Evaluating the Impact of Audio Quality on STT in Unstable Network Conditions

Jaeik Kim
Department of Computer Engineering
Changwon National University
Changwon, South Korea
kimji0129s@gmail.com

Seungwoo Hong
Network Research Department
Electronics and Telecommunications
Research Institute (ETRI)
Daejeon, South Korea
swhong@etri.re.kr

Donghyeok An
Department of Computer Engineering
Changwon National University
Changwon, South Korea
donghyeokan@changwon.ac.kr

I. Introduction

Speech recognition technology has become an essential user interface in a wide range of applications including intelligent speakers, smart devices, and call centers. These speech-to-text (STT) systems generally rely on high-quality audio inputs to achieve accurate transcription. However, when operating under limited or unstable network conditions, seamless transmission of audio data is challenging, which can significantly degrade recognition accuracy.

Semantic communication, a recently highlighted paradigm, seeks to optimize communication efficiency by focusing on the meaning of the transmitted data rather than the raw data itself [1]. By understanding the semantic content and selectively transmitting only necessary information, bandwidth consumption can be greatly reduced. This is particularly advantageous for real-time applications in environments where network resources are constrained.

The goal of this research is to maintain real-time transmission performance under network constraints by intentionally lowering audio data quality to reduce transmission size. Using OpenAI's Whisper model [2], a state-of-the-art STT system, we analyze the relationship between audio quality parameters and recognition accuracy. This study experimentally verifies that by minimizing unnecessary data transmission and focusing on essential information, Whisper-based STT systems can be effectively operated even in restricted network environments.

II. Related Work

Efforts to overcome network bandwidth limitations in audio and speech processing have been continuously pursued. Previous studies can be broadly categorized into three approaches.

First, audio compression techniques reduce transmission data volume [3]. Codecs such as MP3, AAC, and Opus utilize psychoacoustic models to remove frequency components that are less perceptible to humans, thereby efficiently reducing file size.

Second, preprocessing filtering methods improve recognition accuracy by removing noise and enhancing signal quality [4]. These methods primarily focus on cleaning audio signals captured in noisy environments.

Third, edge-cloud distributed processing structures allocate lightweight audio preprocessing and simple speech recognition models on edge devices to reduce network load and minimize latency [5].

However, these studies have largely emphasized system or technical improvements without extensively evaluating the robustness of specific models such as Whisper under varying audio quality conditions. This work aims to fill that gap by quantitatively analyzing how audio quality parameters—including sample rate, bit depth, and bitrate—impact recognition accuracy of contemporary STT models, thereby proposing optimal quality parameters for efficient real-time transmission.

III. Methodology

We employed the "Extreme Noise Speech Recognition Dataset" provided by AI Hub [6], which contains over 12,000 hours of Korean speech collected across six environments: transportation, construction sites, factories, indoor/outdoor facilities, natural, and mechanical noise. These diverse noise conditions realistically reflect harsh real-world scenarios, making the dataset suitable for evaluating STT model performance under noisy conditions.

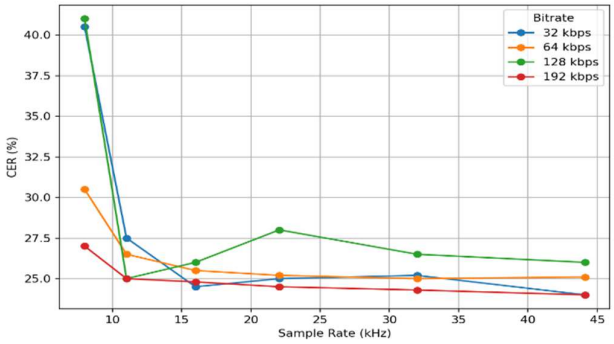
Our experiments utilized the OpenAI Whisper base model, which features a Transformer-based encoder-decoder architecture and demonstrates high performance in multilingual speech recognition and translation. The Whisper base model offers a lightweight parameter size while maintaining robustness against various noise environments.

We compared recognition accuracy by adjusting audio quality parameters—sample rate, bitrate, and format—and measured performance using Character Error Rate (CER).

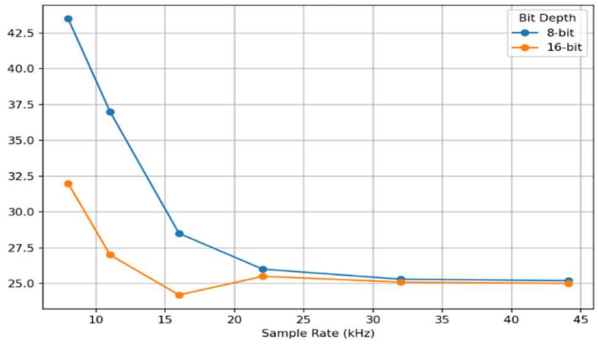
$$CER = \frac{S+D+I}{N} \quad (1)$$

CER is defined as where S denotes substitutions, D deletions, I insertions, and N the total number of characters in the reference transcript. CER is chosen over Word Error Rate (WER) as it captures fine-grained errors more accurately for character-based languages like Korean.

The experimental procedure involved converting the original audio into two formats: WAV and MP3. For WAV,



(A) MP3



(B) WAV

Figure 1 Speech recognition accuracy across various sample rates and bit rates

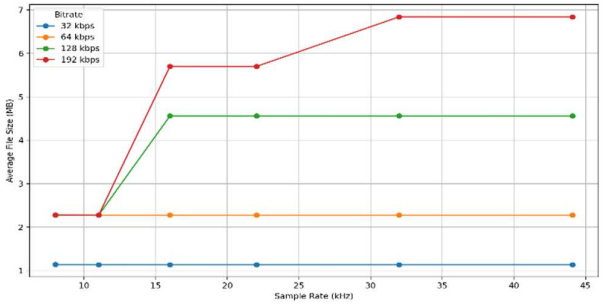


Figure 2 MP3 file size across various sample rates and bit rates

sample rates of 8, 11.025, 16, 22.05, 32, and 44.1 kHz were tested, combined with 8-bit and 16-bit depths, yielding 12 quality configurations. For MP3, the same sample rates were used with bitrates of 32, 64, 128, and 192 kbps, totaling 24 configurations. Each audio version was fed into the Whisper model to obtain CER, and file sizes were compared.

IV. Evaluation

The experimental results indicate that when the sample rate and bit rate of the audio data exceed certain thresholds, the speech recognition accuracy remains largely unaffected. Fig. 1 and Fig. 2 present the speech recognition performance and file sizes of MP3 and WAV audio data under various sample rate and bit depth/bit rate conditions.

Fig. 1(A) shows the results for MP3 audio. When the sample rate is at least 16 kHz, the Character Error Rate (CER) remains stable at 24–27%, regardless of the bit rate. This suggests that a sufficiently high sample rate preserves the phonetic cues necessary for accurate recognition, while lower bitrates mainly remove perceptual redundancies. For instance, a 16 kHz/64 kbps configuration reduces file size by approximately 75% compared to 44.1 kHz/192 kbps, with only a 1.2 percentage-point increase in CER. Therefore, by appropriately reducing the sample rate and bit rate, it is possible to maintain recognition accuracy while minimizing transmission volume, enabling real-time speech transmission even under unstable network conditions.

Fig. 1(B) shows the results for WAV audio. At lower sample rates (8 kHz and 16 kHz), bit depth significantly affects CER, with a reduction observed when increasing from 8-bit to 16-bit. However, when the sample rate exceeds 22.05 kHz, the CER difference between 8-bit and 16-bit audio becomes negligible (within 1 percentage point). Similar to MP3, this demonstrates that selecting appropriate sample rates and bit depths for WAV audio enables efficient data transmission without significant degradation in recognition performance. These results suggest that adjusting audio quality can minimize latency in speech recognition across various network conditions.

V. Conclusion

This work experimentally demonstrates that audio quality adjustment improves transmission efficiency in bandwidth-limited environments. Both MP3 and WAV formats achieve stable recognition accuracy when the sample rate exceeds 16 kHz and 22.05 kHz, respectively. This suggests the feasibility of real-time speech recognition under unstable network conditions. Future work will extend the evaluation to diverse datasets and scenarios, and will develop algorithms for dynamic audio quality selection to maintain accuracy while adapting to network constraints.

Acknowledgment

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by ICT R&D program of MSIT/IITP [2021-0-00715, Development of End-to-End Ultra-high Precision Network Technologies]. This work was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (2410010873, The Competency Development Program for Industry Specialist)

References

[1] D.-S. Kwon and J. Doe, "Trends in semantic communications," *Electronics and Telecommunications Trends*, 2022, [in Korean].

[2] Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv:2212.04356, 2022.

[3] L. Drude, J. Heymann, A. Schwarz, and J.-M. Valin, "Multi-channel Opus compression for far-field automatic speech recognition with a fixed bitrate budget," arXiv, 2021.

[4] S.-J. Lee and H.-Y. Kwon, "A Preprocessing Strategy for Denoising of Speech Data Based on Speech Segment Detection," *Applied Sciences*, vol. 10, no. 20, p. 7385, 2020.

[5] Y. Kang, J. Hauswald, C. Gao, and L. Tang, "Task Offloading for Automatic Speech Recognition in Edge-Cloud Computing Based Mobile Networks," *Computer Architecture News*, Apr. 2017.

[6] AI Hub, "Extreme Noise Speech Recognition Dataset," AI Hub, Accessed:Aug.2025.[Online].Available:https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=7141