Assessing ASR Robustness for Burmese: Impacts of Missing Speech Segments and Interruptions

Ankit Maurya[1] **Manikandan Ravikiran**[1] **Rohit Saluja**[1]
[1]Indian Institute of Technology Mandi
{s24090,erpd2301}@students.iitmandi.ac.in, rohit@iitmandi.ac.in

Abstract

The study explores the performance, robustness, and effects of automatic speech recognition systems when speech is missing or interrupted, with a specific focus on Burmese, a low-resource language. This study addresses several key research questions: How does missing or interrupted speech affect the accuracy of ASR? What is the link between the length of missing speech and the accuracy of the transcription? How are errors propagating when speech is masked or interrupted? By fine-tuning Wav2vec-bert2.0 and MMS-Zeroshot-300M (Massively Multilingual Speech) on a regular speech dataset (OpenSLR) of Burmese, the study answers these questions by evaluating the models on OpenSLR and 2 other datasets (FLEURS and Bloom) on common ASR metrics like Word Error Rate and Character Error Rate. The results reveal significant insights into error propagation, positional error patterns, and dataset-specific robustness. The study provides a baseline and methodological insights for future ASR research in interrupted settings for low-resource languages. The study's findings can inform the development of more robust ASR systems for real-world applications in low-resource languages.

1 Introduction

While automatic speech recognition (ASR) has enabled applications ranging from voice assistants (Dubiel et al., 2018; Sim et al., 2019) to automated transcription services (Jeffries et al., 2024), its performance often degrades under real-world conditions involving missing or incomplete speech segments (Barker et al., 2013; Gemmeke et al., 2011). Missing speech in these scenarios commonly arises from network packet loss during VoIP or streaming (Dissen et al., 2024; Lee and Kang, 2013; Kumalija and Nakamoto, 2022), recording interruptions caused by hardware issues or user-generated noise, and transmission errors from corrupted media. This challenge is particularly acute for low-resource languages (LRLs), which have not seen the same focus on robustness as high-resource languages (Baevski et al., 2020; Rubenstein et al., 2023; Radford et al., 2022), raising important questions about ASR reliability in these contexts.

Accordingly, this study focuses on Burmese, a Tibeto-Burman language spoken by approximately 42.9 million people and notably underrepresented in ASR research (Wikipedia contributors, 2025b; Li and Jian, 2024). The linguistic complexity of Burmese makes it a compelling case for robustness analysis. As a tonal language with an agglutinative morphology (Wikipedia contributors, 2025a), meaning is conveyed through subtle changes in pitch and duration, while grammatical information is often encoded in extended word forms. Consequently, the limited data available for training robust models poses a critical challenge, as even brief gaps in audio can lead to significant loss of semantic and grammatical information. Enhancing ASR robustness for Burmese can thus bridge communication gaps and enable technological inclusion for millions of speakers. To investigate these challenges, this paper aims to:

- Quantify the performance impact of missing speech on a fine-tuned Burmese ASR model;
- Perform a detailed error analysis to understand the nature and distribution of errors caused by missing data; and
- Evaluate model robustness across multiple relevant datasets.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes our experimental setup, including the datasets and methodology. Finally, Section 5 concludes the paper and outlines future work.

2 Related Works

The challenge of handling missing audio in automatic speech recognition (ASR) has long been a tough problem. In the past, researchers mostly focused on techniques like robust feature extraction and model adaptation to reduce the effects of noise and distortions (Ming and Crookes, 2014). More recently, the focus has shifted to what's called speech inpainting or speech reconstruction, which involves filling in or estimating missing segments of audio. These methods range from simple interpolation approaches (Kauppinen et al., 2001) to more advanced models like Gaussian Mixture Models (GMMs) (Cooke et al., 2001) and deep learning techniques using autoencoders or Generative Adversarial Networks (GANs) (Wali et al., 2021).

Another related idea comes from self-supervised learning, where models are trained to predict masked parts of the input data. This was first popularized in natural language processing by models like BERT (Devlin et al., 2019) and later adapted for speech in models like Wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021). These models learn robust representations that help in handling missing data. Though primarily used for pre-training, masking is also used as a data augmentation strategy to make models more robust (Rebuffi et al., 2021). One recent example is SpeechPainter (Zalan Borsos and Matthew Sharifi and Marco Tagliasacchi, 2022), which uses a generative diffusion model to do high-quality speech inpainting.

However, most of these efforts have focused on high-resource languages. In our study, we explore how masking and gapping two strategies for simulating missing speech, differently affect ASR error patterns and robustness, specifically for Burmese.

3 Experimental Setup

This section outlines the experimental framework designed to assess the performance and robustness of ASR systems for Burmese with missing speech.

3.1 Dataset

To evaluate the impact of missing speech on Burmese ASR, we utilize publicly available Burmese speech datasets, representing different domains, recording conditions, and potentially varying levels of annotation quality. The primary datasets considered are: (a) **OpenSLR Burmese (SLR80)** (Oo et al., 2020): Contains approximately 4 hours of read speech, used here for fine-tuning and evaluation. Characteristics often include relatively clean recordings. (b) **Few-shot Learning Evaluation of Universal Representations of Speech (FLEURS)** (Conneau et al., 2023; Goyal et al., 2022):An n-way parallel dataset with 12 hours of speech per language, representing a more diverse source. (c) **Bloom Dataset** (Leong et al., 2022): Contains 1 hour of burmese read speech from book paragraphs. Table 1 presents the dataset splits we use in our study.

Dataset	Split Used	No. of Samples
OpenSLR	Test + Val	196+206
FLEURS	Test	880
BLOOM	Test + Val	50+50

Table 1: Dataset and Split Configuration for the Study.

To analyze ASR model robustness to incomplete speech, we simulate missing data using two controlled techniques: masking and gapping. These are applied over specified durations $D_m \in \{0.25\text{s}, 0.5\text{s}, 0.75\text{s}, 1\text{s}\}$, chosen to reflect real-world interruption lengths, such as network packet loss in VoIP systems (e.g., 10–40 packets of 25ms) (Zhang et al., 2024). Let an original discrete-time audio signal be $S_{orig}[n]$, with N_{orig} samples. The gap/mask duration D_m corresponds to N_D samples. If N_{orig} is not a multiple of N_D, $S_{orig}[n]$ is zero-padded at the end to form a signal $S[n]$ of length N, where N is the smallest multiple of N_D such that $N \geq N_{orig}$. Subsequent operations refer to this signal $S[n]$ of length N.

Masking Masking simulates data loss by setting a segment of N_D samples to zero amplitude, with the masked signal $S'_{mask,k}[n]$ retaining the length N. For each signal $S[n]$, N/N_D distinct masked versions are generated. The k^{th} masked signal is formed by zeroing out samples in the segment $n \in [kN_D, (k+1)N_D - 1]$, where $k \in [0, \ldots, (N/N_D) - 1]$. This is described by:

$$S'_{mask,k}[n] = \begin{cases} S[n] & \text{for } 0 \leq n < kN_D \\ 0 & \text{for } kN_D \leq n < (k+1)N_D \\ S[n] & \text{for } (k+1)N_D \leq n < N \end{cases}$$

This technique is analogous to how lost data packets replaced by silence are handled and is fundamental in applying time-frequency masks in speech enhancement (Kim, 2021). Alternatively, it is an element-wise multiplication $S'_{mask,k} = S \odot M_k$, where $M_k[n]$ is a binary mask.

Gapping Gapping simulates interruptions by inserting a silent segment of N_D zero-valued samples into $S[n]$, thereby extending its total duration to $N + N_D$ samples. For each signal $S[n]$, $(N/N_D) + 1$ distinct gapped versions are generated. In the j^{th} gapped signal, $S'_{gap,j}[m]$, silence is inserted at a position corresponding to jN_D in the timeline of $S[n]$, where $j \in [0, \dots, N/N_D]$. The resulting signal $S'_{gap,j}[m]$ for $m \in [0, N+N_D-1]$ is constructed as:

$$S'_{gap,j}[m] = \begin{cases} S[m], & \text{for } 0 \le m < jN_D, \\ 0, & \text{for } jN_D \le m < jN_D + N_D, \\ S[m-N_D], & \text{for } jN_D + N_D \le m < N + N_D \end{cases}$$

This method alters the signal's temporal structure, testing the model's ability to handle unexpected pauses. The key distinction is that masking overwrites existing audio data while preserving effective signal duration (N), whereas gapping inserts new silent data, increasing the overall signal length and specifically challenging resilience to pauses and temporal shifts.

3.2 Models and Parameters

In this work, we selected Wav2Vec-BERT-2.0 (Chung et al., 2021) and MMS-Zeroshot-300M [1] (Pratap et al., 2024) due to their extensive multilingual pre-training and strong performance on LRLs. Wav2Vec-BERT-2.0's prior exposure to Burmese and MMS-Zeroshot-300M's training on over 1,000 languages make them suitable for this study. We fine-tune publicly available checkpoints for our experiments. This study fine-tunes both models on the OpenSLR dataset for up to 16 epochs using the AdamW optimizer (learning rate 5e-5, batch size 8), utilizing pretrained checkpoints available on Hugging Face (Face, 2025) and utilizing the Transformers library for the fine-tuning process. For our experiments, we select the best-performing fine-tuned checkpoints of both Wav2Vec-BERT-2.0 and MMS-Zeroshot-300M on the OpenSLR validation set with the best validation WER and use it for all subsequent experiments. All subsequent evaluations use the data splits detailed in Table 1. All the experiments were done on a single NVIDIA RTX A5000 GPU.

3.3 Evaluation Metrics

We assess ASR performance using the following metrics to quantify transcription errors and to analyze the impact of missing speech:

- **WER (Word Error Rate)**: Measures word-level transcription errors.
- **CER (Character Error Rate)**: Measures character-level transcription errors.
- **Error Percentage (Error %)**: The proportion of samples where the transcription differs from the clean audio baseline after simulating missing speech.
- **Edit Distance Distribution**: Examines how severe the errors are by analyzing the distribution of edit distances for each affected sample compared to the baseline.
- **Positional Error Analysis**: Looks at where errors tend to occur, focusing on the first and last segments around the missing speech region.

Dur.	**Masking**		**Gapping**	
	Samples	**Err. (%)**	**Samples**	**Err. (%)**
0.25s	9144	69.14	9546	77.92
0.50s	4458	84.28	4860	77.72
0.75s	2924	88.68	3326	78.35
1s	2134	91.38	2536	74.80

Table 2: Total processed samples and error percentage (%) for Masking vs. Gapping on the OpenSLR dataset by duration. (Model Used: Wav2Vec-BERT-2.0)

4 Results and Analysis

Across all three datasets, OpenSLR, Fleurs, and Bloom, our experiments highlight that the impact of missing speech (via masking and gapping) is closely tied to baseline dataset characteristics and duration of gap/mask. Trends across datasets are captured in Appendix Figures 7, 8 (overall WER/CER), Figure 1 (OpenSLR edit distance), Figures 4, 5 (Fleurs), Figures 3, 6 (Bloom), and Tables 3 and 4.

OpenSLR (Appendix Figures 1, 2, 7): This relatively clean dataset exhibits clear, progressive increases in both WER and CER as mask duration grows. Masking is notably more detrimental than gapping, with Wav2Vec-BERT-2.0's CER rising from 6.79% to 21.02% at 1s (compared to 8.89% for gapping). Edit distance distributions (Figure 1) confirm more severe character-level errors for longer mask durations. Positional analysis

[1] The *MMS-Zeroshot-300M* model checkpoint was used for fine-tuning. The name denotes the pre-trained model and does not imply a zero-shot evaluation methodology in our experiments.

(Figure 2) reveals that masking the final segment of an utterance dramatically increases edit distances compared to initial-segment masking, underscoring the importance of utterance-end information for ASR. This can be attributed to the model's need for forward and backward context to correctly decode speech. When the end of an utterance is masked, the model loses critical cues for disambiguation, which is especially important for a tonal and agglutinative language like Burmese where meaning and grammatical information can be conveyed by subtle changes at the end of words. The absence of this final information results in a significantly higher character-level error rate. Error percentages (Table **??**) mirror these patterns, especially for masking (69.14% to 91.38% with duration).

Fleurs (Appendix Figures 4, 5): Fleurs is substantially more challenging, with baseline WERs already above 179% and CERs over 21% even without missing speech. Masking consistently elevates CER (Wav2Vec-BERT-2.0: 21.26% to 25.35%), while WER shows only minor, sometimes negative shifts likely due to error saturation. Edit distance distributions (Figure 4) and average edit distances (Figure 5) confirm more fine-grained character errors from masking, especially at utterance ends. Gapping causes a severe but consistent level of degradation across all gap durations, with WER and CER remaining largely stable (Wav2Vec-BERT-2.0 WER ≈184%, CER ≈22%). Error percentages (Table 4) remain extremely high (>90%) across all gap/mask types and durations.

Bloom (Appendix Figures 3, 6): The Bloom dataset poses the greatest challenge, with baseline CERs over 66% and WERs exceeding 150%. Neither masking nor gapping meaningfully alters the already saturated error rates (CER ≈67-68%, WER ≈150-160%). Edit distance distributions and error percentages (near 100% for all conditions) confirm that the models nearly uniformly fail on this dataset, irrespective of the missing speech scenario. Overall, our key takeaways include:

- **Masking** generally introduces more severe character-level errors (as seen in CER and edit distance shifts), particularly on cleaner datasets like OpenSLR and moderately so on Fleurs. This highlights that real-world ASR systems deployed in environments with short-duration or partial occlusions (e.g., coughs, short microphone dropouts) are likely to see disproportionately larger transcription errors, especially on simpler, cleaner audio inputs.

- **Gapping** causes more stable but consistently high errors, with less sensitivity to gap/mask duration. This suggests that silent gaps or short audio losses (e.g., packet loss in VoIP or poor connectivity) might degrade performance consistently across a range of scenarios, rather than in a duration-dependent manner.

- **Dataset Difficulty Dominates:** On challenging datasets (Fleurs, Bloom), extremely high baseline errors overshadow the incremental effects of missing speech, leading to error saturation. This indicates that for real-world ASR robustness, improving baseline model performance (e.g., adapting to domain-specific vocabularies, reducing dataset-domain mismatch) is critical, as gap/mask effects become secondary when baseline transcription itself is unreliable.

These insights reinforce that efforts to improve ASR robustness in real-world scenarios must prioritize both baseline domain adaptation and gap or mask specific resilience, especially for character-level fidelity and utterance-end information crucial for downstream tasks.

5 Conclusion

In conclusion, this study reveals significant challenges in applying any ASR system to LRLs when encountering simulated missing speech segments. Our findings indicate a clear degradation in performance, measured by WER and CER, as the duration or ratio of missing data increases. Notably, the structure of the missing data matters; our results suggest that masking the audio signal generally causes more severe degradation than inserting gaps. Furthermore, the errors induced by these missing segments are not localized; they often propagate beyond the immediate vicinity of the gap or mask, primarily manifesting as deletions and substitutions, which aligns with observations from edit distance analysis. Finally, the system's resilience is not uniform, varying significantly across the different datasets tested (OpenSLR, Fleurs, Bloom), underscoring the influence of acoustic conditions, speaking styles, and domain specificity on robustness. These results highlight the need for targeted strategies to improve the robustness of Burmese ASR systems against various forms of missing speech data encountered in real-world scenarios.

Limitations

This study's limitations include the simulation of missing speech (zero-masking, gapping), which may not fully mirror real-world data loss complexities. Findings are also specific to the chosen pre-trained models (Wav2Vec-BERT-2.0, MMS-Zeroshot-300M) potentially differing for other ASR architectures or training methods. Furthermore, the employed Burmese datasets, while valuable, may not encompass the language's complete dialectal or acoustic diversity, which could affect the broader generalizability of the observed robustness levels.

Future Works

Future work will focus on three key areas. First, we will explore advanced speech inpainting and reconstruction techniques to better mitigate the effects of missing speech. Second, we plan to investigate more dynamic and realistic interruption patterns, such as randomly distributed or non-uniform segment lengths, to provide a more accurate assessment of ASR robustness. Finally, we will consider a broader range of Burmese datasets to improve the generalizability of our findings. We also plan to evaluate models with shorter durations, such as 0.10s, to capture more nuanced effects on performance.

Acknowledgments

We thank Bharat Electronics Limited (BEL) for funding the research work.

Ethics Statement

This research exclusively utilizes publicly available speech datasets (OpenSLR, Fleurs, and Bloom) intended for academic use, and no new data was collected from human subjects. The primary goal of this work is to foster positive technological inclusion by analyzing and improving ASR robustness for the low-resource Burmese-speaking community. The findings are specific to the models and datasets employed, which may not encompass the full linguistic and dialectal diversity of the Burmese language, potentially leading to performance disparities across different speaker populations. The computational cost required to train and run these large models also represents an environmental consideration.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Jon Barker, Emmanuel Vincent, Ning Ma, Heidi Christensen, and Phil Green. 2013. The pascal chime speech separation and recognition challenge. *Computer Speech Language*, pages 621–633.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Martin Cooke, Phil Green, Ljubomir Josifovski, and Ascension Vizinho. 2001. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Yehoshua Dissen, Shiry Yonash, Israel Cohen, and Joseph Keshet. 2024. Enhanced asr robustness to packet loss with a front-end adaptation network. *ArXiv*, abs/2406.18928.

Mateusz Dubiel, Martin Halvey, and Leif Azzopardi. 2018. A survey investigating usage of virtual personal assistants. *ArXiv*, abs/1807.04606.

Hugging Face. 2025. Wav2vec2-bert model documentation. `https://huggingface.co/docs/transformers/en/model_doc/wav2vec2-bert`. Accessed: 2025-04-25.

Jort F. Gemmeke, Maarten Van Segbroeck, Yujun Wang, Bert Cranen, and Hugo Van Hamme. 2011. Automatic speech recognition using missing data techniques: Handling of real-world data. In *Robust Speech Recognition of Uncertain or Missing Data*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

Nat Jeffries, Evan King, Manjunath Kudlur, Guy Nicholson, James Wang, and Pete Warden. 2024. Moonshine: Speech recognition for live transcription and voice commands.

I. Kauppinen, Jyrki K. Kauppinen, and Pekka Saarinen. 2001. A method for long extrapolation of audio signals. *Journal of The Audio Engineering Society*, 49:1167–1180.

Gibak Kim. 2021. Review of time–frequency masking approach for improving speech intelligibility in noise. *IETE Technical Review*, 39:1–12.

Elhard Kumalija and Yukikazu Nakamoto. 2022. Performance evaluation of automatic speech recognition systems on integrated noise-network distorted speech. *Frontiers in Signal Processing*, 2:999457.

Min-Ki Lee and Hong-Goo Kang. 2013. Speech quality estimation of voice over internet protocol codec using a packet loss impairment model. *The Journal of the Acoustical Society of America*, 134:EL438–44.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yaxuan Li and Yang Jian. 2024. Low-resource burmese speech synthesis based on visual text embedding and diffusion model. In *Proceedings of the 2024 4th International Conference on Artificial Intelligence, Automation and Algorithms*, AI2A '24, page 34–40, New York, NY, USA. Association for Computing Machinery.

Ji Ming and Danny Crookes. 2014. Speech enhancement from additive noise and channel distortion - a corpus-based approach. In *Interspeech*.

Yin May Oo, Theeraphol Wattanavekin, Chenfang Li, Pasindu De Silva, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, Oddur Kjartansson, and Alexander Gutkin. 2020. Burmese Speech Corpus, Finite-State Text Normalization and Pronunciation Grammars with an Application to Text-to-Speech. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6328–6339, Marseille, France. European Language Resources Association (ELRA).

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. 2021. Data augmentation can improve robustness. In *Neural Information Processing Systems*.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023. Audiopalm: A large language model that can speak and listen.

Khe Chai Sim, Petr Zadrazil, and Françoise Beaufays. 2019. An investigation into on-device personalization of end-to-end automatic speech recognition models. In *Interspeech 2019*. ISCA.

Aamir Wali, Zareen Alamgir, Saira Karim, Ather Fawaz, Mubariz Ali, Muhammad Adan, and Malik Mujtaba. 2021. Generative adversarial networks for speech processing: A review. *Computer Speech & Language*, 72:101308.

Wikipedia contributors. 2025a. Burmese grammar — Wikipedia, the free encyclopedia. [Online; accessed 18-May-2025].

Wikipedia contributors. 2025b. Burmese language — Wikipedia, the free encyclopedia. [Online; accessed 2-April-2025].

Zalan Borsos and Matthew Sharifi and Marco Tagliasacchi. 2022. SpeechPainter: Text-conditioned Speech Inpainting. In *Interspeech 2022*, pages 431–435.

Zihan Zhang, Jiayao Sun, Xianjun Xia, Chuanzeng Huang, Yijian Xiao, and Lei Xie. 2024. Bsplcnet: Band-split packet loss concealment network with multi-task learning framework and multi-discriminators. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 23–24. IEEE.

A Appendix

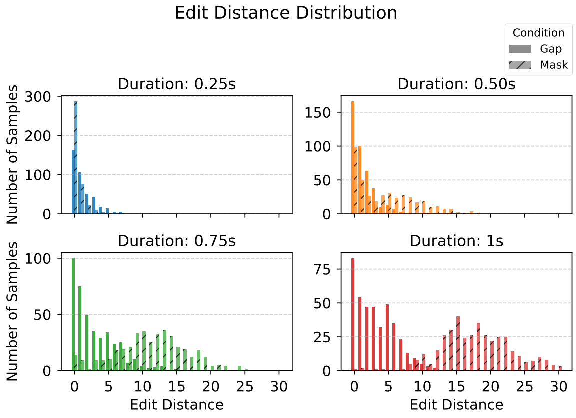

Figure 1: Edit distance distribution across samples for OpenSLR dataset. (Model: Wav2Vec-BERT-2.0)

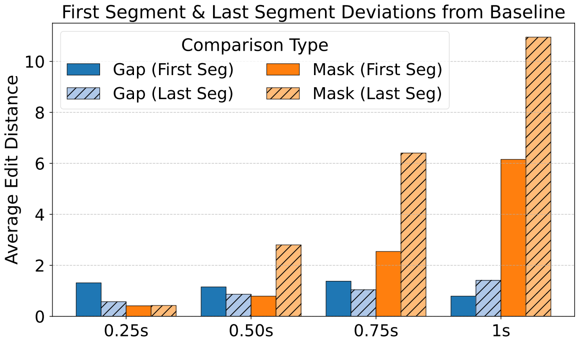

Figure 2: Average Edit distance for First and Last segment of mask and gap audio segments. (Model= Wav2Vec-BERT-2.0, Dataset = OpenSLR) (Baseline = Prediction without mask/gap)

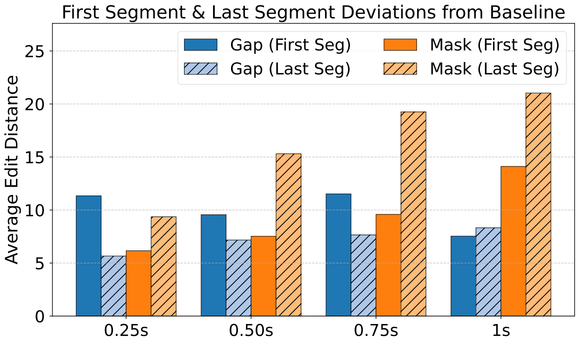

Figure 3: Average Edit distances for First segment and Last Segment. (Model = Wav2Vec-BERT-2.0, dataset = Bloom)

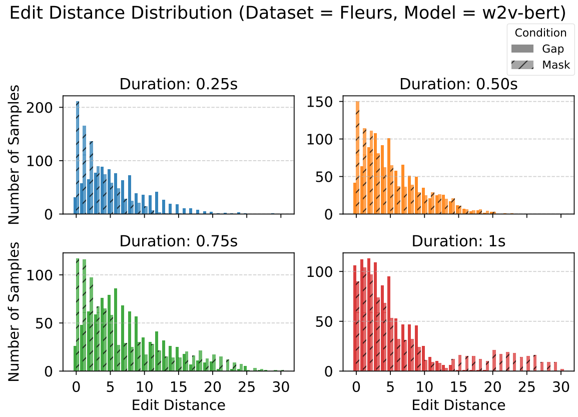

Figure 4: Edit Distance distribution of Fleurs dataset. (Model: Wav2Vec-BERT-2.0)

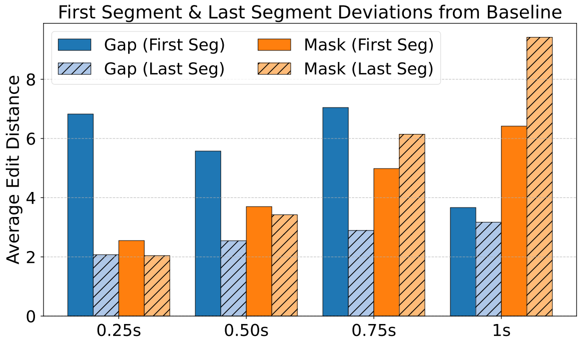

Figure 5: Average Edit distances for First segment and Last Segment. (Model = Wav2Vec-BERT-2.0, dataset = Fleurs)

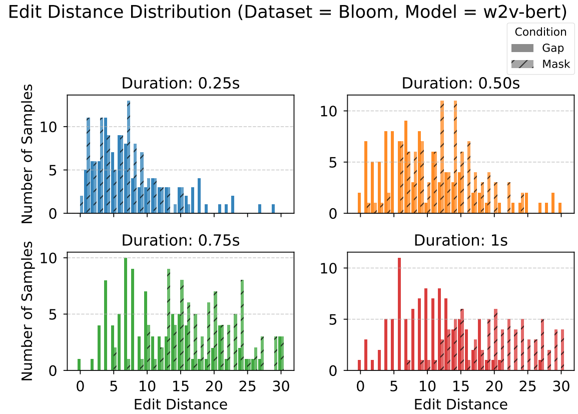

Figure 6: Edit Distance distribution of Bloom dataset. (model = Wav2Vec-BERT-2.0)

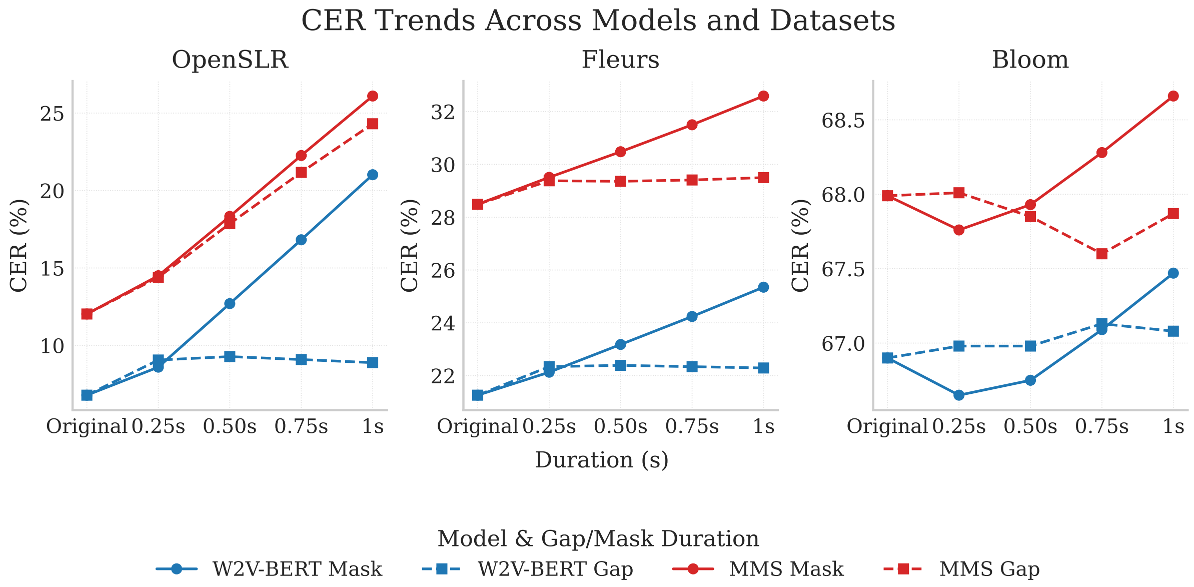

Figure 7: CER Trend for all datasets for different gap and mask duration for both models.

Table 3: Detailed WER (%) and CER (%) Results by Duration for Masking vs. Gapping Across Datasets and Models. Models were fine-tuned (FT) on OpenSLR.

			Masking		**Gapping**	
Dataset	**Model**	**Duration of mask/gap**	**WER (%)**	**CER (%)**	**WER (%)**	**CER (%)**
OpenSLR	Wav2Vec-BERT-2.0 (FT)	0s	37.35	6.79	37.35	6.79
		0.25s	40.85	8.60	41.39	9.06
		0.50s	44.75	12.70	41.29	9.28
		0.75s	47.69	16.82	41.53	9.09
		1s	50.46	21.02	41.12	8.89
	MMS-Zeroshot-300M (FT)	0s	56.37	12.03	56.37	12.03
		0.25s	59.94	14.50	59.78	14.40
		0.50s	61.74	18.33	61.29	17.86
		0.75s	63.81	22.26	63.00	21.17
		1s	65.58	26.10	64.38	24.31
Fleurs	Wav2Vec-BERT-2.0 (FT)	0s	181.00	21.26	181.00	21.26
		0.25s	182.60	22.13	184.00	22.34
		0.50s	180.10	23.18	184.00	22.39
		0.75s	177.00	24.24	184.00	22.34
		1s	174.40	25.35	184.00	22.29
	MMS-Zeroshot-300M (FT)	0s	179.24	28.49	179.24	28.49
		0.25s	178.50	29.51	180.29	29.38
		0.50s	175.42	30.48	179.90	29.36
		0.75s	172.34	31.50	179.48	29.41
		1s	169.05	32.59	178.95	29.50
Bloom	Wav2Vec-BERT-2.0 (FT)	0s	182.45	66.90	182.45	66.90
		0.25s	183.66	66.65	186.10	66.98
		0.50s	180.94	66.75	185.32	66.98
		0.75s	178.19	67.09	185.22	67.13
		1s	175.10	67.47	186.24	67.08
	MMS-Zeroshot-300M (FT)	0s	159.27	67.99	159.27	67.99
		0.25s	157.45	67.76	159.27	68.01
		0.50s	154.67	67.93	159.21	67.85
		0.75s	151.84	68.28	159.01	67.60
		1s	148.55	68.66	158.13	67.87

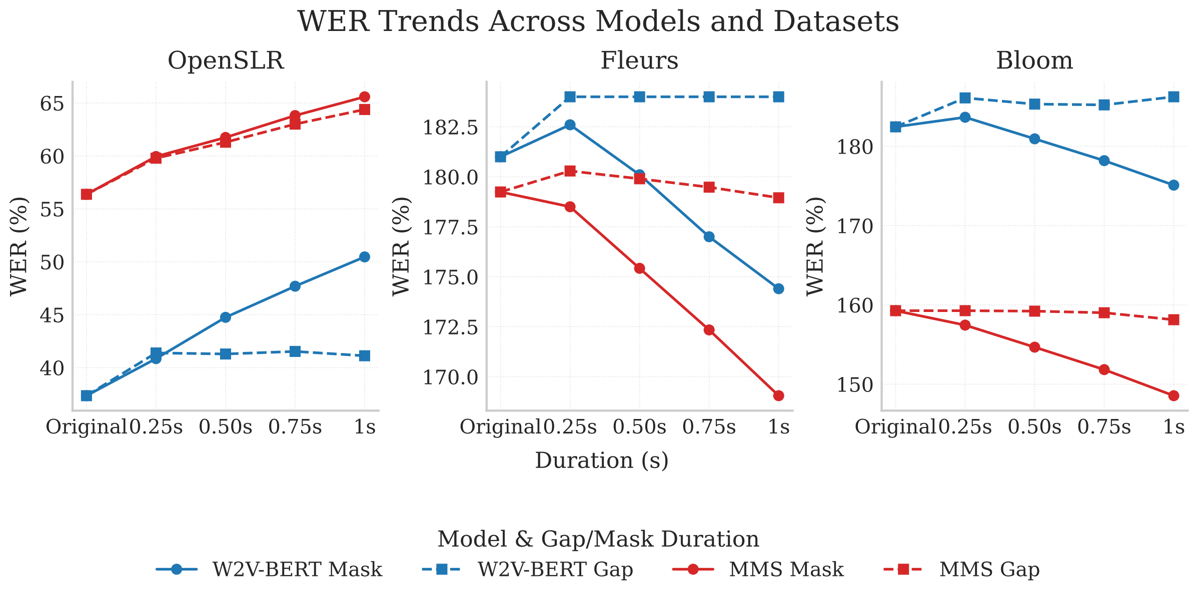

Figure 8: WER Trend for all datasets for different gap and mask durations for both models.

Table 4: Total Processed Samples and Percentage of Samples with Errors/Prediction Changes Across Datasets, Models, and Conditions. Models were fine-tuned (FT) on OpenSLR.

			Masking		Gapping	
Dataset	**Model**	**Duration of mask/gap**	**Total Samples**	**Error (%)**	**Total Samples**	**Error (%)**
OpenSLR	Wav2Vec-BERT-2.0 (FT)	0.25s	9144	69.14	9546	77.92
		0.50s	4458	84.28	4860	77.72
		0.75s	2924	88.68	3326	78.35
		1s	2134	91.38	2536	74.80
	MMS-Zeroshot-300M (FT)	0.25s	9144	79.01	9546	76.79
		0.50s	4458	87.01	4860	82.88
		0.75s	2924	90.94	3326	84.97
		1s	2134	92.83	2536	86.24
Fleurs	Wav2Vec-BERT-2.0 (FT)	0.25s	54 515	88.93	55 395	94.00
		0.50s	27 037	94.02	27 917	94.63
		0.75s	17 882	95.76	18 762	95.81
		1s	13 293	96.83	14 173	93.69
	MMS-Zeroshot-300M (FT)	0.25s	54 515	82.24	55 395	91.27
		0.50s	27 037	88.77	27 917	87.45
		0.75s	17 882	91.98	18 762	94.89
		1s	13 293	94.21	14 173	92.85
Bloom	Wav2Vec-BERT-2.0 (FT)	0.25s	3108	99.81	3208	99.75
		0.50s	1524	99.93	1625	99.75
		0.75s	1008	99.90	1108	99.55
		1s	737	100.00	838	99.40
	MMS-Zeroshot-300M (FT)	0.25s	3108	97.52	3208	99.13
		0.50s	1524	99.61	1625	98.58
		0.75s	1008	99.90	1108	99.46
		1s	737	100.00	838	99.52