How Much Context Is Enough? Evaluating the Role of Audio and Textual Context in ASR Systems

Anonymous ACL submission

Abstract

Automatic Speech Recognition (ASR) systems often process audio in short segments, limiting their ability to leverage broader context. This work systematically explores how increasing both audio and textual context length affects ASR performance. We evaluate multiple architectures—including Fast Conformer with CTC and RNN-T 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 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.

Our study demonstrates that expanding audio context beyond short windows initially enhances ASR performance, reducing WER. However, our experiments reveal a distinct threshold beyond which extending audio context offers minimal further gains and begins to compromise efficiency due to increased GPU memory demands and inference latency. This highlights the importance of identifying an optimal context length. Notably, we observe that including transcribed-textual context degrades the performance of multi-modal models. This suggests that current architectures, not having been predominantly trained with such contextual inputs, struggle to integrate them effectively, leading to a cascade of transcription errors.

2 Related Work

Recent ASR research, demonstrates that broader context improves model performance (Li et al., 2022; Fox et al., 2024; Flynn and Ragni, 2023). For instance, Flynn and Ragni (2023) showed the advantages of long-form audio training for transcription. Complementary work has integrated textual context with audio. Chang et al. (2023) explored fusing acoustic and text information in RNN-T architectures. Lakomkin et al. (2024); Chen et al.

(2024); Yang et al. (2024); Cheng (2024) investigated in-context learning for the multi-modal LLM decoder, such as adding keywords and video descriptions. Radford et al. (2023) proposed Whisper, showing the benefit of feeding previously transcribed text for long-form transcription, though specifics 30-second speech input limit were noted.

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.

3 Methodology

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

3.1 Research Objectives

This study investigates the impact of varying audio and textual context length on ASR model performance during inference. While prior work indicates benefits from additional context, the optimal amount and potential for diminishing returns remain 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 (previous transcriptions) 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, experiments utilize context window sizes from 1 second to 15 minutes of continuous audio, and for multimodal models, varying lengths of prior transcribed text is provided during inference.

3.2 Experimental Setup

Datsets. For data preparation, in each context length setting, we build data pairs (audio and ground-truth transcription of the audio). We use three datasets in the experiments, TED-LIUM (Hernandez et al., 2018), Earnings 22 (Del Rio et al., 2022), and AIHub Korean Lectures (Kim et al., 2021). The details about the dataset are described in appendix A

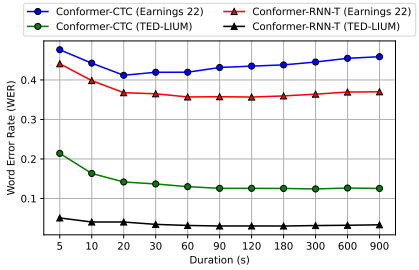
Models. The ASR models employed in this study are categorized based on their input types:
Audio-Only Model: For audio-only input models, traditional ASR architectures of NVIDIA Fast Conformer model (Rekesh et al., 2023) are used. We evaluate both both CTC based and RNN-T based Fast Conformer. We abbreviate them as Conformer-CTC and Conformer-RNN-T in figures, respectively.
Audio + Text Model: For this type of model, Whisper (Radford et al., 2022) and Qwen2-Audio (Chu et al., 2024) are used in the experiment, which is the LLM-based ASR model. We abbreviate them as Whisper and Qwen2 in figures, respectively.
Evaluation Metrics. We use Word Error Rate (WER), Inference Speed and Memory Usage to measure its performance. 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.
Implementation Details. We obtained the pretrained model from Nvidia Nemo[1] and Hugging Face Hub[2]. All the experiments were conducted in NVIDIA A100 80GB GPU.

4 Key Findings

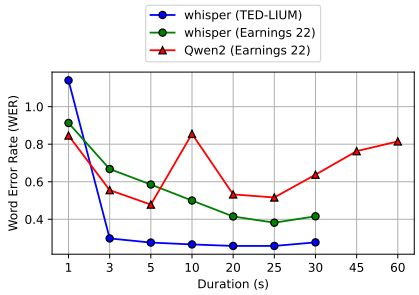
4.1 Effect of Audio Context Length

ASR Performance: We evaluate how varying the duration of the input audio affects ASR performance across different models on Earnings-22 and TED-LIUM datasets. Figure 1a shows the impact of varying input audio duration on ASR models' WER using the NVIDIA Fast Conformer CTC and RNN-T models. The figure indicates that performance begins to saturate after an audio duration of 20 and 90 seconds, accordingly for the Fast Conformer CTC and RNN-T. Similarly, Figure 1b illustrates the effect of varying input audio duration using the Whisper-small and Qwen2-Audio models, where performance saturation is observed after 20 seconds. Unlike Qwen2-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

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

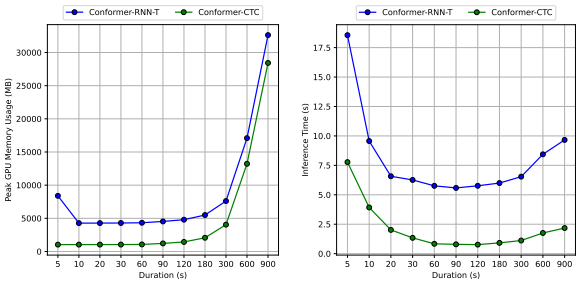


(a) NVIDIA Fast Conformer CTC and RNN-T model



(b) Whisper-small and Qwen2-Audio model

Figure 1: Performance vs context length for TED-LIUM and Earnings 22 dataset.



(a) Peak GPU Memory usage (b) Inference Time.

Figure 2: Peak GPU Memory usage and Inference Time for NVIDIA Fast Conformer CTC and RNN-T with with varying audio durations.

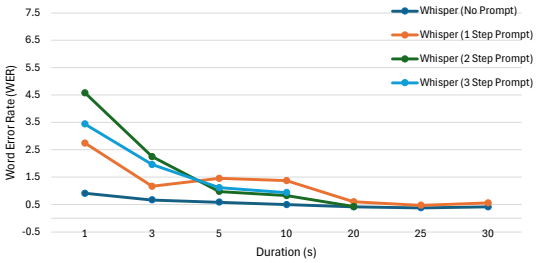
attributed to differences in model architecture or the way it processes the audio segments and bias in the training data.

GPU peak memory usage: Figure 2a indicates GPU peak memory usage with varying audio duration for NVIDIA Fast Conformer CTC and RNN-T models. For both models, the memory usage remains almost constant for shorter intervals but rises sharply beyond 300 seconds, becoming impractical for longer audio durations due to excessive memory demands.

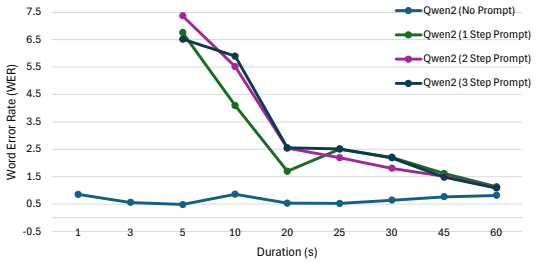
Inference time: Figure 2b demonstrates the impact of different input time-frame lengths on the overall processing time for transcribing the complete audio input. When using shorter time-frame (e.g., 1s), the increased number of segmentation leads to frequent ASR model invocations, causing significant computational overhead. Additionally, we observe saturation effects in model performance (WER) depending on the input time-frame length: for Whisper, performance tends to saturate around 20-second windows, while for Fast Conformer CTC 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.

4.2 Effect of Textual Context

We investigate whether feeding previous transcription results improves ASR in the audio. The figure 3 presents WER under different prompt settings. Across all configurations, 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. Notably, at early steps (e.g., 1 or 3 seconds), transcriptions often include short, possibly error-prone segments,



(a) Whisper



(b) Qwen2

Figure 3: 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.

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 input length constraint, we were unable to evaluate prompts with a large number of steps (e.g. 3 steps 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 non-meaningful outputs. Therefore, we excluded these results from our evaluation.

4.3 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). The experimental results (Figure 4) revealed an interesting pattern. Injecting preceding textual context yielded improved WER performance only for extremely short, 1-second target segments and shown performance degradation after that. We assume that 1-second segments showed improved performance due to frequent number pronunciations, which are often ambiguous in Korean because of its dual numeral systems (Native Korean and Sino-Korean). 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. Consequently, we presume 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.

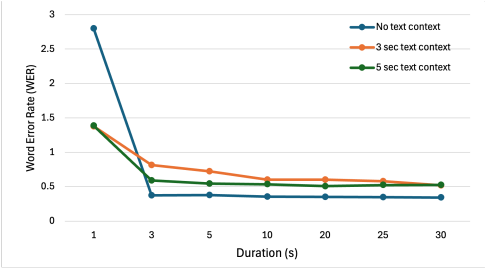


Figure 4: Performance vs context length for AI-Hub Korean dataset using Whipser-small. 'n-sec text context' indicate that the transcribed text from n second earlier is provided as input for the current prediction time frame.

4.4 Impact of Noise Levels on ASR Performance

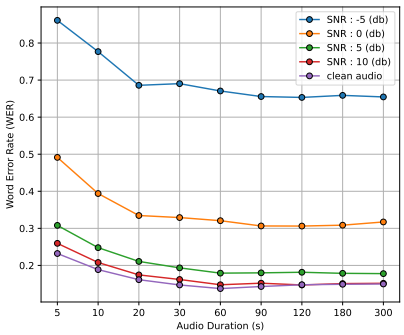


Figure 5: Performance of Conformer under various noise conditions.

Figure 5 illustrates how the WER of a Conformer-based ASR 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 stabilizes. 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 work, we evaluated how varying audio and textual context lengths affect ASR performance across different models and datasets. Our findings reveal that the actual benefits in terms of ASR performance are frequently limited and highly conditional on the specific model, dataset, and duration. Performance gains often saturated relatively early in our tests, indicating 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. Future work could focus on optimizing the use of moderate context lengths or developing models that are inherently more robust to local ambiguities, reducing the need for extensive historical information.

6 Limitations

Despite offering insights into the effect of audio and textual context lengths in ASR systems, our study is constrained by several limitations. The Whisper model imposes a hard limit of 30 seconds on audio input length. As a result, we were unable to evaluate Whisper's performance on longer context windows, which restricted our ability to fully compare its behavior with other models under extended audio conditions. Moreover, while our experiments included both English and Korean datasets, the majority of the analysis focused on English speech from well-structured sources such as TED talks and earnings calls. This limited linguistic and domain diversity may reduce the generalizability of our findings to more conversational, noisy, or code-switched speech data.

7 Ethics Statement

We foresee no ethical concerns with our work. The datasets employed in our research are publicly available, and it does not contain any personal information.

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*.

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.

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.

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.

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.

A Additional Details on Datasets

The primary focus of this work 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. Despite this limitation, to facilitate multilingual performance evaluation and quantify our model with diverse data, we incorporated a Korean dataset (Kim et al., 2021) into our experiments. Table 1 shows overview of the dataset we used in the experiments. To be specific we used following dataset:

TED-LIUM: The TED-LIUM dataset (Hernandez et al., 2018) employed for one of base long-form 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.

Earnings-22: 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. Our experimental subset of the Earnings-22 dataset comprises randomly selected segments, which collectively span 11 hours of audio.

AI Hub Korean lecture: 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.

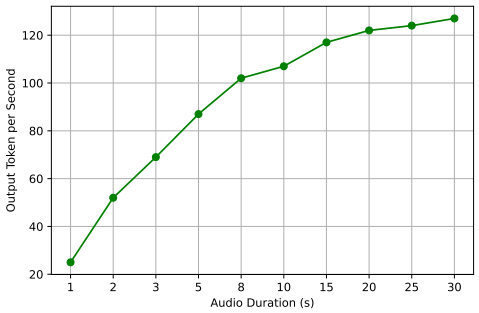
B Additional Details on ASR Models

Table 2 shows the overall details of the model used in the evaluation. To be specific, pre-trained Fast Conformer-CTC is obtained from the NVIDIA Nemo platform. Other models are from the Hugging Face model hub.

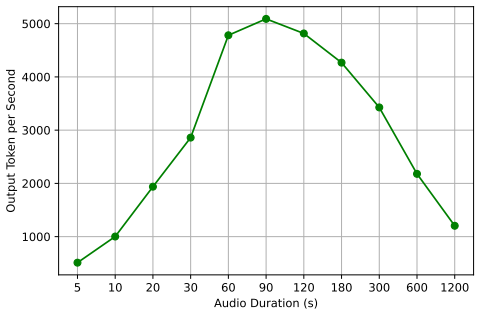
C More Results

C.1 Output Token Per Second vs audio length

Figure 6 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.



(a) Whisper-small



(b) NVIDIA FastConformer

Figure 6: Output Token generated per second. All experiments were done in NVIDIA A100 80GB GPU.

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

Model Name	Parameters	Architectures	Implementation/Source
Fast Conformer-CTC	115M	Fast Conformer with CTC decoder	nvidia/stt_en_fastconformer_ctc_large
Fast Conformer-RNN-T	1.1B	Fast Conformer with RNN-T decoder	nvidia/parakeet-rnnt-1.1b
Whisper-small	244M	Transformer based encoder-decoder model	openai/whisper-small
Qwen2-Audio	7B	Audio encoder with Qwen LM	Qwen/Qwen2-Audio-7B

Table 2: Overview of models used in the experiments