozen large language model (LLM) [36] as an auto-regressive text decoder. It extracts features using a frozen pretrained encoder [8], [37] and maps them into the LLM's text space using

[2]https://github.com/rishabhjain16/lipreading-data-guide/tree/main/Candor

[3]https://github.com/yakhyo/uniface

TABLE I: WER (%) performance comparison across modalities (VO, AO, and AV) on LRS3 and Candor-LR datasets.

Model	Learning Type	Train Hours	LRS3				Candor-LR			
			VO	AO	AV	Δ	VO	AO	AV	Δ
AV-HuBERT	Self-Supervised	1,759	28.69	1.95	1.47	+0.48	78.35	24.32	22.76	+1.56
Llama-AVSR	LLM-based Decoder	1,759	26.20	0.74	0.79	-0.05	111.26	16.92	16.92	0.00
Auto-AVSR-S	Supervised	1,759	24.60	1.00	–	–	75.95	16.83	–	–
Auto-AVSR-L	Supervised	3,448	19.10	1.00	0.90	+0.10	70.48	17.26	16.69	+0.57

Note: $\Delta = AO - AV$, a positive delta indicates that the inclusion of visual information reduced the WER. All models were trained on LRS3 and VoxCeleb2, except Auto-AVSR-L, which included LRW, LRS2, LRS3, AVSpeech, and VoxCeleb2. Auto-AVSR-S AV omitted due to lack of public checkpoint.

modality-specific low-rank adaptation (LoRA) [38] projectors. It is trained on the same 1,759-hour corpus. Finally, **Auto-AVSR** [10] is a supervised approach designed to scale AVSR by leveraging automatically generated labels. To bypass the expensive process of manual annotation, this framework utilizes powerful pretrained audio-only ASR models to transcribe massive unlabeled datasets. We use Auto-AVSR in two configurations: **Auto-AVSR-S**, trained on the 1,759-hour corpus, and **Auto-AVSR-L**, trained on a 3,448-hour dataset across five corpora (LRW [39], LRS2, LRS3, AVSpeech [40], and VoxCeleb2 [34]).

Table I shows WER performance across video-only (VO), audio-only (AO), and audio-visual (AV) modalities. On LRS3, all models achieve strong AO performance (WER: 0.74–1.95%), reflecting the dataset's scripted and acoustically controlled nature. On Candor-LR, performance drops significantly (AO WER: 16.83–24.32%), showing that dyadic conversational speech in Candor-LR provides a much more difficult benchmark. Llama-AVSR maintains the best AO performance on both datasets (LRS3: 0.74, Candor-LR: 16.92).

VO performance is consistently poor on Candor-LR. On LRS3, Auto-AVSR-L achieves the best VO performance (19.10%) compared to AV-HuBERT (28.69%) and Llama-AVSR (26.20%), suggesting that larger and more diverse training data improves VO recognition on scripted speech. However, this advantage narrows on Candor-LR (Auto-AVSR-L: 70.48% vs AV-HuBERT: 78.35%), indicating that VO recognition is especially brittle for natural conversational speech, where spontaneous articulation creates a much more difficult setting than scripted speech.

The AV modality provides only marginal gains over AO on both LRS3 and Candor-LR. On LRS3, this is unsurprising given AO performance is already near-perfect, leaving little room for visual cues to contribute. On Candor-LR, absolute AV gains remain small ($\Delta \leq +1.56$), and given the already high AO WER, it is difficult to draw conclusions about the contribution of visual information from these zero-shot results alone. Llama-AVSR shows no visual benefit on either dataset ($\Delta = 0.00$ on Candor-LR, -0.05 on LRS3), suggesting its LLM decoder does not effectively leverage visual features. Auto-AVSR-L achieves the best overall Candor-LR performance (AV WER: 16.69%, $\Delta = +0.57$), yet its AO WER on Candor-LR (17.26%) is slightly higher than Auto-AVSR-S (16.83%), despite being trained on nearly double the data. This suggests that additional training data from diverse scripted sources does not necessarily improve AO performance on naturalistic conversational speech.

TABLE II: VO, AO, and AV WER (%) performance comparison across training sets (AV-HuBERT).

Test Set	VO	AO	AV	Δ
Trained on LRS3: 433 Hours				
LRS3	28.60	2.99	1.65	**+1.34**
LRS2	38.00	9.43	7.25	**+2.18**
Candor-LR	78.35	26.63	24.86	+1.77
Trained on LRS2+LRS3: 657 Hours				
LRS3	28.29	2.60	1.60	+1.00
LRS2	26.58	**3.63**	2.85	+0.78
Candor-LR	76.83	22.72	21.32	+1.40
Trained on Candor-LR: 713 Hours				
LRS3	39.03	6.88	6.02	+0.86
LRS2	47.09	10.72	9.15	+1.57
Candor-LR	69.60	**14.81**	10.50	+4.31
Trained on LRS2+LRS3+Candor-LR: 1370 Hours				
LRS3	**27.89**	**2.42**	**1.42**	+1.00
LRS2	**24.47**	3.76	**2.54**	+1.22
Candor-LR	**69.50**	16.37	**9.83**	**+6.54**

Note: $\Delta = AO - AV$; a positive Δ indicates that the inclusion of the visual modality reduced the WER. All models in this table were **trained without additional noise augmentation**, in contrast to the best AV-HuBERT checkpoint reported in Table I.

B. Cross-Domain Generalization and Multi-Dataset Training

We further analyze AV-HuBERT under different training data configurations to examine how Candor-LR affects cross-training generalization. AV-HuBERT was chosen for this because it has a comparatively lightweight decoding setup and lower computational requirements for training than both Auto-AVSR and Llama-AVSR. It also avoids the additional complexity introduced by an LLM-based decoder, while also showing the largest visual benefit on Candor-LR in Section IV-A. Table II reports VO, AO, and AV WER when the model is trained on LRS3 only, LRS2+LRS3, Candor-LR only, and the combined LRS2+LRS3+Candor-LR setup.

VO performance is consistently poor on Candor-LR across all training setups, showing that visual information alone is insufficient for this dataset. AO performance varies significantly by training data, ranging from 26.63% (LRS3-only) down to 14.81% (Candor-LR-only), with the combined setup falling in between at 16.37%. Training only on LRS3 gives strong in-domain performance on LRS3 (2.99% AO) but performance drops substantially on Candor-LR (26.63% AO). Adding LRS2 to training improves results slightly on Candor-LR (22.72%

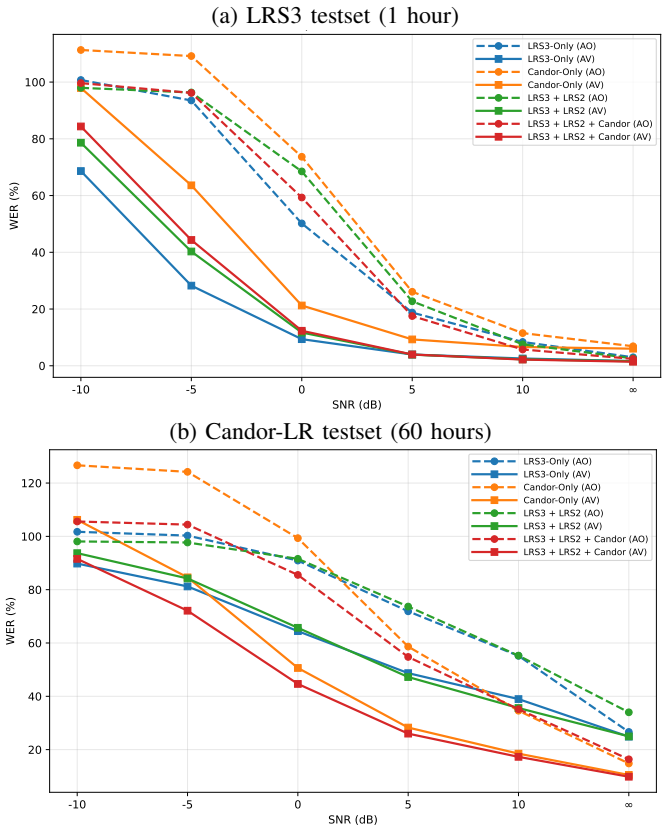


Fig. 4: WER across SNR levels for LRS3 and Candor-LR test sets when **trained without noise augmentation**, where ∞ dB denotes clean speech and different colors indicate training data combinations.

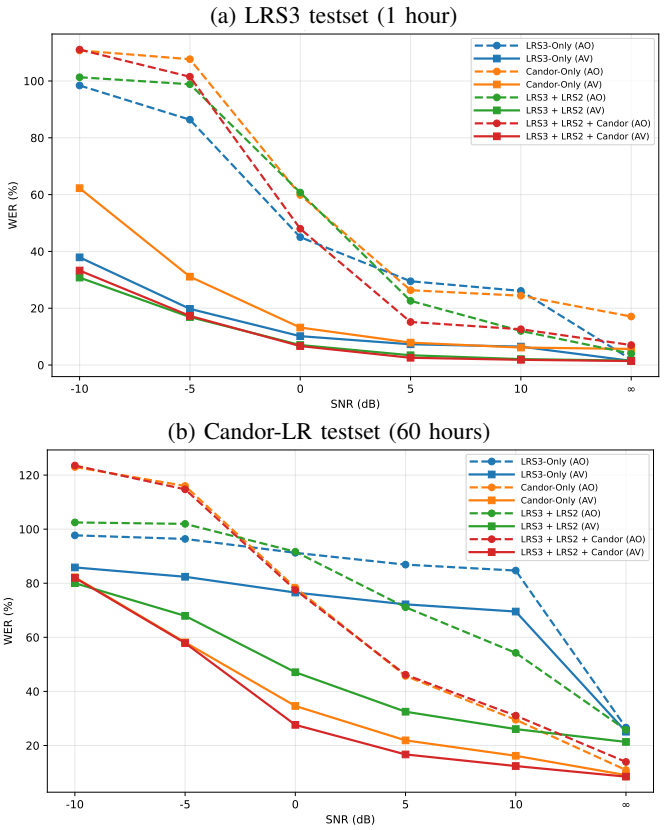


Fig. 5: WER across SNR levels for LRS3 and Candor-LR test sets when **trained with noise augmentation**, where ∞ dB denotes clean speech and different colors indicate training data combinations.

AO). Training on Candor-LR alone yields much better Candor-LR performance, with AO WER falling to 14.81% and AV WER to 10.50%. Lastly, the combined LRS2+LRS3+Candor-LR training gives the best overall Candor-LR result, reaching 9.83% AV WER. This setup also maintains strong performance on LRS2 (2.54% AV) and LRS3 (1.42% AV).

The visual benefit (Δ) on Candor-LR is smallest when trained on scripted data alone (+1.77 LRS3-only, +1.40 LRS2+LRS3) but rises sharply once Candor-LR is included in training (+4.31 Candor-LR, +6.54 combined). This shows that visual cues become substantially more important once the model is exposed to realistic conversational data, rather than simply scaling scripted data volume. In contrast, LRS3 visual gains remain stable across all training setups, showing that adding Candor-LR to training does not hurt out-of-domain performance on LRS3.

C. Noise Robustness for Candor-LR

To better understand the characteristics of the Candor-LR dataset under challenging acoustic conditions, we evaluate its performance trends under noise and compare them against the standard LRS3 benchmark. Using AV-HuBERT as our baseline model, we augment both datasets with synthetic babble noise. The noise generation follows the same procedure used by the AV-HuBERT authors for LRS3 [8], [41], where babble noise is synthesized from multiple overlapping speech sources from training data and added at different signal-to-noise ratio (SNR) levels (-10dB to $+10$dB). We experiment with two setups: applying this noise augmentation only at inference (with clean training), and incorporating it during both training and inference. Figure 4 and Figure 5 plot the WER performance across SNR levels for the inference-only noise and joint train-and-test noise configurations, respectively, across both the LRS3 and Candor-LR test sets. While performance predictably drops on both datasets as noise increases, the widening gap between them highlights the unique difficulty of Candor-LR.

1) Evaluation with Inference-Only Noise: As shown in Figure 4, WER decreases for both datasets as SNR increases, but Candor-LR shows a higher sensitivity to noise. In completely clean conditions (no added noise), the AO LRS3 model achieves a 2.99% WER on the LRS3 test set, while the Candor-LR model records a 14.81% WER on Candor-LR. At 10 dB SNR, LRS3 degrades slightly to 8.39% WER, whereas Candor-LR increases to 34.60% WER. At -10 dB SNR, LRS3 reaches 100.73% WER, while Candor-LR reaches 126.66% WER, indicating high insertion errors. This indicates that Candor-LR is inherently more vulnerable to noise.

Introducing the visual modality provides consistent improvements, but examining the gap between AO and AV performance reveals that visual cues become increasingly

important as noise increases. For models trained exclusively on in-domain data (i.e., trained and tested on the same corpus), incorporating the visual modality at 10 dB SNR reduces the Candor-LR WER by 16.11% (34.60% AO vs 18.49% AV). The gap widens at 0 dB SNR, where speech and noise are equally loud. At this level, the AO model on Candor-LR reaches a 99.33% WER, while the AV model maintains a 50.63% WER, highlighting a 48.7% WER difference bridged by visual data. At -10 dB SNR, visual modalities become necessary for Candor-LR to mitigate severe degradation. Expanding the training dataset to the combined LRS2+LRS3+Candor-LR corpus further validates this observation. This combined dataset yields a 17.29% AO WER on the Candor-LR test set at 10 dB. Yet, under the severe -10 dB condition, the model exhibits a 91.56% AO WER, showing that diverse, large-scale data alone cannot replace structured exposure to acoustic noise during training.

Cross-dataset generalization is highly asymmetric. Training exclusively on Candor-LR and testing on LRS3 yields a 6.62% AV WER at 10 dB (and 6.02 in clean conditions). Conversely, training on LRS3 and testing on Candor-LR yields a 38.96% AV WER at 10 dB. While the larger volume of training data in Candor-LR naturally contributes to this improved generalization, it also demonstrates that Candor-LR's combined scale and naturalistic distribution effectively encompass the more controlled LRS3 domain.

2) Evaluation with Joint Train-and-Test Noise: Building on this sensitivity to noise observed in Figure 4, we next examine whether exposing the model to noise during training can mitigate these effects. Following the noise augmentation approach used in AV-HuBERT, our second configuration trains with noise, randomly mixing babble noise into the training data with a 25% probability. This noise-trained setup substantially narrows the AO-AV performance gap observed earlier, as shown in Figure 5. Most notably, augmentation increases the model's reliance on visual information precisely when acoustic conditions deteriorate. For the LRS2+LRS3 model evaluated on the LRS3 test set at -10 dB, the AV WER drops from 78.62% in the clean-trained setup to 30.79% in the noise-trained setup, a relative improvement that far exceeds anything achieved by scaling data alone. A similar, though smaller, gain appears on Candor-LR, where AV WER improves from 93.70% to 80.03% under the same conditions. This result suggests that noise exposure during training does more than make the model robust to noise; it teaches the model to rely on noise-invariant visual features, compensating for audio degradation that clean-trained models cannot handle.

Interestingly, the data reveals a penalty for applying noise augmentation to narrow, single-domain datasets. When the LRS3-only model is evaluated on Candor-LR at 10 dB, the clean-trained version achieves a 38.96% WER, but the noise-trained version degrades sharply to 69.53%. This suggests that adding synthetic noise to a limited dataset causes the model to overfit to that specific acoustic distribution, actively hurting its ability to generalize to out-of-domain, naturalistic speech.

Expanding the training data to the full LRS2+LRS3+Candor-LR corpus removes this penalty and reveals a more nuanced relationship between data diversity and noise tolerance. At 10 dB SNR, combining diverse training data with noise augmentation drops the LRS3-only AV WER from 6.48% to a near-perfect 1.84%. For the Candor-LR test set, however, moving from the Candor-LR model to the combined model brings only a modest gain at 10 dB, from 16.19% to 12.40% AV WER. The smaller gain on Candor-LR indicates that scripted data diversity helps LRS3 more than it helps Candor-LR at moderate noise levels.

As noise increases to the -10 dB extreme, though, diverse training data combined with noise augmentation becomes critical for avoiding total model failure. The Candor-LR model trained without noise augmentation reaches a 106.20% AV WER at -10 dB, while the fully optimized model (combined data and noise training) reduces this to 82.00% AV WER. In-domain data is therefore necessary for handling unscripted conversational speech, but diverse out-of-domain data paired with noise augmentation becomes essential once noise interference gets severe. Overall, Candor-LR proves far tougher than LRS3 as a benchmark for evaluating model robustness in real-world conditions.

To provide a definitive baseline for future research, we summarize the best performance on the Candor-LR test set across all acoustic conditions in Table III.

TABLE III: Benchmark WER (%) comparison across SNR levels on the Candor-LR test set (trained on LRS2+LRS3+Candor-LR). ∞ represents clean audio.

	SNR (dB)					
Modality	**-10**	**-5**	**0**	**5**	**10**	∞
VO	69.50					
AO	123.54	114.74	77.53	46.17	31.04	13.92
AV	82.00	57.88	27.63	16.69	12.40	8.48

V. Conclusion

In this work, we provide an open-source pipeline to derive AVSR-ready data from the CANDOR dataset. Using this pipeline, we present Candor-LR, a large-scale AVSR benchmark for conversational settings built from natural dyadic videoconference data. It contains 713.5 hours of training data, 10.1 hours of validation data, and 60.1 hours of test data, capturing spontaneous turn-taking speech and realistic acoustic conditions that are largely absent from scripted datasets. We benchmark pretrained AVSR models and evaluate cross-domain generalization under varying training configurations and noise conditions. Our results show that current AVSR models struggle with unscripted, conversational speech, with visual modality and in-domain exposure playing a key role in narrowing this gap. Candor-LR offers a challenging real-world benchmark that moves AVSR beyond scripted datasets toward conversational speech understanding.

AI-Generated Content Disclosure

Gemini was used during the preparation of this work for minor grammatical edits and to enhance the clarity of the writing.

References

[1] K. R. Prajwal, T. Afouras, and A. Zisserman, "Speech Recognition Models are Strong Lip-readers," in *Interspeech 2024*, 2024, pp. 2425–2429.

[2] Y. J. Ahn, J. Park, S. Park, J. Choi, and K.-E. Kim, "SyncVSR: Data-Efficient Visual Speech Recognition with End-to-End Crossmodal Audio Token Synchronization," in *Interspeech 2024*, 2024, pp. 867–871.

[3] P. H. Seo, A. Nagrani, and C. Schmid, "Avformer: Injecting vision into frozen speech models for zero-shot av-asr," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22 922–22 931.

[4] M. Kit Khinn Teng, H. Zhang, and T. Saitoh, "Phoneme-level visual speech recognition via point-visual fusion and language model reconstruction," *arXiv e-prints*, pp. arXiv–2507, 2025.

[5] J. H. Yeo, H. Rha, S. J. Park, and Y. M. Ro, "MMS-LLaMA: Efficient LLM-based audio-visual speech recognition with minimal multimodal speech tokens," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 20 724–20 735. [Online]. Available: https://aclanthology.org/2025.findings-acl.1065/

[6] A. Rouditchenko, Y. Gong, S. Thomas, L. Karlinsky, H. Kuehne, R. Feris, and J. Glass, "Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation," in *Interspeech 2024*, 2024, pp. 2420–2424.

[7] J. H. Yeo, M. Kim, C. W. Kim, S. Petridis, and Y. M. Ro, "Zero-avsr: Zero-shot audio-visual speech recognition with llms by learning language-agnostic speech representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 6693–6703.

[8] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations (ICLR)*, 2022.

[9] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," in *arXiv preprint arXiv:1809.00496*, 2018.

[10] P. Ma, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[11] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, " Deep Audio-Visual Speech Recognition ," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 44, no. 12, pp. 8717–8727, Dec. 2022. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/TPAMI.2018.2889052

[12] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," *Journal of Phonetics*, vol. 48, pp. 1–12, 2015, the Impact of Stylistic Diversity on Phonetic and Phonological Evidence and Modeling. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447014000941

[13] J. Linke, B. C. Geiger, G. Kubin, and B. Schuppler, "What's so complex about conversational speech? a comparison of hmm-based and transformer-based asr architectures," *Computer Speech & Language*, vol. 90, p. 101738, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230824001219

[14] R. Jain and N. Harte, "The lipreading gap: Do vsr models perceive visual speech like human lipreaders?" 2026. [Online]. Available: https://arxiv.org/abs/2606.07435

[15] ——, "From hype to insight: Rethinking large language model integration in visual speech recognition," in *Proceedings of 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[16] Z. Lin and N. Harte, "Uncovering the visual contribution in audio-visual speech recognition," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[17] K. R. Prajwal, S. Hegde, and A. Zisserman, "Scaling multilingual visual speech recognition," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[18] W. Dai, S. Cahyawijaya, T. Yu, E. J. Barezi, P. Xu, C. T. Yiu, R. Frieske, H. Lovenia, G. I. Winata, Q. Chen *et al.*, "Ci-avsr: A cantonese audio-visual speech datasetfor in-car command recognition," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 6786–6793.

[19] Y. Wang, X. Meng, Y. Wang, J. Liang, Q. Liu, and D. Zhao, "Friends-mmc: A dataset for multi-modal multi-party conversation understanding," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, p. 25425–25433, Apr. 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/34731

[20] J. Zhao, Y. Jia, S. Wang, J. Zhou, H. Wang, and Y. Qin, "Chinese-lips: A chinese audio-visual speech recognition dataset with lip-reading and presentation slides," in *2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025, pp. 1–6.

[21] Y. A. D. Djilali, S. Narayan, E. LeBihan, H. Boussaid, E. Almazrouei, and M. Debbah, "Do vsr models generalize beyond lrs3?" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 6635–6644.

[22] D. Kwak, J. Choi, S. Lee, and J. S. Chung, "Lrs-voxmm: A benchmark for in-the-wild audio-visual speech recognition," *arXiv preprint arXiv:2604.27866*, 2026.

[23] T.-B. Nguyen, N.-Q. Pham, and A. Waibel, "Cocktail-Party Audio-Visual Speech Recognition," in *Interspeech 2025*, 2025, pp. 1828–1832.

[24] T.-B. Nguyen, K. Zmolikova, P. Ma, N. Q. Pham, C. Fuegen, and A. Waibel, "A cocktail-party benchmark: Multi-modal dataset and comparative evaluation results," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 19 502–19 506.

[25] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022. [Online]. Available: https://doi.org/10.1038/s42256-022-00550-z

[26] Y. Huang, J. Xue, L. Jiajun, D. Li, T. Zhang, Z. Yi, Y. Ren, and K. Li, "When avsr meets video conferencing: Dataset, degradation, and the hidden mechanism behind performance collapse," *arXiv preprint arXiv:2603.22915*, 2026.

[27] H. Chen, J. Du, Y. Dai, C.-H. Lee, S. M. Siniscalchi, S. Watanabe, O. Scharenborg, J. Chen, B. Yin, and J. Pan, "Audio-Visual Speech Recognition in MISP2021 Challenge: Dataset Release and Deep Analysis," in *Interspeech 2022*, 2022, pp. 1766–1770.

[28] L. T. P. Nguyen, Z. Yu, S. L. Y. Hang, S. An, J. Lee, Y. Ban, S. Chung, T.-H. Nguyen, J. Maeng, S. Lee, and Y. J. Lee, "See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, June 2026, pp. 2272–2283.

[29] A. Reece, G. Cooney, P. Bull, C. Chung, B. Dawson, C. Fitzpatrick, T. Glazer, D. Knox, A. Liebscher, and S. Marin, "The candor corpus: Insights from a large multimodal dataset of naturalistic conversation," *Science Advances*, vol. 9, no. 13, p. eadf3197, 2023. [Online]. Available: https://www.science.org/doi/abs/10.1126/sciadv.adf3197

[30] Speechmatics Ltd., "Speechmatics speech-to-text platform," https://www.speechmatics.com, 2026.

[31] S. O. Russell, I. Gessinger, A. Krason, G. Vigliocco, and N. Harte, "What automatic speech recognition can and cannot do for conversational speech transcription," *Research Methods in Applied Linguistics*, vol. 3, no. 3, p. 100163, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772766124000697

[32] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 5202–5211.

[33] Z. Lin, S. Petridis, M. Pantic, and N. Harte, "Assessing true generalisability of audio-visual speech recognisers," 2026. [Online]. Available: https://arxiv.org/abs/2606.07259

[34] J. S. Chung *et al.*, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.

[35] U. Cappellazzo *et al.*, "Large language models are strong audio-visual speech recognition learners," in *2025 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.

[36] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," in *in arXiv preprint arXiv:2307.09288*, 2023.

[37] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[38] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models." *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[39] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 87–103.

[40] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, Jul. 2018. [Online]. Available: https://doi.org/10.1145/3197517.3201357

[41] B. Shi, W.-N. Hsu, and A. Mohamed, "Robust Self-Supervised Audio-Visual Speech Recognition," in *Interspeech 2022*, 2022, pp. 2118–2122.