Context Length of Speech and Text for ASR

Hyunho Ahn
ahn.377@osu.edu

Hansol Lee
lee.11514@osu.edu

Shakhrul Iman Siam
siam.5@osu.edu

Abstract

Automatic Speech Recognition (ASR) systems often process audio in short segments, limiting their ability to leverage broader context. To explore this, the paper systematically explores how increasing both audio and textual context length affects ASR performance. We evaluate multiple architectures—including RNN-T, FastConformer, and multimodal models like Whisper and Qwen2—Audio across a range of context windows from a few seconds up to fifteen minutes. Empirical results on both short and long-form English context, as well as a Korean lecture dataset, reveal that longer context windows can significantly reduce transcription errors and improve coherence. However, excessive context sometimes saturates or even harms performance due to computational overhead and error propagation. Our findings highlight the importance of carefully balancing context length to maximize ASR performance while mitigating potential drawbacks.

1 Introduction

Automatic Speech Recognition (ASR) systems have become integral to a wide range of applications, from voice assistants to live video transcription services. ASR models usually process short audio clips of about 30 seconds or less, which makes it hard to use broader context effectively (Flynn and Ragni, 2023). This lack of context can lead to transcription errors, such as misinterpreting homophones (e.g., "their" vs. "there"), struggling with pronoun references, or losing coherence in long-form speech. Without previous context, ASR systems may also fail to recognize speaker intent, misattribute dialogue, or inconsistently transcribe names and technical terms. Incorporating context can significantly enhance transcription accuracy by allowing the model to reference prior words and phrases, leading to more coherent and accurate outputs (Tang and Tung, 2024). This study explores how increasing the length of audio context will affect ASR performance. With the questions of relevance between the length and context (both audio and text), we hypothesize that processing longer speech segments and providing extended context will reduce transcription errors (e.g., WER) and improve recognition quality. However, we also anticipate that beyond a certain threshold, there is a point where adding more context no longer benefits accuracy and would even degrade performance. Our experiments measure these effects to determine the optimal context length for different ASR applications, balancing accuracy with efficiency.

2 Related Work

In natural language processing, many studies have demonstrated that incorporating additional context improves model performance. In-context learning, for example, enables models to better understand a task by leveraging few-shot prompts (Dong et al., 2022). Similar to these context-aware strategies in text-based LLMs, researchers have explored various methods for incorporating additional context into ASR tasks to address the inherently limited information capacity of short, localized speech segments (Li et al., 2022; Fox et al., 2024).

Focusing on additional speech-based context, (Tsunoo et al., 2019) proposed subsidiary context embedding vectors in the transformer architecture, allowing the model to utilize phonetic information over the inputs. (Flynn and Ragni, 2023) demonstrated that trained with long-form audio has an advantage compared to the short-form only trained models in the task of transcription. Meanwhile, (Huang et al., 2022) introduced a Recurrent Neural Network Transducer (RNN-T) model capable of predicting semantically meaningful segments before transcribing them, thereby capturing more context from each audio portion. Also, (Jia et al., 2025) explored the efficient attention window size

and CTC-aligned methods that can represent the context in the speech utilizing the LLM model.

Complementary work has explored incorporating textual context alongside the audio. (Chang et al., 2023) proposed three ways of fusing the acoustic and text information in the RNN-T architecture for the ASR task. (Lakomkin et al., 2024) utilized the video description and titles along with the audio tokens for the input of the LLM decoder for recognizing the speech. Similarly, (Chen et al., 2024) explored in-context learning methods that can be used in ASR tasks, utilizing the keywords in the speech. Closely related to our work, (Radford et al., 2023) proposed the model Whisper, and showed the effectiveness of feeding the previously transcribed text for the long-form transcription, but it does not clearly explain the length of the text, and the maximum speech input is limited to 30 seconds. The capability of prompting of Whisper is evaluated in the research (Yang et al., 2024; Cheng, 2024), but they are focused on the prompts, not considering the length of the speech input.

Beyond explicit context injection, some architectures inherently have temporal dependencies. For instance, (Graves et al., 2006, 2013; Graves and Jaitly, 2014) leveraged recurrent neural networks to capture sequential patterns in speech. More recently, the Conformer model (Gulati et al., 2020) has proven effective by combining self-attention for global context states with convolutional layers that capture local detail, enabling simultaneous modeling of short- and long-range dependencies in speech.

While previous research has demonstrated the benefits of contextual information in ASR, to the best of our knowledge, no study has systematically analyzed how the length and modality (textual and acoustic) of recursive context affects inference performance. Our work aims to address these gaps by evaluating the impact of recursive context length on ASR performance under inference-time constraints.

3 Methodology

In this section, we describe the proposed methodologies for evaluating the performance of the ASR task.

3.1 Research Objectives

The primary objective of this study is to evaluate how varying the length and modality of context, specifically audio and textual inputs, affects the performance of Automatic Speech Recognition (ASR) models during inference. While prior work has shown that additional context can improve transcription quality, a systematic understanding of how much context is beneficial and whether there is a point of diminishing or negative returns remains underexplored.

To this end, we focus on the following key research questions: **Q1**: How does increasing the length of preceding audio input affect the transcription accuracy of ASR models? **Q2**: How does incorporating textual context (e.g., previous transcriptions or prompts) influence ASR performance? **Q3**: Do different types of ASR models (e.g., audio-only vs. audio+text multimodal) respond differently to increasing context? **Q4**: What are the computational trade-offs (e.g., inference latency, memory usage) associated with using longer context?

To answer these questions, we experiment with a wide range of context window sizes, from very short segments (as little as 2 seconds) to extended sequences up to 15 minutes of continuous audio. For models that support textual inputs, we also vary the length of prior transcribed text provided during inference. This allows us to study how both temporal and semantic context impact model accuracy, robustness, and scalability.

3.2 Dataset Selection & Preprocessing

The primary focus of the projects is open-sourced English datasets. Large-scale non-English speech datasets, such as HKUST (Liu et al., 2006) and CSJ (Maekawa et al., 2003), exist; however, these datasets are difficult to verify as a non-native speaker of these languages. However, to evaluate the multi-lingual performance and quantify the model with the diverse datasets, we also tried using the Korean dataset (Kim et al., 2021). Some of the datasets are in-house, and we exclude these cases because of the limitation of the accesses (Fox et al., 2024). The ASR dataset can be divided into two categories: short-form and long-form. Short-form dataset is made up of utterances that are pre-segmented. Long-form is not segmented and contains real-word-like pauses during the conversations. We created Short-form datasets from long-form corpora by partitioning them into small segments. This approach allows to maintain a consistent dataset source while examining different context lengths. The TED-LIUM dataset (Hernandez et al., 2018) employed for one of base long-form

Dataset Name	Language	Source Domain	Notes / Purpose
TED-LIUM 1	English	TED Talks	Realistic pauses, segmented from talks
Earnings-22	English	Corporate earnings calls	Multiple speakers, segmented by words
AIHub Korean Lectures	Korean	Academic lectures	Used to analyze language-general effects

Table 1: Overview of speech datasets with various languages and domains

datasets, as it contains extensive speech segments reflective of real-world pauses. This dataset is created from the TED talks, contains about 118 hours of speech. The Earnings-22 dataset (Del Rio et al., 2022), derived from corporate earnings calls, is included in our experiments due to its realistic long-form speech content and detailed annotations. A key advantage of this dataset is its precise word-level timestamps, which help avoid partial-word segmentation during preprocessing and enable accurate alignment between audio and transcription. Additionally, a Korean lecture dataset from AI Hub (Kim et al., 2021) is incorporated to include a non-English corpus. This dataset is delivered in sentence- or word-level segments and can easily be merged to create fully contextualized long-form audio or split into very short segments, providing flexibility in examining how context length impacts ASR across different languages.

3.3 Context Window Design

To systematically evaluate how varying amounts of prior context influence ASR performance, four levels of context windows based on audio duration are defined in Table 2.

Context Length	Time Range
Extremely Short Audio Context	0–5 sec
Short Audio Context	5–30 sec
Medium Audio Context	0.5–5 min
Long Audio Context	5–15 min

Table 2: Context length categories and corresponding time ranges.

3.4 Models

The ASR models used in the experiments are categorized into two main groups based on the type of input they accept.

3.4.1 Models with Audio + Text Input

For these type of model, Whisper (Radford et al., 2022) and Qwen2-Audio (Chu et al., 2024) are used in the experiment which are Large Language Model (LLM)-Based ASR models. These models leverage transformer-based encoder-decoder architectures and can be explicitly conditioned on both audio and textual context and provide a framework to investigate how explicit context (either past audio or text) influences transcription performance, especially in long-form speech.

3.4.2 Models with Audio-Only Input

For audio-only input models, traditional ASR architectures such as the Recurrent Neural Network Transducer (RNN-T) (Jain et al., 2020) and NVIDIA FastConformer model (Rekesh et al., 2023) are used. These models are designed to process streaming audio and inherently capture temporal dependencies through their internal mechanisms—such as recurrence or self-attention—without relying on explicit textual feedback. While they do not accept external textual context, they are capable of accumulating information over time within the model itself. This implicit context modeling enables them to maintain coherence in longer utterances and makes them effective for real-time or low-latency ASR scenarios.

4 Evaluation and Analysis

In this section, we evaluate the performance of the ASR models across varying input lengths and analyze the results.

4.1 Experimental Setup

The experiments perform inference with separate models in each context condition, discussed in the section 3. For data preparation, in each context length setting, we build data pairs (audio, ground-truth transcription of the audio). To save time, we randomly select a total of 10 hours of audio from the Earnings-22 dataset. We obtain the pre-trained model from Nvidia Nemo[1] and Hugging Face Hub[2]. To be specific, Qwen2-Audio (abbreviated as Qwen2 in figures) and (abbreviated as Whis-

[1]https://github.com/NVIDIA/NeMo
[2]https://huggingface.co

per) is from Hugging Face, and FastConformer (abbreviated as Conformer) and RNN-T is from Nvidia Nemo platform. The raw text output (transcription) for each segment is saved for subsequent evaluation.

4.2 Metrics

Given the output of the model, which is transcribed text of speech input, we use the following metrics to measure its performance.

Word Error Rate (WER) The ratio of incorrectly transcribed words to the total number of words in the reference transcription. This metric reflects the ability of the model accurately understands the words in the speech.

Character Error Rate (CER) While WER measures errors at the word level, CER is calculated based on character-level differences between the inference and the reference. Unlike WER, CER is particularly useful when the speech contains mispronunciations or minor spelling errors.

Inference Speed and Memory Usage Inference speed is measured as the time taken to generate each token during decoding. Memory usage is recorded for each input size to assess the scalability of the models. These metrics help us understand the system requirements and limitations associated with using longer context windows.

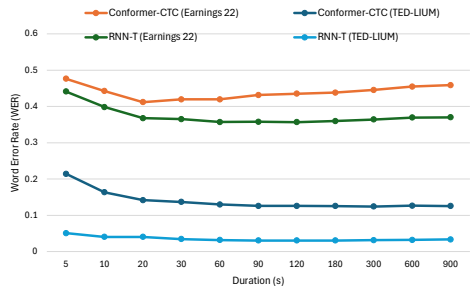
4.3 Effect of Audio Context Length

We evaluate how varying the duration of the input audio affects ASR performance across different models. Using fixed-length audio chunks, we observe a general trend of improved WER with longer context. However, Figure 1 and 2 shows that extremely short segments (1–5 seconds) tend to suffer from unnatural word splits and this leads to artificially high error rates. Figure 1 shows the impact of varying input audio duration on ASR performance (WER and CER) using the NVIDIA FastConformer and RNN-T model. The figure indicates that performance begins to saturate after an audio duration of 20 and 90 seconds, accordingly for the FastConformer and RNN-T. Similarly, Figure 2 illustrates the effect of varying input audio duration using the Whisper-small model, where performance saturation is observed after 20 seconds. Unlike Qwne2-Audio, Whisper model was not able to operates over 30 seconds. While the other model handles the 10-second input range relatively robustly, Qwen2-Audio shows increased variability in performance. This may be attributed to differences in model architecture or the way it processes the audio segments and bias in the training data. Figure 3 shows the output tokens per second for Whisper-small and NVIDIA FastConformer models with varying audio duration. For Whisper-small, the output tokens per second increase with longer audio. On the other hand, NVIDIA FastConformer produces significantly more output tokens per second compared to Whisper-small, it peaks at mid-range audio durations and then decreases for longer audio. Figure 5 indicates GPU peak memory usage with varying audio duration for both models. Memory usage for NVIDIA FastConformer remains almost constant for shorter intervals but rises sharply beyond 300 seconds, becoming impractical for longer audio durations due to excessive memory demands. Figure 4 illustrates the effect of varying input time-frame lengths on the transcription performance of audio-only models. Inference time is measured as the total processing time required to transcribe the entire audio, segmented into 1-second chunks. When using shorter time-frame (e.g., 1s), the increased number of segmentations leads to frequent ASR model invocations, causing significant computational overhead. Additionally, we observe saturation effects in model performance depending on the input time-frame length: for Whisper, performance tends to saturate around 20-second windows, while for FastConformer and RNN-T, saturation is observed at approximately 90 seconds. This suggests that longer input durations can improve efficiency up to a certain point, beyond which further increasing the time-frame does not yield meaningful gains, or even worsen. Figure 5 illustrates the GPU peak memory usage across different input time-frame lengths. We observe that memory consumption increases proportionally with the length of the input time-frame. At shorter time-frames, there are occasional spikes in memory usage, which we attribute to instability caused by the frequent invocation of the ASR model. This suggests that excessively small input window can lead to inefficient and unstable memory behavior during inference.

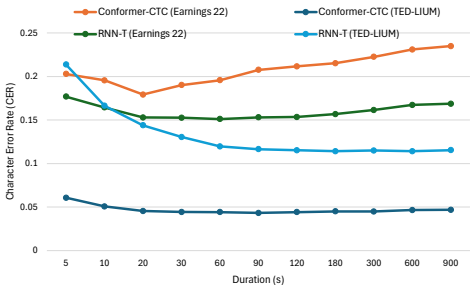
4.4 Effect of Textual Context

For models capable of feeding both audio and text inputs (i.e., Whisper and Qwen2-Audio), we investigate whether feeding previous transcription results improves recognition in the audio.

The figure 6 presents the Word Error Rate (WER) and Character Error Rate (CER) under different prompt settings. Across all configurations,

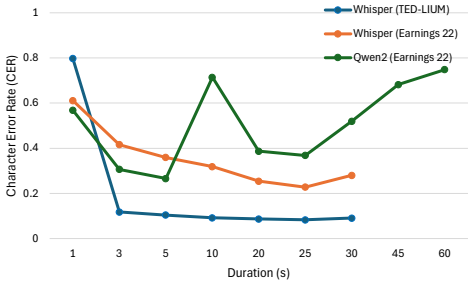


(a) Word Error Rate (WER)

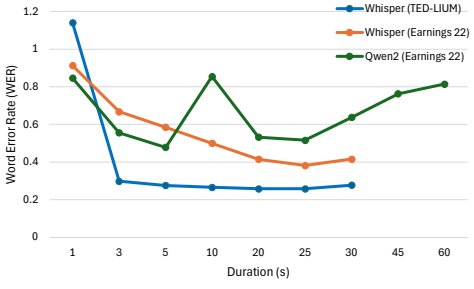


(b) Character Error Rate (CER)

Figure 1: Performance vs context length for TED-LIUM and Earning 22 dataset using NVIDIA Fast Conformer and RNN-T model.

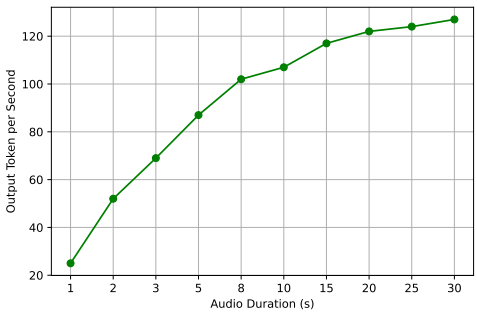


(a) Character Error Rate (CER)

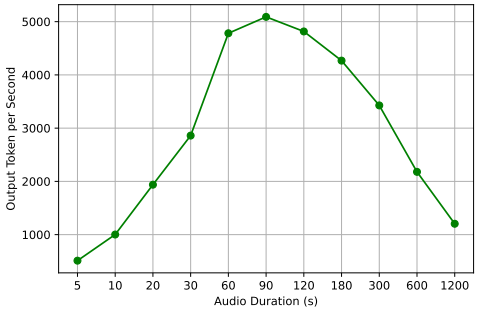


(b) Word Error Rate (WER)

Figure 2: Performance vs context length for TED-LIUM and Earnings 22 dataset using Whisper-small and Qwen2-Audio model.

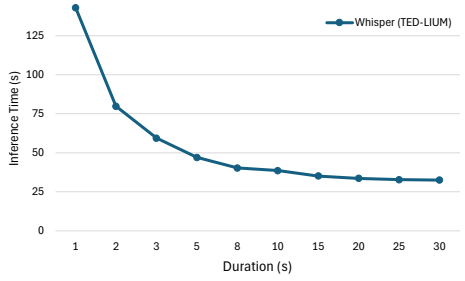


(a) Whisper-small

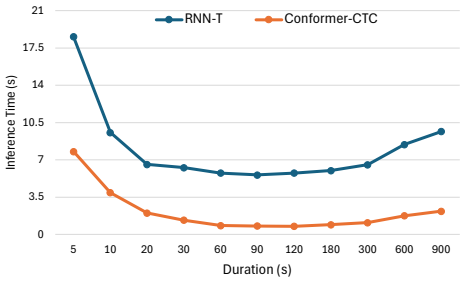


(b) NVIDIA FastConformer

Figure 3: Output Token generated per second. All experiments were done in NVIDIA A100 80GB GPU.



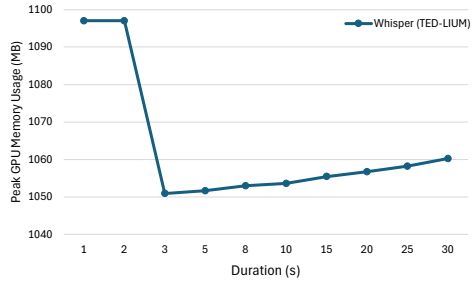
(a) Whisper-small



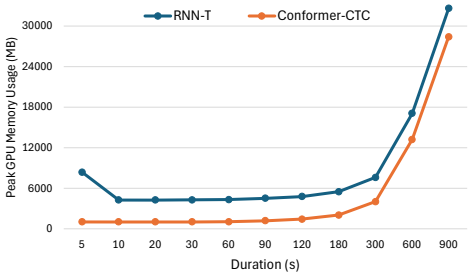
(b) NVIDIA FastConformer and RNN-T

Figure 4: Inference Time with varing audio durations. All experiments were done in NVIDIA A100 80GB GPU.

we observe that feeding prior text prompts does not outperform the baseline where only speech input is provided. However, supplying longer textual context generally leads to a gradual improvement in WER and CER values. Notably, at early steps (e.g., 1 or 3 seconds), transcriptions often include short, possibly error-prone segments, and as the number of steps increases, these early transcription errors accumulate, negatively impacting performance. In the case of Whisper, due to the model's maximum



(a) Whisper-small



(b) NVIDIA FastConformer and RNN-T

Figure 5: Peak GPU Memory usage with varying audio duration. All experiments were done in NVIDIA A100 80GB GPU.

input length constraint, we were unable to evaluate prompts with a large number of steps (e.g. 3 step on 20 seconds), as the combined input exceeded the allowable time frame. For Qwen2-Audio, when prompts ranging from 1 to 3 seconds were used, the generation process took an unusually long time. This was likely due to error accumulation during decoding, as the model failed to fully capture all words in the audio, leading to progressively longer outputs. Therefore, we excluded these results from our evaluation.

4.5 Cross-Language Evaluation

To evaluate the generalization performance of context-aware ASR models beyond English, we utilized a Korean lecture dataset from AI Hub (Kim et al., 2021). This dataset comprises lectures averaging approximately 30 minutes, provided as sentence- or word-level segments. This structure offers flexibility in constructing audio data with varying context lengths by merging consecutive segments.

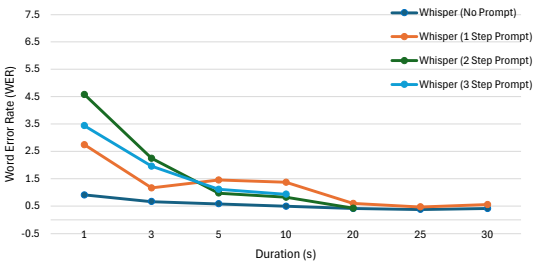
However, the segmented nature of the dataset imposed limitations on using arbitrary lengths of preceding textual context, unlike some approaches explored in English experiments. Therefore, to investigate the effect of textual context in the Korean setting, we experimentally adopted a strategy of injecting the transcription from the immediately preceding 3-second or 5-second segment as context before evaluating WER and CER on the target segment.

The experimental results (Figure 7) revealed an interesting pattern. Injecting preceding textual context yielded improved WER and CER performance only for extremely short, 1-second target segments. Conversely, for target segments longer than one second, this method of context injection resulted in performance degradation.

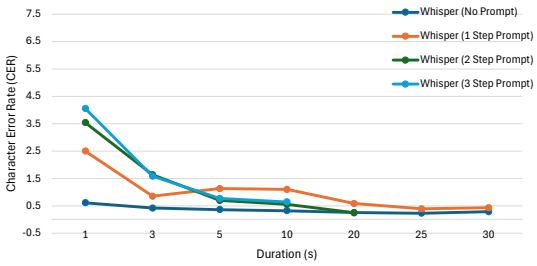
We hypothesize the reason for the performance improvement observed specifically for 1-second segments is as follows. Direct analysis of these 1-second segments revealed a high frequency of simple number pronunciations. The Korean language features two distinct numeral systems (Native Korean and Sino-Korean), and number pronunciations can be homophonic with other monosyllabic words. This linguistic characteristic can make it challenging for the ASR system to discern whether a short utterance is a number, which numeral system it belongs to, or if it represents a different word entirely. While analogous to the need to distinguish 'to' from 'two' in English, the dual numeral system in Korean potentially increases the likelihood of such confusion by effectively doubling the possibilities for numerical homophones. Consequently, we postulate that injecting the short (3-second or 5-second) preceding context aided in disambiguating the meaning of these short, often ambiguous numerical utterances found in 1-second segments, leading to improved recognition accuracy. Conversely, for longer segments containing more complex information, this fixed, short preceding context might have hindered overall contextual understanding, thus degrading performance.

4.6 Impact of Noise Levels on ASR Performance

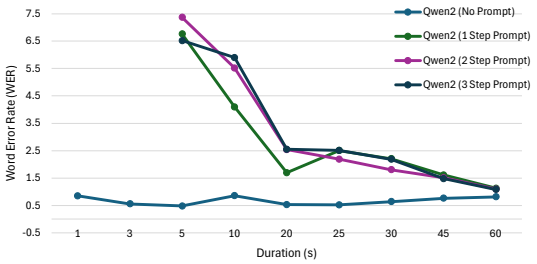
Figure 8 illustrates how the Word Error Rate (WER) of a Conformer-based Automatic Speech Recognition system varies with audio duration across different Signal-to-Noise Ratio (SNR) conditions. As noise increases (i.e., as SNR decreases), WER rises, indicating a decline in recognition accuracy. However, increasing the audio duration, thus providing more context, consistently helps reduce WER across all noise levels. This improvement is most pronounced between 5 and 60 seconds of audio. Beyond this point, the benefits of additional context begin to level off, and WER stabi-



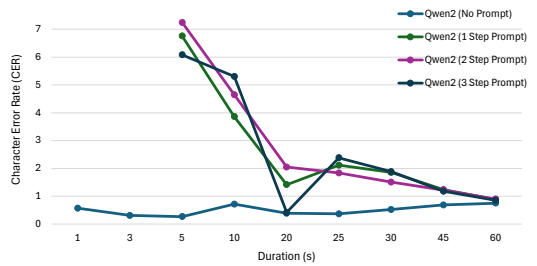
(a) Word Error Rate (WER)



(b) Character Error Rate (CER)

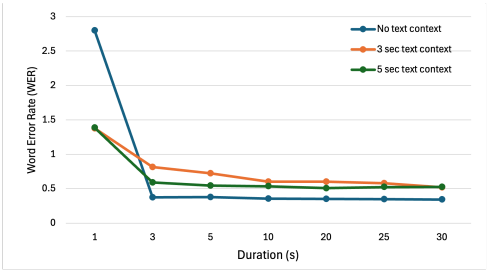


(c) Word Error Rate (WER)

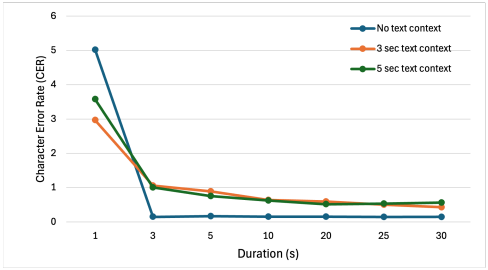


(d) Character Error Rate (CER)

Figure 6: Performance vs context length for Earnings22 dataset using Qwen2-Audio and Whisper-small. 'No prompts' indicates feeding only speech data. 'n-step prompts' indicate that the transcribed text from n steps earlier is provided as input for the current prediction time frame.



(a) Korean dataset Word Error Rate (WER)



(b) Korean dataset Character Error Rate (CER)

Figure 7: Performance vs context length for AI-Hub Korean dataset using Whipser-small. 'n-sec text context' indicate that the transcribed text from n second earlier is provided as input for the current prediction time frame.

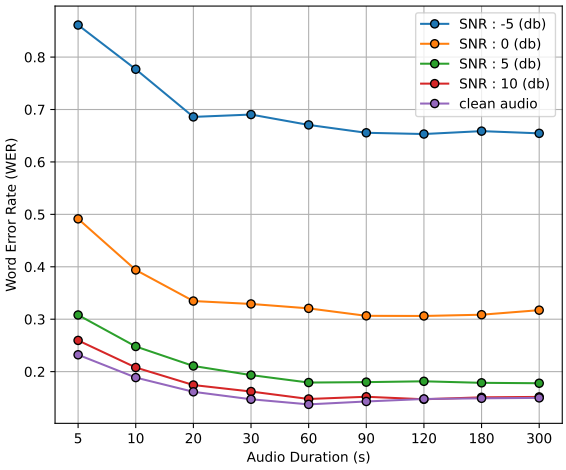


Figure 8: Performance of Conformer under various noise conditions.

lizes. Clean audio yields the lowest WER throughout, representing the system's optimal performance. Notably, even under moderate noise levels, such as 10 dB SNR, the ASR model can approach clean audio performance when given enough context.

5 Conclusion

In this experiment, we systematically evaluated the impact of audio and textual context length on ASR performance across diverse models and datasets. While longer context is often assumed to improve performance, our findings reveal that the actual benefits in terms of WER and CER reduction are frequently limited and highly conditional on the specific model, dataset, and duration. Performance gains often saturated relatively early in our tests, in-

dicating rapidly diminishing returns beyond moderate context lengths. Moreover, attempts to leverage longer context introduced significant drawbacks, notably increased computational demands, particularly memory usage which became prohibitive for models like Fast-Conformer at extended durations. Crucially, under certain experimental conditions, such as our cross-lingual tests or potentially due to error propagation, utilizing longer context even resulted in performance degradation.

Overall, our results suggest that while context does influence ASR, the practical advantages of aggressively increasing context length appear constrained and often overshadowed by computational costs and potential accuracy trade-offs. The pursuit of ever-longer context windows may not be the most efficient path to enhancing ASR systems. Future work could perhaps focus more on optimizing the effective use of moderate context lengths or developing models inherently more robust to local ambiguities, lessening the reliance on extensive historical information. Investigating adaptive or selective context mechanisms, rather than simple window extension, might also yield more practical benefits. Ultimately, developing efficient and reliable ASR systems that perform well without excessive dependence on long-range context seems crucial for widespread, real-world deployment.

Contribution

The contributions of the authors are as follows:

- **Hyunho Ahn:** Focused on the analysis and evaluation of the Earnings 22 dataset using Whisper, Conformer, RNN-T, and Qwen2-Audio. Finalized and organized the results, including tables and figures.
- **Hansol Lee:** Identified and prepared the Korean AI Hub dataset for the cross-lingual analysis; Designed and implemented the Korean data loading pipeline to handle variable segment lengths for context experiments; Conducted the inference experiments across Korean specified ASR models, Korean datasets, and context conditions.
- **Shakhrul Iman Siam:** Conducted analysis on the TED-LIUM dataset using Whisper, Conformer, and RNN-T models. Evaluated and compared peak and average GPU memory consumption during inference across various ASR models with varying input audio lengths. Designed and executed experiments to investigate the impact of different noise levels on ASR performance.

Acknowledgments

Portions of this report, including minor edits and phrasing improvements, were assisted by ChatGPT. All core content, analysis were written independently by the author.

References

Shuo-Yiin Chang, Chao Zhang, Tara N Sainath, Bo Li, and Trevor Strohman. 2023. Context-aware end-to-end asr using self-attentive embedding and tensor fusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525. IEEE.

Jian Cheng. 2024. Context-aware speech recognition using prompts for language learners. In *Proc. Interspeech 2024*, pages 4009–4013.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. 2022. Earnings-22: A practical benchmark for accents in the wild. *arXiv preprint arXiv:2203.15591*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Robert Flynn and Anton Ragni. 2023. How much context does my attention-based asr system need? *arXiv preprint arXiv:2310.15672*.

Jennifer Drexler Fox, Desh Raj, Natalie Delworth, Quinn McNamara, Corey Miller, and Migüel Jetté. 2024. Updated corpora and benchmarks for long-form speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13246–13250. IEEE.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data

with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1764–1772, Bejing, China. PMLR.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer.

W Ronny Huang, Shuo-Yiin Chang, David Rybach, Tara Sainath, Rohit Prabhavalkar, Cal Peyser, Zhiyun Lu, and Cyril Allauzen. 2022. E2e segmenter: Joint segmenting and decoding for long-form asr. In *Proc. Interspeech 2022*, pages 4995–4999.

Mahaveer Jain, Kjell Schubert, Jay Mahadeokar, Ching-Feng Yeh, Kaustubh Kalgaonkar, Anuroop Sriram, Christian Fuegen, and Michael L. Seltzer. 2020. Rnn-t for latency controlled asr with improved beam search. *Preprint*, arXiv:1911.01629.

Junteng Jia, Gil Keren, Wei Zhou, Egor Lakomkin, Xiaohui Zhang, Chunyang Wu, Frank Seide, Jay Mahadeokar, and Ozlem Kalinli. 2025. Efficient streaming llm for speech recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yoonsung Kim, Yoonsu Park, TmaxSoft, IcecreamEdu, Korea Edutech Industry Association, and Namu Technologies. 2021. Ai-hub lecture transcription dataset: Korean long-form speech corpus for asr. In *Open AI Dataset Project (AI-Hub), Republic of Korea*, page Available online. Ministry of Science and ICT. Available at `https://www.aihub.or.kr/aidata/105`.

Egor Lakomkin, Chunyang Wu, Yassir Fathullah, Ozlem Kalinli, Michael L Seltzer, and Christian Fuegen. 2024. End-to-end speech recognition contextualization with large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12406–12410. IEEE.

Zehan Li, Haoran Miao, Keqi Deng, Gaofeng Cheng, Sanli Tian, Ta Li, and Yonghong Yan. 2022. Improving streaming end-to-end asr on transformer-based causal models with encoder states revision strategies. *arXiv preprint arXiv:2207.02495*.

Yi Liu, Pascale Fung, Yongsheng Yang, Christopher Cieri, Shudong Huang, and David Graff. 2006. Hkust/mts: A very large scale mandarin telephone speech corpus. In *International Symposium on Chinese Spoken Language Processing*, pages 724–735. Springer.

Kikuo Maekawa et al. 2003. Corpus of spontaneous japanese: Its design and evaluation. In *Proc. ISCA & IEEE workshop on spontaneous speech processing and recognition*, volume 2003, pages 7–12.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *Preprint*, arXiv:2305.05084.

Yixuan Tang and Anthony KH Tung. 2024. Contextualized speech recognition: rethinking second-pass rescoring with generative large language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6478–6485.

Emiru Tsunoo, Yosuke Kashiwagi, Toshiyuki Kumakura, and Shinji Watanabe. 2019. Transformer asr with contextual block processing. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 427–433. IEEE.

Chih-Kai Yang, Kuan-Po Huang, and Hung-yi Lee. 2024. Do prompts really prompt? exploring the prompt understanding capability of whisper. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1–8. IEEE.