Comparative Evaluation of Speech-to-Text Models for Lithuanian Transcription: Effects of Audio Formats and Recording Environments

Dovydas Šablevičius, Asta Slotkienė

Vilniaus Gedimino technikos universitetas,
Saulėtekio al. 11, LT-10223 Vilnius
dovydas.sablevicius@stud.vilniustech.lt

Summary. This study evaluates the performance of various speech-to-text models for Lithuanian transcription, focusing on how audio formats and recording environments affect their accuracy. Among the models tested, Google's Chirp-2 demonstrated the highest accuracy under optimal conditions. However, its performance declined with increased playback speeds and in environments with significant background noise, highlighting the importance of controlled recording conditions for effective deployment of STT systems in real-world applications.

Keywords: Speech-to-text model, audio format, transcription accuracy, recording environment, Lithuanian language.

1 Introduction

In the past decade, artificial intelligence accelerated the application of speech-to-text (STT) and inspired the development of STT services that convert natural spoken language into text. STT is applicable to various areas where human interaction with computers and other digital devices is necessary, such as healthcare [1, 2, 3], education [4,5], etc. This technology enhances system services, improves content accessibility, and enables automated documentation without the need for human interpretation. However, a major challenge remains: most STT models have been predominantly developed and evaluated for English, which limits their effectiveness when applied to other languages, including Lithuanian [6]. In recent years, efforts have been made to develop increasingly accurate Lithuanian STT models despite limited training data [7]. Additionally, large-scale projects such as LIEPA-3 are creating extensive Lithuanian speech corpora to further improve

DOI: https://doi.org/10.15388/LMITT.2025.23

training resources for STT systems [8]. The objective of this research is twofold. First, it seeks to determine which existing STT model yields the highest transcription accuracy for the Lithuanian language. Second, it aims to analyze how various audio parameters—including technical specifications like sampling rate and bit depth, as well as environmental conditions such as room acoustics and ambient noise—influence transcription accuracy.

2 Research methodology

To describe the standardized methodology applied throughout this research is essential because this approach ensured consistent recording conditions and established a robust pipeline for transcription evaluation.

All recordings utilized an identical 226-word Lithuanian script focused on topics related to local nature and culture, including specific references to regional objects. A native Lithuanian speaker conducted every recording from a fixed distance of 60 centimeters from the microphone. Room noise levels were measured using the NIOSH Sound Level Meter App, which resulted in three distinct sets of recordings: one set consisted of three baseline recordings in a quiet room at approximately 30 dB, with durations of 2.02 min, 2.06 min, and 2.03 min. Another set comprised three recordings with "Coffee Shop Background Noise," maintaining a constant ambient noise level of around 50 dB and their durations were 1.50 min, 1.55 min, and 2.02 min. The final set involved three recordings in a naturally echoey room, also at approximately 30 dB but with noticeable acoustic reverberations and their durations were 2.04 min, 1.04 min, and 2.01 min.

The baseline audio files then underwent additional processing based on experimental conditions, as illustrated below in Figure 1. In Test 1, audio files were exported in FLAC format at 48 kHz, with a 24-bit depth and a mono channel. Test 2 maintained these FLAC settings but varied the channel counts between mono and stereo. Tests 3 and 4 followed the configurations detailed in Figure 1, involving variations in sampling rates, bit depths, and audio compression formats. Meanwhile, Tests 5 and 6 utilized FLAC files at 16 kHz and 16-bit depth (mono), which were modified to assess playback speed variations along with additional environmental factors such as echo and background noise.

Controlled parameters—such as the speaker-to-microphone distance and ambient noise measurements—were essential to ensure that any observed differences in transcription accuracy could be attributed solely to the tested experimental variables.

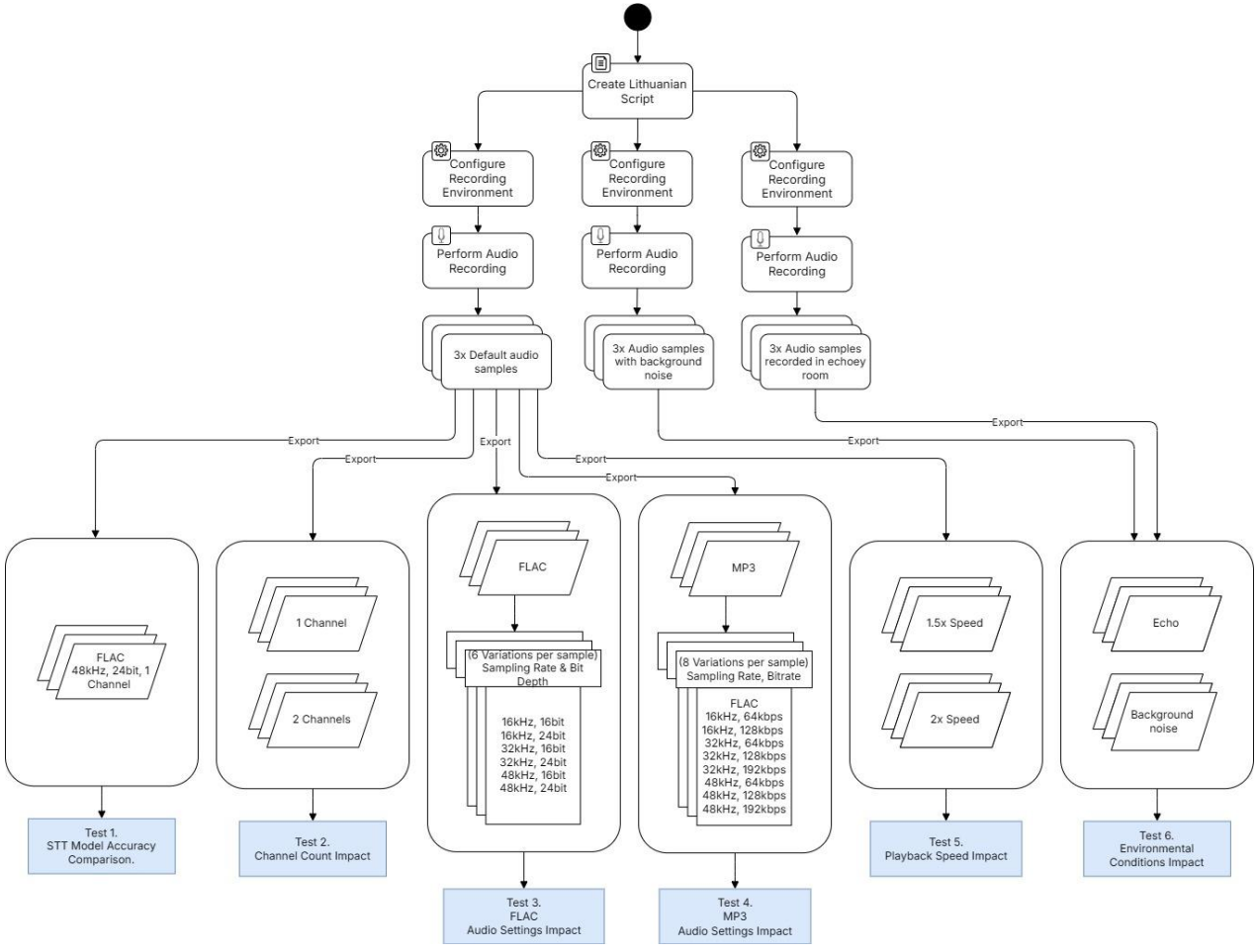

Figure 1. Audio sample preparation process.

Following the generation of the audio samples, all recordings were processed through a uniform transcription pipeline, as outlined below in Figure 2. In this phase, each audio file was asynchronously submitted to the selected STT models operating in their default configurations. The resulting transcripts, along with a reference transcript, underwent a standardized normalization process that included:

1. Converting all text to lowercase
2. Removing punctuation
3. Eliminating extra spaces (leading, trailing, or redundant)

Accuracy metrics were calculated using the JiWER Python library, comparing the normalized model outputs against the normalized reference transcripts. The following metrics were recorded: Word Error Rate (WER), Character Error Rate (CER), and Real-Time Factor (RTF). These metrics calculation and interpretation are detailed in section 3.

Figure 2 encapsulates this transcription and evaluation workflow, underscoring the methodical approach taken to process each audio input and compute the metrics.

Figure 2 outlines this transcription and evaluation workflow, emphasizing the systematic approach taken to accurately assess the impact of different audio characteristics on transcription performance. Overall, the described methodology ensures reliable and comparable results across varied testing conditions, thereby enabling clear insights into how specific audio attributes influence the accuracy and efficiency of STT models.

3 Metrics for Speech-based Models

In this investigation, the results were validated with a wide range of evaluation metrics, such as acoustic and text-based metrics, which allows to compare different STT models. One of them is word error rate (WER) is calculated using Levenshtein distance between word [9]. In Levenshtein distance, we count the number of insertions (I), substitutions (S) and deletions (D) performed to equal two-word sequences. WER is calculated by this formula present in (1). Lower values of WER are preferred since they indicate an STT model that makes fewer errors.

$$WER = \frac{S+D+I}{N} \tag{1}$$

Where S = the number of substitutions required to change the hypothesis string to the reference string, D = the number of necessary deletions, I = the

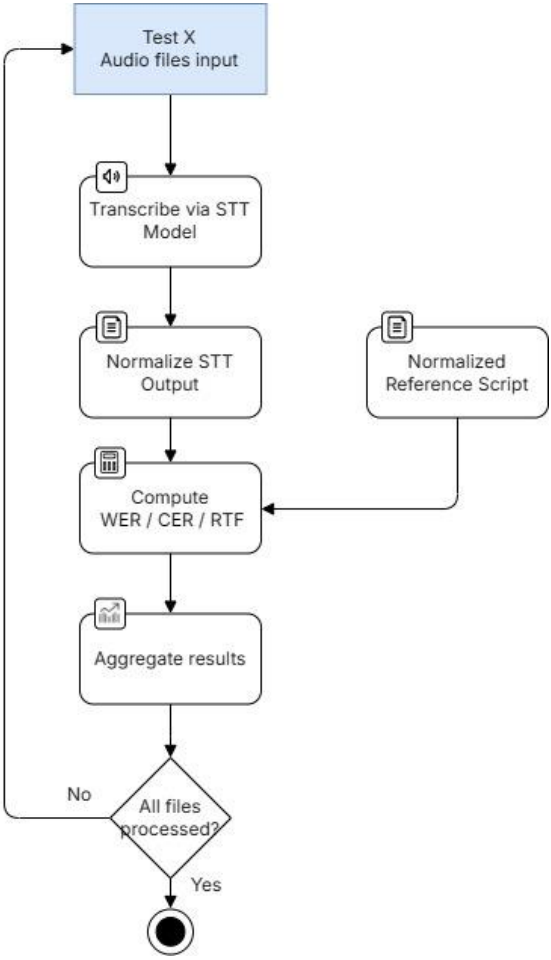

Figure 2. Audio transcription and evaluation pipeline.

number of insertions, and N = the total number of words in the reference string [10]. Other metrics that complement the widely used WER to evaluate the performance of STT model are character error rate (CER). CER measures the edit distance between the recognized text and the reference text at the character level, where substitutions (S), deletions (D), and insertions (I),

respectively, at the character level [11]. The CER is calculated as (2) formula, and with lower values indicating better performance of STT model.

$$CER = \frac{S+D+I}{N} \tag{2}$$

In addition to the accuracy indicator, the real time factor (RTF) is another important performance indicator and it applied to measure the speed of a system that processes an input audio signal. The RTF is the ratio of the time taken to process a speech to the duration of that speech, and the lower the value is, the better the real-time performance of text extraction. The RTF can be defined as:

$$RTF = \frac{T}{D} \tag{3}$$

Where T is time to transcribe the audio file and D is duration of the audio file [8]. Values of RTF<1.0 are preferred since values≥1.0 indicate that the decoding (transcribing) an audio file takes a larger amount of time than the duration of the audio itself.

4 Analysis of Experimental Results

This section provides an in-depth analysis of the experimental results, beginning with the identification of the optimal STT model for Lithuanian transcription and continuing with a detailed examination of how various audio configurations and environmental conditions affect transcription accuracy.

Test 1 identified the best STT Model for the Lithuanian language.

The goal of this test was to determine which of the five selected STT models provides the highest transcription accuracy for the Lithuanian language. By comparing performance across these models, the optimal candidate was chosen to serve as the baseline for subsequent experiments involving modifications in audio quality. The selected models were based on their explicit support for Lithuanian language. The models evaluated were:

1. OpenAI: GPT-4o-transcribe
2. Google: Chirp-2
3. Amazon Web Services: unknown
4. Microsoft: Whisper Large V2
5. Tilde: unknown

The selection criteria emphasized the providers' market scale and reputation, as well as Tilde's specific focus on Baltic region languages.

The results presented in Table 1 reflect the average performance across the three audio recordings.

Table 1. STT Model Accuracy Comparison.

Provider	Model	WER	CER	RTF
OpenAI	GPT-4o-transcribe	0.196	0.053	0.083
Google	Chirp-2	0.093	0.033	0.086
Amazon Web Services	Unknown	0.133	0.044	0.112
Microsoft	Whisper Large V2	0.342	0.217	0.519
Tilde	Unknown	0.323	0.211	0.591

The results indicate that Google's Chirp-2 outperforms the other models, achieving the lowest WER (0.093) and CER (0.033). Although OpenAI's GPT-4o-transcribe demonstrated a slightly lower RTF (0.083) compared to Chirp-2 (0.086), its error rates were considerably higher, making it less favorable for accurate transcription. The unnamed models from Amazon Web Services, Microsoft, and Tilde showed progressively poorer performance with higher error rates and, in the case of Microsoft and Tilde, substantially higher RTFs, suggesting slower processing times unsuitable for efficient transcription tasks.

In conclusion, the evaluation establishes Google's Chirp-2 as the optimal STT model for Lithuanian transcription under controlled conditions. Its superior accuracy, combined with competitive processing speed, renders it the preferred choice for further experiments involving the impact of audio quality modifications and environmental factors.

For Tests 2 through 5, the same set of audio recordings and normalization procedures as described in Test 1 were employed. In these tests, only the performance of Google's Chirp-2 was evaluated, and transcription accuracy was quantified solely using Word Error Rate (WER) and Character Error Rate (CER). Each result represents the average performance across the three audio recordings.

Test 2 investigated whether the number of audio channels influences transcription accuracy. The recordings were processed in both mono (one channel) and stereo (two channels) configurations using Chirp-2 model. The results, summarized in Table 2 below, indicate that there is no discernible

difference in performance between the two configurations. Both mono and stereo recordings yielded identical error rates (WER of 0.093 and CER of 0.033), suggesting that mono recordings are sufficient for accurate transcription while offering advantages in terms of file size.

Table 2. Channel Count Impact.

STT model	Channel count	WER	CER
Chirp-2	1	0.093	0.033
	2	0.093	0.033

Test 3 examined the impact of varying FLAC audio settings on transcription accuracy. In this test, the FLAC files were exported using different combinations of sampling rates (16kHz, 32kHz, and 48kHz) and bit depths (16-bit and 24-bit). Table 4 shows that the optimal performance was achieved with both 16kHz/16-bit and 16kHz/24-bit configurations, each resulting in a WER of 0.088 and a CER of 0.031. Since lower sampling rates and bit depths produce smaller file sizes without compromising accuracy, the 16kHz, 16-bit configuration was selected as the optimal setting for subsequent evaluations.

Table 3. FLAC Audio Settings Impact.

STT model	Sampling Rate & Bit Depth	WER	CER
Chirp-2	16kHz, 16bit	0.088	0.031
	16kHz, 24bit	0.088	0.031
	32kHz, 16bit	0.089	0.032
	32kHz, 24bit	0.092	0.033
	48kHz, 16bit	0.092	0.033
	48kHz, 24bit	0.092	0.033

Test 4 focused on the effect of MP3 compression settings on transcription accuracy. The original recordings were converted to MP3 format under various configurations, varying both sampling rates (16kHz, 32kHz, and 48kHz) and bitrates (64, 128, and 192 kbps), while maintaining mono audio. Table 3 presents the average WER and CER for each configuration. The findings demonstrate that transcription accuracy remains relatively stable across the different MP3 settings. With an average WER of approximately

0.091 and a CER of about 0.032, these results imply that MP3 compression does not significantly impair the performance of Chirp-2, thereby providing flexibility in the choice of audio compression without a major impact on accuracy. Note: The configuration 16kHz, 192kbps is not included in the table because Audacity did not support exporting audio with this specific combination.

Table 4. MP3 Compression Settings Impact.

STT model	Sampling Rate & Bitrate	WER	CER
Chirp-2	16kHz, 64kbps	0.088	0.032
	16kHz, 128kbps	0.089	0.032
	16kHz, 192kbps	-	-
	32kHz, 64kbps	0.093	0.034
	32kHz, 128kbps	0.093	0.034
	32kHz, 192kbps	0.093	0.033
	48kHz, 64kbps	0.091	0.032
	48kHz, 128kbps	0.092	0.032
	48kHz, 192kbps	0.092	0.032

Test 5 assessed the effect of altering the playback speed of the recordings on transcription accuracy. Under baseline conditions—using the normal playback speed with optimal FLAC settings (16kHz, 16bit)—the transcription achieved a Word Error Rate (WER) of 0.088 and a Character Error Rate (CER) of 0.032. The audio samples were then artificially accelerated to 1.5x and 2.0x their original speed while retaining these settings. As shown in Table 5, at a 1.5x speed-up, the WER increased to 0.140 and the CER to 0.058, corresponding to an approximate increase of 59% in WER and 81% in CER relative to the baseline. When the playback speed was doubled to 2.0x, transcription performance deteriorated even further, with the WER rising to 0.286—a 225% increase over baseline—and the CER climbing to 0.148, representing an increase of approximately 362.5%. These results underscore the model's sensitivity to changes in speech rate, confirming that acceleration negatively impacts transcription accuracy compared to the normal playback regime.

Table 5. Playback Speed Impact.

STT model	Speed	WER	CER
Chirp-2	1.5x	0.140	0.058
	2.0x	0.286	0.148

Test 6 evaluated the impact of environmental factors on transcription accuracy. Recordings were conducted under two distinct conditions: one in an echoey room selected for its natural acoustic reverberations, and another in a setting where "Coffee Shop Background Noise" was played to create a constant ambient noise level of approximately 50 dB. These conditions were compared against the baseline performance achieved under optimal recording conditions (WER of 0.088 and CER of 0.031).

As depicted in Table 6, under the echo condition, the transcription accuracy decreased slightly, with the WER increasing to 0.105 and the CER to 0.035—representing approximate increases of 19% and 13%, respectively, relative to baseline. In contrast, the background noise condition had a more pronounced effect, with the WER rising to 0.162 and the CER to 0.075. This corresponds to increases of about 84% in WER and 142% in CER compared to the optimal baseline. These results clearly indicate that while a modest echo exerts a minor impact on transcription accuracy, significant background noise substantially degrades performance, highlighting the importance of maintaining controlled acoustic environments for high-accuracy STT applications.

Table 6. Environmental Conditions Impact.

STT model	Room setting	WER	CER
Chirp-2	Echo	0.105	0.035
	Noise	0.162	0.075

5 Conclusions

In evaluating various STT models for Lithuanian transcription, Google's Chirp-2 emerged as the top performer, achieving the lowest WER of 0.093 and CER of 0.033. Subsequent analyses provided several key insights:

Firstly, the configuration of audio channels—whether mono or stereo—showed no impact on transcription accuracy, with both setups yielding identical WER and CER values. This suggests that mono recordings are

sufficient for accurate transcription while offering advantages in terms of smaller file sizes.

Secondly, among the FLAC audio settings tested, both 16kHz/16-bit and 16kHz/24-bit configurations achieved optimal performance, each resulting in a WER of 0.088 and a CER of 0.031. Given that the 16-bit setting produces smaller file sizes without compromising accuracy, it is recommended for efficient storage and processing.

Thirdly, varying MP3 compression settings, including different sampling rates and bitrates, demonstrated no significant influence on transcription accuracy. All configurations consistently yielded approximately a 0.091 WER and 0.032 CER, comparable to the lossless FLAC format. Notably, for both MP3 and FLAC formats, lower sampling rates resulted in slightly better accuracy.

However, increasing playback speed adversely affected transcription quality. At 1.5× speed, the WER increased by 59% and the CER by 81%. At 2.0× speed, the degradation was even more pronounced, with WER rising by 225% and CER by 362.5%.

Environmental conditions also played a significant role in transcription accuracy. Moderate echo led to a modest 19% increase in WER, while substantial background noise resulted in severe degradation, with WER increasing by 84% and CER by 142%.

Overall, while Chirp-2 demonstrates great performance under ideal conditions, it remains sensitive to variations in playback speed and environmental noise. These findings underscore the importance of controlled recording environments and appropriate playback settings for the effective real-world deployment of STT systems.

References

[1] Zhao, R., Choi, A. S., Koenecke, A., & Rameau, A. (2025). Quantification of Automatic Speech Recognition System Performance on d/Deaf and Hard of Hearing Speech. The Laryngoscope, 135(1), 191-197.

[2] Adedeji, A., Joshi, S., & Doohan, B. (2024). The sound of healthcare: Improving medical transcription asr accuracy with large language models. arXiv preprint arXiv:2402.07658.

[3] Afonja, T., Olatunji, T., Ogun, S., Etori, N. A., Owodunni, A., & Yekini, M. (2024). Performant ASR models for medical entities in accented speech. arXiv preprint arXiv:2406.12387.

[4] Boateng, G., Mensah, J. A., Yeboah, K. T., Edor, W., Mensah-Onumah, A. K., Ibrahim, N. D., & Yeboah, N. S. (2024, July). Brilla AI: Ai contestant for the national science and maths quiz. In International Conference on Artificial Intelligence in Education (pp. 214-227). Cham: Springer Nature Switzerland.

[5] Kaulage, A., Walunj, A., Bhandari, A., Dighe, A., & Sagri, A. (2024, May). Edu-lingo: A Unified NLP Video System with Comprehensive Multilingual Subtitles. In 2024 Second International Conference on Data Science and Information System (ICDSIS) (pp. 1-8). IEEE.
[6] Yang, Y., Song, Z., Zhuo, J., Cui, M., Li, J., Yang, B., ... & Chen, X. (2024). GigaSpeech 2: An Evolving, Large-Scale and Multi-domain ASR Corpus for Low-Resource Languages with Automated Crawling, Transcription and Refinement. arXiv preprint arXiv:2406.11546.
[7] Navickas, G., Raškinis, G., Mikulėnienė, D., Kardelis, V., Makauskaitė, I., Kasparaitis, P., ... & Korvel, G. (2024). Development of a Large Lithuanian Speech Corpus for speech recognition, artificial intelligence, and other innovative language technologies. In DAMSS: 15th conference on data analysis methods for software systems, Druskininkai, Lithuania, November 28-30, 2024. (pp. 74-75). Vilniaus universiteto leidykla.
[8] Pipiras, L., Maskeliūnas, R., & Damaševičius, R. (2019). Lithuanian speech recognition using purely phonetic deep learning. *Computers*, *8*(4), 76.
[9] Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, *33*(1), 31-88.
[10] Kheddar, H., Hemis, M., & Himeur, Y. (2024). Automatic speech recognition using advanced deep learning approaches: A survey. Information Fusion, 102422.
[11] Naqvi, S. M. R., Tahir, M. A., Javed, K., Khan, H. A., Raza, A., & Saeed, Z. (2024). Code-mixed street address recognition and accent adaptation for voice-activated navigation services. IEEE Access.