Visual-Aware Speech Recognition for Noisy Scenarios

Anonymous ACL submission

Abstract

Humans have the ability to utilize visual cues, such as lip movements and visual scenes, to enhance auditory perception, particularly in noisy environments. However, current Automatic Speech Recognition (ASR) or Audio-Visual Speech Recognition (AVSR) models often struggle in noisy scenarios. To solve this task, we propose a model that improves transcription by correlating noise sources to visual cues. Unlike works that rely on lip motion and require the speaker's visibility, we exploit broader visual information from the environment. This allows our model to naturally filter speech from noise and improve transcription, much like humans do in noisy scenarios. Our method re-purposes pretrained speech and visual encoders, linking them with multi-headed attention. This approach enables the transcription of speech and the prediction of noise labels in video inputs. We introduce a scalable pipeline to develop audio-visual datasets, where visual cues correlate to noise in the audio. We show significant improvements over existing audio-only models in noisy scenarios. Results also highlight that visual cues play a vital role in improved transcription accuracy.

1 Introduction

Automatic Speech Recognition (ASR) models have applications in many voice-enabled applications, including audio-video calls, intelligent virtual assistants, and media processing. These models are expected to work well in noisy conditions for their effective use in real-world scenarios. Several studies demonstrate that the human brain uses both audio and visual streams (e.g. lip motion, visual scenes) for listening, particularly when the speech is noisy (Sumby and Pollack, 1954; McGurk and MacDonald, 1976; Boots et al., 2020). These models have applications where the visual stream is also available as additional input. These observations have led to the development of audio-visual speech recognition (AVSR) models.

Several AVSR models show that transcription can be improved in the noisy scenario by attending to lip-region movement (Burchi and Timofte, 2023; Shi et al., 2022) and exploiting the correlation of visual scenes with spoken content (Seo et al., 2023). Recently Luo et al. (2024) show that background scenes can help in improving transcription in a given environment. However, its dependence on a manually collected dataset and limited alignment between visual context and audio hinder its scalability and effective utilization of visual cues.

Building on these insights, we address these limitations by proposing a scalable data creation pipeline and finetuning method that utilizes pretrained checkpoints. Our automated pipeline allows the mixing of audio-visual noise datasets with clean speech at variable noise ratios, eliminating the need for specialized datasets. In this work, we propose an architecture that integrates pretrained audio and visual encoders via Multi-Headed Attention. We hypothesize that training AVSR models with visual cues of the noise sources will improve speech recognition in noisy scenarios.

We use AudioSet (Gemmeke et al., 2017) mixed with a clean speech corpus, People speech (Galvez et al., 2021) for finetuning purposes. We extract speech embeddings for each time-step in audio and then calculate enhanced representations by attending to visual features obtained from CLIP visual encoder (Radford et al., 2021). Our model takes (audio, video) pairs and finetunes the speech encoder for multi-modal speech recognition and noise label prediction jointly using CTC loss (Graves and Graves, 2012). We hypothesize that leveraging the correlation between noise sources and visual cues will lead to more accurate transcription by providing richer context than background scene awareness alone.

The resultant finetuned model improves transcription quality while also predicting noise labels.

Ablation experiments further suggest that these improvements in transcription accuracy, are primarily due to our model's ability to attend to visual cues. The main contributions of this work are two-fold, (i) We propose a scalable dataset creation pipeline to develop audio-visual datasets, where visual cues correlate to noise sources in the audio. (ii) This work introduces a finetuning method that is visually aware of the noise while doing transcription. The dataset and code will be made publicly available. (iii) Finally, we present extensive ablation experiments to analyze our model.

2 Related Work

Audio only noisy speech recognition. Noise can be removed as a pre-processing step before being fed to ASR systems for improved transcription. Noise removal can be done either via signal enhancement techniques (Steinmetz et al., 2023) and via source separation methods (Rouard et al., 2023; Défossez, 2021; Petermann et al., 2023). Recent advancements have explored integrating speech enhancement modules directly into end-to-end ASR systems, allowing joint optimization for both enhancement and recognition tasks. This approach aims to mitigate the distortions introduced by separate enhancement stages and improve overall recognition performance in noisy environments (Zhu et al., 2022). However, purely audio-based models still face difficulties in extreme noise conditions, highlighting the need for multi-modal approaches, such as AVSR, which leverage visual cues to handle noise better.

Audio-visual Speech Recognition. Recent studies propose AVSR models capable of exploiting visual cues for improved performance. Multiple works have focused on exploiting lip motion as additional information along with audio to improve transcription (Shi et al., 2022; Huang and Kingsbury, 2013; Burchi and Timofte, 2023). In the context of full frame features, some works show that having visual cues related to the topics spoken helps with better word disambiguation (Gabeur et al., 2022; Seo et al., 2023). However these works only see visual information to correlate with actual spoken content, instead, we focus on exploiting visual context as a cognition enhancer for ASR systems.

3 Dataset Creation Pipeline

We aim to create a dataset where audio noise is closely correlated with the video content and each noise instance is uniquely annotated along ground truth transcriptions. To facilitate this, we have developed a dataset creation pipeline that selectively filters AudioSet (Gemmeke et al., 2017) for videos and corresponding noise audio with annotated labels. We then mix noise-labeled videos with the People's Speech dataset (Galvez et al., 2021), that have ground-truth transcriptions. Further details are discussed below.

Filtering AudioSet. *AudioSet* (Gemmeke et al., 2017) comprises of 2 million human-labelled, 10-second audio clips from YouTube, categorized into 632 audio event classes arranged hierarchically. This work targets only the videos associated with a noise label; thus, we exclude any video labelled with speech or human voice. We limit our scope to videos that only have a single noise label. We found that there is a big skew in the class distribution of noise labels, therefore we only select labels having at least 750 samples. This filtered subset of AudioSet has 44 unique noise labels (e.g. car, water, fireworks).

Mixing with People's Speech. *People's Speech* (Galvez et al., 2021) is an ASR dataset featuring 30K hours of transcribed English speech from a diverse range of speakers. We utilize clean subset of it for our dataset. Since AudioSet videos are of 10 seconds each, we select speech samples longer than 10 seconds and then trim both audio and transcripts. We take a clean speech sample and run an off-the-shelf forced aligner from the NeMo toolkit (Kuchaiev et al., 2019). The forced-aligned output provides word time stamps, allowing us to trim both audio and transcripts to a 10-second duration. We append the noise label as the final word to the transcripts, enabling the model to learn both transcription and noise label prediction for each sample. We process our filtered AudioSet (10-second video clips) and clean speech recordings to generate samples consisting of: video (without audio), corresponding noisy audio, clean speech, and corresponding transcripts. A noisy speech is obtained by mixing the clean speech recording with the original noisy audio extracted from the same video clip in a one-to-one correspondence.

Finally, we divide the dataset curated into train-

ing, validation, and testing subsets, ensuring each set contains a uniform distribution of noise samples from AudioSet. We refer to this dataset as the Visual-Aware Noisy Speech (*VANS*) dataset in further sections. The current VANS dataset contains 28K samples, providing 75 hours of training data, and 2K samples each contributing 6.1 hours for validation and testing. It is important to note that this dataset is scalable and can be expanded by incorporating more samples from AudioSet that may contain multiple labels, as well as more samples from People's Speech. Furthermore, we can enhance the dataset by dynamically altering the sample mixing mappings during model training to create augmentations.

4 Method

To enhance ASR robustness in noisy conditions, we adopt a late fusion approach inspired by recent multi-modal studies (Gabeur et al., 2022; Burchi and Timofte, 2023). Our model leverages a pretrained Conformer-based E2E ASR encoder[1] to extract audio embeddings $\mathbf{H_a}$ from noisy input speech. Visual features $\mathbf{H_v}$ are obtained using CLIP's ViT-L/14 image encoder (Radford et al., 2021). While both encoders are frozen, we enhance the speech encoder with trainable adapters. As shown in Figure 1, dense layers $\mathbf{W_A}$ and $\mathbf{W_V}$ project $\mathbf{H_a}$ and $\mathbf{H_v}$ into a shared space, producing $\mathbf{A_t}$ and $\mathbf{V_t}$ respectively. Formally,

$$\overline{\mathbf{A}_t} = \mathbf{W_A}\mathbf{H_a} + \mathbf{E_A^M} + \mathbf{E_A^T}, \quad (1)$$

$$\overline{\mathbf{V}_t} = \mathbf{W_V}\mathbf{H_v} + \mathbf{E_V^M} + \mathbf{E_V^T}. \quad (2)$$

$\mathbf{E_A^T}$ and $\mathbf{E_V^T}$ represent the positional embeddings for the audio and video time series, respectively. We use separate positional embeddings for audio and visual features to enhance the system's ability to track context across both modalities. Additionally, $\mathbf{E_A^M}$ for audio and $\mathbf{E_V^M}$ for video are modality embeddings, enabling the system to effectively distinguish between audio and visual information.

$\overline{\mathbf{A_t}}$ and $\overline{\mathbf{V_t}}$ from (1) and (2) are then passed through a standard transformer encoder block, facilitating Multi-Head Self-Attention across the modalities (Vaswani, 2017). This cross-modal interaction yields outputs $\mathbf{Z}_a$ for audio and $\mathbf{Z}_v$ for video respectively. For our task, we only utilize the visual-aware audio outputs $\mathbf{Z}_a$ and ignore $\mathbf{Z}_v$. $\mathbf{Z}_a$ is then processed through a convolutional decoder and then

[1] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_large

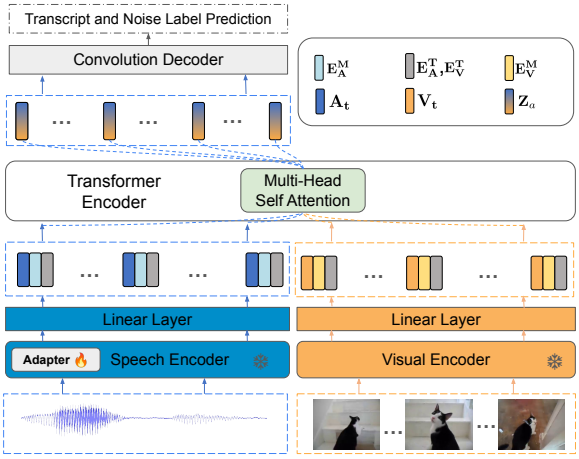

Figure 1: A visualization of our architecture. Speech and Visual representations are first obtained from their respective encoders, then aligned and enhanced via a Transformer-based Multi-Head Self-Attention mechanism. The output is then decoded using a convolutional decoder for simultaneous transcript and noise label prediction.

optimized for transcription task using standardized CTC loss. In our case, the last word in the transcripts refers to the noise label.

Base Model Pretraining. Existing ASR models and tokenizers typically include only transcription-related tokens, whereas our model requires the final token to represent noise label, which is not covered by the pretrained ASR tokenizer. Following (Karan et al., 2023), we extended the tokenizer to include special tokens for noise labels, necessitating the reinitialization of the prediction layer in the convolutional decoder. To adapt the model, we performed pretraining on 420 hours of People's Speech data using CTC loss and the extended tokenizer. This resulted in a pretrained speech encoder capable of jointly predicting the transcription followed by a noise label as last token in the final output.

5 Experiments & Results

Implementation details. Our experiments utilize a pretrained model, initially trained solely on transcription task without visual inputs, as described earlier. For visual information, we extract CLIP features at 5 fps. We use a Transformer Encoder with 4 layers with a dimensionality of 512. We assess model performance using Word Error Rate (WER) for transcription task and noise label prediction accuracy. For each prediction, we first strip away the noise label at the end, if present, and then compare the remaining transcript against the

ground truth transcript of the audio clip. We use the extracted noise label to evaluate the accuracy of the noise label prediction task.

Models. We conducted a series of experiments to demonstrate the improved performance of our model in noisy conditions by leveraging visual information. Thus, we selected 10dB SNR noisy speech samples for our experiments and train audio and audio-visual models. We recognize that it is impractical to train a separate model for each possible noise level, therefore we adopt a uniform sampling strategy to dynamically choose the SNR values in the range of -5 dB to +5 dB for each sample. This method, termed AV-UNI-SNR, ensures that our model encounters a varied but controlled set of noise scenarios, thus enhancing its ability to generalize across similar conditions.

5.1 Results

	Model	SNR (dB)	P_r	V_T	V_I	WER	ACC (%)
1	Conformer-CTC	-	-	-	-	26.99	-
2	A-SNR	10	✓	-	-	23.30	02.98
3	A-UNI-SNR	[-5,5]	✓	-	-	23.11	04.54
4	AV-SNR	10	✓	✓	✓	21.83	**60.95**
5	AV-SNR	10	-	✓	✓	23.59	58.59
6	AV-UNI-SNR	[-5,5]	✓	✓	✓	**20.71**	54.23
7	AV-UNI-SNR	[-5,5]	✓	✓	-	22.29	02.36

Table 1: Model Performance at SNR 10 dB. P_r refers to pretraining, V_T refers to visual information available during training, and V_I refers to visual information available during inference. "A" indicates models using only audio, while "AV" represents models utilizing both audio and video while training. "UNI" refers to models trained with uniformly sampled SNR levels. For details, please refer to section 5.1.

Table 1 presents the results of our experiments. On comparing R2 and R4 shows gains over the audio-only model in transcription accuracy with visual awareness. Notably, results depict a big gain in the correct prediction of noise labels when model learns to exploit cues from visual background. This proves our hypothesis that the correlation of noise with the visual cues helps with improved transcription and noise label predictions. The comparison between R4 and R5 shows the importance of pretraining, in preparing the model for both transcription and noise prediction tasks.

Results for AV-UNI-SNR models show the best performance overall. Performance gains are higher when visual information is provided at both fine-tuning and inference time. However, R7 vs R3 shows our model improves over the audio-only model even when visual information is not provided at inference time. This suggests that models trained with visual guidance for noise detection also perform well when only audio is used during inference. It shows that models trained with visual cues develop a more nuanced understanding of complex acoustic environments than audio-only models. However, it falls short in predicting noise labels without visual input. The model naturally tends to rely on video context for noise prediction, as it offers clearer cues. Consequently, when tested with only audio inputs, the model's performance on the noise prediction task declines. Details on across SNR results, baseline performances, ablations, and compute in section A.

	Models	LS test-clean	LS test-other
1	Conformer-CTC (Gulati et al., 2020)	31.07	39.89
2	A-UNI-SNR *(Ours)*	28.05	37.91
3	AV-UNI-SNR *(Ours)*	**27.86**	**37.47**

Table 2: Models Performance at SNR 0 dB on LibriSpeech (LS) Test Sets.

Out-of-Domain Evaluation. While AV-UNI-SNR is pretrained on People's Speech, and Conformer-CTC is pretrained on a broader range of datasets including People's Speech and LibriSpeech (Panayotov et al., 2015), there may be concerns that AV-UNI-SNR's superior performance on noisy audio is due to its specialized training on People's Speech. To address this, we conducted an additional experiment using LibriSpeech, mixed with AudioSet samples as described in section 3. Importantly, LibriSpeech is within the domain for Conformer-CTC but out-of-domain for our model. As shown in Table 2, our model still outperforms R1 and R2 on this dataset as well, confirming that R3 is robust in noisy environments even with out-of-domain data.

6 Conclusion

In this work, we show that exploiting visual cues with audio signals significantly improves transcription accuracy for noisy scenarios. Our automated dataset creation pipeline, designed to align noise with visual cues, provides a promising foundation for enhancing AVSR models. We show that models trained across varied SNR levels, especially the AV-UNI-SNR model, excel in diverse noise conditions. Our proposed method is easily adaptable to other pretrained architectures and checkpoints.

Limitations

While AudioSet provides a scalable foundation, the success of this approach relies heavily on its fine-grained noise-to-video correlations. These annotations, although extensive, are still manually curated and may not fully capture the complexity of real-world noisy environments. Incorporating visual inputs during inference introduces computational overhead, primarily due to the use of a pretrained CLIP visual encoder. While this overhead exists for achieving the best performance, our approach mitigates this by outperforming audio-only models even when used with only audio inputs during inference. However, for scenarios demanding the highest accuracy, the additional computational cost remains a trade-off.

References

Robert James Boots, Gabrielle Mead, Oliver Rawashdeh, Judith Bellapart, Shane Townsend, Jennifer D. Paratz, Nicholas Garner, Pierre Clement, David Oddy, and On behalf of the Circadian Investigators in Critical Illness. 2020. Circadian hygiene in the icu environment (chie) study. *Critical Care and Resuscitation*, 22:361 – 369.

Maxime Burchi and Radu Timofte. 2023. Audio-visual efficient conformer for robust speech recognition. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2257–2266.

Google Cloud. 2024. Gemini 2.0 flash. `https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash`. Accessed May 19, 2025.

Alexandre Défossez. 2021. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*.

Valentin Gabeur, Paul Hongsuck Seo, Arsha Nagrani, Chen Sun, Alahari Karteek, and Cordelia Schmid. 2022. Avatar: Unconstrained audiovisual speech recognition. In *Interspeech*.

Daniel Galvez, Gregory Frederick Diamos, Juan Ciro, Juan Felipe Cer'on, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *ArXiv*, abs/2111.09344.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Alex Graves and Alex Graves. 2012. Connectionist temporal classification. *Supervised sequence labelling with recurrent neural networks*, pages 61–93.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. *ArXiv*, abs/2005.08100.

Jing Huang and Brian Kingsbury. 2013. Audio-visual deep learning for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7596–7599.

Singla Karan, Jalalv Shahab, Kim Yeon-Jun, Ljolje Andrej, Daniel Antonio Moreno, Bangalore Srinivas, and Stern Benjamin. 2023. 1-step speech understanding and transcription using ctc loss. In *ICON*.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building ai applications using neural modules. *ArXiv*, abs/1909.09577.

Cheng Luo, Yiguang Liu, Wenhui Sun, and Zhoujian Sun. 2024. Multi-modality speech recognition driven by background visual scenes. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10926–10930.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264:746–748.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. 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.

Darius Petermann, Gordon Wichern, Aswin Subramanian, and Jonathan Le Roux. 2023. Hyperbolic audio source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *ArXiv*, abs/2212.04356.

Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid transformers for music source separation. In *ICASSP 23*.

Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. 2023. Avformer: Injecting vision into frozen speech models for zero-shot av-asr. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22922–22931.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdel rahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *ArXiv*, abs/2201.02184.

Christian J Steinmetz, Thomas Walther, and Joshua D Reiss. 2023. High-fidelity noise reduction with differentiable signal processing. *arXiv preprint arXiv:2310.11364*.

William H. Sumby and Irwin Pollack. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26:212–215.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Qiu-Shi Zhu, Jie Zhang, Zi-Qiang Zhang, and Li-Rong Dai. 2022. Joint training of speech enhancement and self-supervised model for noise-robust asr. *arXiv preprint arXiv:2205.13293*.

A Appendix

In this section, we present additional experiments and comparisons with state-of-the-art models across various SNR levels A.1, followed by ablation studies on key design choices A.2. We also analyze the computational costs of our AV model A.3, evaluate class-wise noise prediction accuracy A.4, and outline directions for future work A.5.

A.1 Results across SNRs

The results in Table 3 show that AV-UNI-SNR generalizes well across varying SNR levels, outperforming the individual models in lower SNR conditions (below -5 dB). However, models trained at fixed SNRs perform better at higher SNR values. These findings, along with the results from Table 1, suggest that training on variable SNR values, as in the AV-UNI-SNR model, enables robust performance across noisy conditions, and using visual cues further enhances generalization, even when visual cues are absent during inference.

The results in Table 3 demonstrate the effectiveness of our proposed models, particularly *AV-UNI-SNR*, in achieving consistently low WERs across a wide range of SNR conditions. Despite having significantly fewer parameters (453M) compared to models like Whisper Large V3 (1550M) and Whisper Medium (769M), AV-UNI-SNR delivers competitive or superior performance, especially in low SNR regimes (below 0 dB), where it achieves second-best performance at **20 dB**, **15 dB**, and **5 dB**.

Compared to larger models like Gemini (Cloud, 2024) and Whisper (Radford et al., 2022), which have access to significantly larger datasets and compute costs, our models are more parameter-efficient and adaptable. The AV-SNR variant, though trained at a fixed SNR level, still demonstrates strong performance and surpasses many larger baselines in mid-to-low SNR regions.

This robustness can be attributed to two key factors: (1) training the model across a uniform distribution of SNR levels rather than a fixed SNR value, and (2) incorporating visual modality acts as a additional guide and encourages the model to learn noise-invariant representations. This modeling is more effective than scaling model size alone for robustness in real-world noisy speech scenarios.

Training Details. Our AVSR model was trained for 10 epochs on a single L40S GPU with a batch size of 96, completing in approximately 8 hours. The model employs a 4-layer Transformer Encoder with 8 attention heads and a dimensionality of 512. Linear adapters with a dimensionality of 64 are incorporated into the speech encoder. For all other hyperparameters, we adhere to the NEMO toolkit defaults. We focused on CTC-based experiments in this project to prioritize training simplicity, modularity, and compatibility with external models for rapid prototyping and evaluation.

A.2 Ablations

#	Model	WER	ACC (%)
1	Conformer-CTC (Gulati et al., 2020)	26.99	-
2	AV-UNI-SNR (VL)	22.24	47.62
3	AV-UNI-SNR *(Start)*	20.85	54.07
4	AV-UNI-SNR	**20.71**	**54.23**

Table 4: WER and ACC (%) performance of ablation models at SNR 10 dB.

Table 4 presents ablation results assessing the impact of variable-length training and an alternative prediction strategy where the noise token is predicted as the start token of a transcript. By variable-length training, we refer to randomly cropping and injecting noise at varying positions within human speech samples, aiming to better simulate real-world conditions. However, as seen in Row 3

Model	# Params	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	-10 dB	-15 dB	-20 dB
Conformer-CTC	120M	21.53	23.13	26.99	34.42	47.34	64.58	79.53	87.65	92.06
AV-SNR *((Ours))*	453M	21.76	20.85	21.83	25.38	33.72	50.01	65.70	**78.53**	87.62
AV-UNI-SNR *Ours*	453M	18.50	19.08	20.71	24.96	33.83	50.06	68.52	80.48	87.42
Whisper Medium	769M	17.15	17.47	19.31	24.96	35.72	53.03	80.43	94.87	97.92
Whisper Large V3	1550M	**14.44**	**15.53**	**16.64**	**19.68**	**28.09**	47.31	71.11	85.22	90.22
Gemini 2.0 Flash	-	19.31	19.96	21.26	25.15	34.06	50.63	68.76	81.30	89.95
Gemini 1.5 Flash	-	19.16	19.82	20.95	25.00	31.26	**46.46**	**65.15**	78.76	**86.66**

Table 3: WER (%) performance of different models across varying SNR levels. **Bold** indicates the best (lowest) WER and underlined indicates the second-best per SNR level.

(VL), this approach slightly degrades performance. This is likely because, although AudioSet provides labels indicating the presence of certain noises in videos, it does not specify their exact temporal locations. As a result, random cropping may lead to segments that do not actually contain the intended noise, weakening the correlation between the audio and corresponding visual cues. This weaker alignment can also be observed in the drop in ACC, making it more challenging for the model to associate specific visual content with particular noise types.

Row 4 evaluates an alternative noise prediction method that places the noise token at the start of the transcription rather than the end (our default setup). Interestingly, this strategy yields performance comparable to the default. This is because both audio and visual tokens are contextualized using multi-head self-attention (MHSA), which allows tokens at all time steps to interact, making the position of the noise token less critical. Nonetheless, we choose to place the noise token at the end, as it is more suitable for potential online inference scenarios—where predictions occur in real time—making this setup more practical for deployment.

A.3 Computational costs?

	Models	Params	A	V	WER
1	Conformer-CTC Large	120M	✓	-	26.99
2	Conformer-CTC XLarge (XL)	635M	✓	-	26.15
3	A-UNI-SNR *(Large Backbone)*	150M	✓	-	23.11
4	A-UNI-SNR *(XL Backbone)*	665M	✓	-	22.34
5	AV-UNI-SNR *(Ours)*	453M	✓	✓	20.71
6	AV-UNI-SNR *(Ours)*	150M	✓	-	22.29

Table 5: Comparison of Models, Parameters, Modalities, and WER on Test Set of proposed dataset at 10dB.

We discuss the computational costs of our AV model in Table 5. Using visual inputs at inference requires an additional 300M parameters for CLIP feature extraction R5, increasing computational overhead compared to audio-only models. However, our AV-UNI-SNR model is flexible, supporting both audio-visual and audio-only inference. Notably, when used with only audio R6 it requires just 30M more parameters than the Conformer-CTC Large model (R1). Despite this smaller increase in parameters, our AV-UNI-SNR model outperforms the A-UNI-SNR XL model (R4), trained on audio-only data with 4x more parameters, demonstrating the superior efficiency and performance of our AV framework.

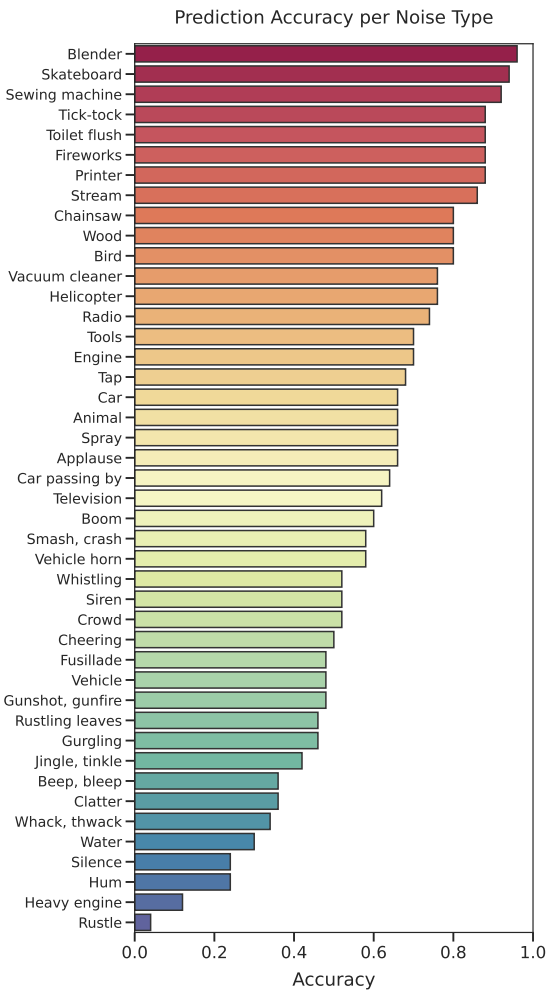

Figure 2: Prediction accuracy per noise type based on final token match between predicted and reference noise labels.

A.4 Noise Prediction Accuracy.

Figure 2 shows the model (AV-UNI-SNR) accuracy in predicting the correct noise type for each sample at 10dB SNR. We observe that categories like *Blender*, *Toilet flush*, and *Sewing machine* show higher accuracy—these objects are typically large, stationary, and visually distinct in their environments, making their presence easier to detect in frames. On the other hand, sounds such as *Rustle*, *Heavy engine*, *Hum* or *Whack, thwack* tend to be visually ambiguous or associated with smaller or transient objects (e.g., foliage, distant vehicles, or quick actions), reducing their visibility and making accurate association with the noise source more challenging for the model. This highlights the importance of clear visual grounding in achieving robust multimodal noise recognition.

A.5 Future Work

We plan to improve our model by exploring additional pretrained speech and visual encoder checkpoints and expanding our dataset pipeline to include AudioSet samples with multiple noise labels, enhancing visual context awareness. Furthermore, we plan to extend this approach to scalable audio-visual speech transcription, incorporating not only noise labels but also other visual cues and related events as tags.

Our framework discussed in section 3 has the potential to scale up and generate over 4000 hours of data by leveraging the full clean subset of People's Speech and AudioSet. This scalability enables the community to adopt and expand our approach for AVSR training, facilitating the development of models that leverage our AV training strategy. Such models could achieve superior performance with audio-only inputs at test time compared to those trained solely with audio.