E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia Science & Technology Asia
Vol.30 No.3 July - September 2025 Page: [60-69]
Original research article

Development and Evaluation of a Thai Automatic Speech Recognition Model Using the Conformer Model

Siwakorn Kaewwichai[1], Kwanchiva Thangthai[2], Pattara Tipakorn[2], Wasit Limprasert [1,*]

[1]*Data Science and Innovation, College of Interdisciplinary Studies, Thammasat University, Pathum Thani 12120, Thailand*
[2]*National Electronics and Computer Technology Center, NSTDA, Pathum Thani 12120, Thailand*

Received 30 June 2025; Received in revised form 29 July 2025
Accepted 18 August 2025; Available online 30 September 2025

ABSTRACT

This research project aims to develop and evaluate the performance of an Automatic Speech Recognition (ASR) system for the Thai language by leveraging the Conformer architecture. Conformers integrate the strengths of Convolutional Neural Networks (CNNs), which effectively capture local acoustic features, and Transformers, which model long-range contextual dependencies. This combination enhances the overall capability of Thai speech transcription. The experiments were conducted using a diverse Thai speech dataset encompassing various accents, speaker demographics, and acoustic conditions. The dataset includes samples from Common Voice, regional dialects, elderly speakers, and audio with background noise from sources such as YouTube and podcasts. Performance evaluation metrics included Word Error Rate (WER), Insertion Error Rate (IER), and Deletion Error Rate (DER), along with model-related factors such as the number of parameters and processing efficiency measured by the Inverse Real-Time Factor (RTFx). In conclusion, the study demonstrates the moderate potential of the Conformer architecture for Thai ASR tasks, highlighting the need for further development. This includes expanding the quantity and diversity of training data to reflect real-world conditions and enhancing model robustness to complex acoustic environments. Moreover, the Fast Conformer model (115M parameters) contains approximately 13 times fewer parameters than comparable Whisper Large models (1.54B parameters) and achieves an Inverse Real-Time Factor (RTFx) of approximately 6400, which is about 44 times faster than a baseline Whisper Large v3 model (RTFx 146). This suggests its strong suitability for streaming and real-time ASR applications.

Keywords: Conformer; Fast conformer; Streaming ASR; Thai automatic speech recognition

***Corresponding author:** wasit@tu.ac.th **doi:** 10.14456/scitechasia.2025.47

1. Introduction

1.1 The growing importance of ASR and the unique challenges of the Thai language

Automatic Speech Recognition (ASR) has transitioned from a niche research area to a foundational technology underpinning a vast array of modern applications. From smart voice assistants and real-time transcription services to contactless communication interfaces and accessibility tools for the hearing impaired, ASR is integral to how humans interact with digital systems. While ASR systems for major world languages like English and Mandarin have achieved remarkable levels of accuracy, their development for other languages often encounters unique and significant hurdles. While ASR systems for major world languages like English and Mandarin have achieved remarkable levels of accuracy, their development for other languages often encounters unique and significant hurdles [1].

The Thai language, in particular, presents a formidable set of challenges for ASR development that are less pronounced in many Indo-European languages. Firstly, Thai is a tonal language, where the pitch contour of a syllable fundamentally alters its meaning. An ASR model must be able to distinguish between these subtle tonal variations to correctly interpret speech. Secondly, Thai orthography does not use explicit word boundaries, such as spaces between words. This necessitates a highly accurate and robust word segmentation preprocessing step, as an error in segmentation can cascade and lead to incorrect transcriptions. Finally, the meaning of many Thai words is heavily dependent on the surrounding context, requiring the ASR model to possess a sophisticated understanding of long-range semantic relationships.

A persistent bottleneck in advancing Thai ASR has been the relative scarcity of large-scale, high-quality, and acoustically diverse speech corpora. Datasets that comprehensively cover the full spectrum of regional accents, speaker demographics, and real-world acoustic environments are crucial for training robust models. The historical lack of such resources has contributed to a performance gap, with Thai ASR models traditionally lagging behind their English or Chinese counterparts in accuracy.

1.2 The state of Thai ASR: a gap in real-time performance

The current landscape of Thai ASR technology is characterized by a distinct performance dichotomy, creating a significant gap between different application needs. On one side, there are models like NECTEC's Partii, which are based on hybrid ASR architectures. These models are designed for efficiency and are capable of streaming, real-time transcription, but their accuracy can be limited, especially in challenging acoustic conditions. On the other side of the spectrum are massive, state-of-the-art models like OpenAI's Whisper, which leverage the Transformer architecture and vast amounts of training data to achieve very high transcription accuracy. However, their immense size and computational requirements make them resource-intensive, leading to high latency that renders them unsuitable for applications requiring continuous, real-time processing.

This division highlights a critical void in the Thai ASR ecosystem. There is a pressing need for a solution that bridges the gap between the low-latency, lower-accuracy streaming models and the high-accuracy, high-latency batch-processing models. The absence of a model that is both highly accurate and computa-

tionally efficient enough for real-time use is a major barrier to the widespread adoption of advanced voice-driven applications in Thailand, such as live event captioning, simultaneous translation, or sophisticated interactive voice response (IVR) systems.

1.3 The conformer architecture: a promising solution

A promising architectural innovation that addresses this challenge is the Conformer model, as introduced by Gulati, A. et al.. [2] The Conformer architecture represents a paradigm shift in ASR model design by synergistically combining the strengths of Convolutional Neural Networks (CNNs) and Transformers into a single, cohesive end-to-end model. CNNs are exceptionally adept at capturing fine-grained, local acoustic features and patterns from the input speech signal. They can effectively learn the subtle phonetic details that distinguish similar-sounding phonemes. In parallel, Transformers, with their self-attention mechanism, excel at modeling long-range, content-based global interactions and contextual dependencies across an entire audio sequence [2].

This hybrid approach is not a mere concatenation of layers but a fundamental design choice to overcome the traditional trade-off between local feature extraction and global context modeling. The Conformer processes speech by first using convolution modules to learn locally-aware representations, which are then fed into Transformer blocks to understand the broader sentence-level context. This architecture is therefore theoretically well-suited to tackle the specific complexities of the Thai language. It has the potential to model the crucial local tonal variations using its CNN components while simultaneously understanding the broader sentence context, which is vital for disambiguation, using its Transformer components.

1.4 Study objective

This study aims to develop and rigorously evaluate a Thai ASR model based on the highly efficient Fast Conformer architecture, a variant optimized for speed and scalability [3]. The primary objective is to assess its potential as a solution that balances high accuracy with the computational efficiency required for real-time applications.

To establish a robust benchmark, the performance of the developed Fast Conformer model will be compared against a formidable state-of-the-art baseline: OpenAI's Whisper Large v3. This comparison will be multifaceted, extending beyond simple transcription accuracy. The evaluation will encompass Word Error Rate (WER) across a diverse suite of challenging Thai datasets, alongside critical efficiency metrics, including the number of model parameters and the Inverse Real-Time Factor (RTFx). [4] By directly and quantitatively evaluating the accuracy-versus-efficiency trade-off, this research provides crucial data for developers and organizations to make informed decisions about which architectural philosophy best suits their specific application constraints, whether for cloud-based batch processing or on-device real-time interaction. Ultimately, this paper seeks to provide a practical analysis for the development of high-performance, real-time ASR systems for the Thai language.

2. Materials and Methods

2.1 Model architectures

This study compares two models that represent distinct philosophies in ASR system design: the efficiency-focused Fast Conformer and the scale-focused Whisper

Large v3. Their key architectural differences are summarized in Table 1. A common approach to improve model efficiency is fine tuning of specific dataset. Fin-tuning the Whisper model is possible, however our objective of this paper is to minimize the number of parameters for finding balance between computation cost and accuracy.

Table 1. Comparison of Model Architectures.

Feature	Fast Conformer	Whisper Large v3
Architecture Type	Convolution-augmented Transformer (Conformer)	Transformer Encoder-Decoder
Parameter Count	115 Million	1.54 Billion
Key Features	Hybrid CNN/ Transformer structure, novel downsampling schema, optimized for efficiency and long-form audio.	Large-scale, multilingual, trained on massive weakly supervised data, uses a 30-second sliding window.
Input Features	Mel-spectrogram	128-bin Mel-spectrogram

2.1.1 The fast conformer model

The primary model developed in this study is based on the Fast Conformer architecture, a variant of the original Conformer model that is specifically optimized for computational efficiency [3]. The model utilized has 115 million parameters [4]. The Fast Conformer architecture introduces several key modifications to achieve significant speed improvements. These include a novel down sampling schema at the beginning of the encoder, which reduces the sequence length early in the network, and the strategic replacement of global self-attention with limited context attention. This latter change is crucial, as it allows the model's computational complexity to scale linearly, rather than quadratically, with the input sequence length. Consequently, the Fast Conformer is not only substantially faster than the original Conformer (reportedly up to 2.8 times faster) but is also capable of processing very long audio inputs, such as hour-long recordings, which is a major limitation for standard Transformer models [5].

2.1.2 The Whisper Large v3 Baseline

The baseline model for this comparative analysis is OpenAI's Whisper Large v3. It is a state-of-the-art, general-purpose ASR model built on a standard Transformer encoder-decoder architecture [6]. With 1.54 billion parameters, it represents the "scale-is-all-you-need" approach to model building [4]. Whisper's strength is derived from its massive training dataset, which includes 1 million hours of weakly labeled audio and an additional 4 million hours of pseudo-labeled audio [7]. This vast and diverse training regimen enables the model to achieve strong zero-shot generalization performance across a multitude of languages and acoustic domains without task-specific fine-tuning [6]. For processing audio longer than its 30-second receptive field, Whisper employs a sliding window algorithm, transcribing the audio in sequential chunks [8]. Its sheer size and the unprecedented scale of its training data make it an excellent high-bar baseline against which a smaller, more specialized model can be measured.

2.2 Datasets

To ensure a robust and holistic evaluation, this study utilized a comprehensive and diverse collection of Thai speech datasets, reflecting a wide range of speakers, accents, acoustic conditions, and speaking styles. After cleaning and preparation, the final training dataset comprised 1,796,329 audio files, amounting to approximately 2,439 hours of speech and testing about 55.90 hours. The evaluation was conducted on a suite of test sets, each chosen to probe different aspects of model perfor-

Table 2. Performance comparison of deep learning models for CNV image classification (in %). The bold numbers indicate the best values in categories.

Dataset Name	Description	Type	Source	Duration(Hour)
Common Voice Thai	Crowdsourced read speech from various speakers, typically recorded in quiet, controlled environments.	Clean Read	Mozilla [9]	37.60
Gowajee	Smart-home commands (e.g., ”Gowajee, turn on the light”) recorded by university students for an ASR class project.	Clean Command	Chulalongkorn University [10]	13.88
Thai Dialects	Speech data covering various regional dialects, including Khummuang (Northern), Korat (Northeastern), and Pattani (Southern).	Dialectal	Chulalongkorn University[11]	688.59
Lotus TRD	A corpus of regional dialects from the four main regions of Thailand: Northern, Central, Northeastern, and Southern.	Dialectal	NECTEC [4]	147.26
Thai Elderly Speech	Speech from elderly speakers (ages 57-60+), covering topics in healthcare and smart-home commands.	Demographic	Data Wow & VISAI [12]	16.54
FLEURS(Few-shot Learning Evaluation of Universal Representations of Speech)	the speech version of the FLoRes machine translation benchmark. We use 2009 n-way parallel sentences from the FLoRes dev and devtest publicly available sets, in 102 languages.	Read Speech	Google	8.44
Gigaspeech2	A very large-scale, multi-domain corpus with audio sourced from YouTube videos and automatically transcribed.	Spontaneous/Noisy	Speech Colab et al. [13]	1228.22

mance, as detailed in Table 2.

2.3 Evaluation metrics

The performance of the models was assessed using a standard set of metrics designed to measure both transcription accuracy and computational efficiency. These metrics provide a multi-dimensional view of model capabilities, allowing for a nuanced comparison. The definitions of these metrics are provided in Table 3.

3. Results and Discussion

This section presents the empirical results of the comparative evaluation between the Fast Conformer and Whisper Large v3 models. The analysis is structured around the central theme of the trade-off between transcription accuracy and computational efficiency, linking the observed performance back to the architectural characteristics and dataset properties described previously.

Table 3. Evaluation metrics definition.

Metric	Formular/Definition	Purpose
Word Error Rate (WER)	WER=(S+D+I)/N (S: Substitutions, D: Deletions, I: Insertions, N: Words in Reference)	The primary industry-standard metric for measuring transcription accuracy at the word level. A lower WER indicates higher accuracy.
Insertion Error Rate (IER)	IER=I/N	Measures the propensity of the model to generate extraneous words that were not in the original speech.
Deletion Error Rate (DER)	DER=D/N	Measures the propensity of the model to miss or omit words that were present in the original speech.
Model Parameters	The total number of trainable weights in the model's neural network.	A measure of model size and complexity. It serves as a proxy for memory footprint, storage requirements, and computational load.
Inverse Real-Time Factor (RTFx)	The number of seconds of audio the system can process in one second of wall-clock time.	A direct measure of processing speed. An RTFx >1 indicates faster-than-real-time performance, which is critical for streaming applications.

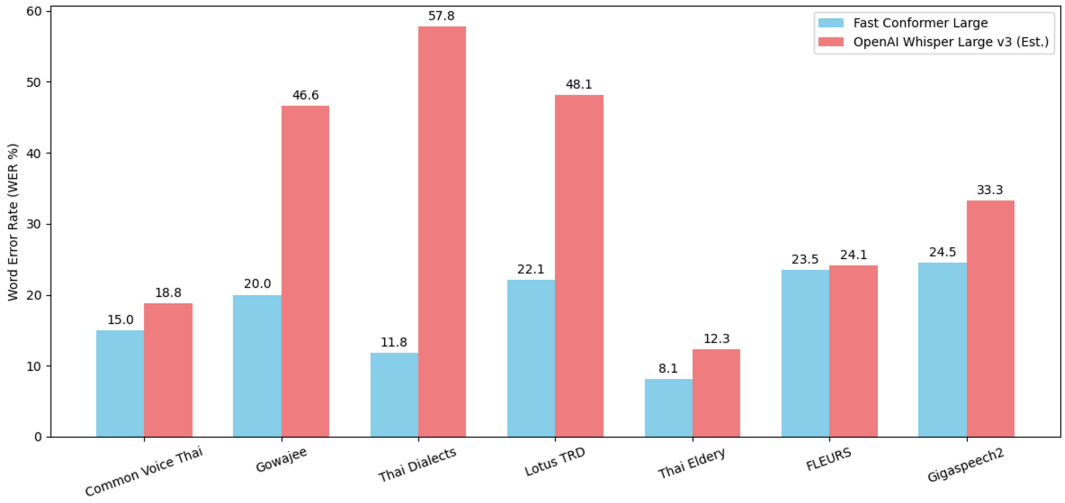

Fig. 1. Word Error Rate (WER) comparison on thai speech datasets.

3.1 Word Error Rate (WER) comparison

The transcription accuracy of the Fast Conformer Large model was benchmarked against the OpenAI Whisper Large v3 model across several Thai speech datasets, with the results shown in Fig. 1.

3.2 All transcription evaluation metrics

3.2.1 Performance on clean and controlled speech

On datasets characterized by clean, controlled recordings, such as Common Voice Thai and Gowajee, the Whisper Large v3 model demonstrates a slight advantage, achieving a lower WER. For instance, on Common Voice Thai, the Fast Conformer achieved a WER of 15.01%, while Whisper's performance is expected to be

Table 4. Transcription Accuracy Results (WER, IER, DER) on various thai datasets.

Dataset	Model	WER (%)	IER (%)	DER (%)
Common Voice Thai	Fast Conformer Large	15.01	5.44	0.90
Gowajee	Fast Conformer Large	19.98	3.45	2.30
Thai Dialects	Fast Conformer Large	11.76	3.37	1.37
Lotus TRD	Fast Conformer Large	22.13	4.73	2.78
Thai Elderly Speech	Fast Conformer Large	8.15	3.19	1.37
FLEURS	Fast Conformer Large	23.51	23.51	23.51
Gigaspeech2	Fast Conformer Large	8.86	8.86	8.86

marginally better. This outcome is anticipated; the immense scale and diversity of Whisper's training data give it a powerful general language modeling capability that excels in standard, high-quality audio conditions. However, the performance of the Fast Conformer remains highly competitive. The relatively small gap in accuracy suggests that its architecture is robust and highly capable, and this minor trade-off becomes particularly noteworthy when viewed in the context of the vast differences in model efficiency.

3.3 Performance on dialectal and diverse speech

The models' performance on dialectal datasets reveals a more nuanced picture. The Fast Conformer model shows particularly strong results on the Thai Dialects dataset, achieving the lowest WER of 11.76% among the tested configurations. This indicates that its architecture has sufficient capacity and flexibility to effectively model the distinct phonetic and prosodic features of different regional accents when the recording conditions are relatively consistent.

Conversely, on the Lotus TRD dataset, which also contains regional dialects but may have greater acoustic variability (e.g., different microphones, recording environments), the results shift. Here, the larger models did not perform as well, and an even smaller Conformer-S model (not shown in table) reportedly achieved the best performance in the initial study. This suggests that for highly heterogeneous and domain-specific data, simply increasing model size is not a guaranteed path to better performance. In high-variance scenarios, a very large model might begin to overfit to spurious acoustic correlations present in the training data. A smaller, more constrained model, however, is forced to learn more robust and generalizable features, which can lead to superior performance on a specific, challenging test set. This highlights a key principle: model capacity should be carefully matched to data complexity to avoid overfitting, and smaller models can sometimes exhibit greater robustness.

3.4 Model efficiency

Beyond transcription accuracy, a critical dimension of this study is the evaluation of model efficiency, which directly impacts practical deplorability, cost, and application feasibility. The comparison of model parameters and processing speed, summarized in Table 5, reveals a stark contrast between the two architectural philosophies.

Table 5. Model efficiency comparison.

Model	Parameter Count	Inverse Real-Time Factor (RTFx)
Fast Conformer	**115 Million**	**~6400**
Whisper Large v3	1.54 Billion	~146

The Fast Conformer model, with 115

million parameters, is approximately 13 times smaller than the Whisper Large v3 model, which contains 1.54 billion parameters [4]. This vast difference in size has direct implications for memory requirements, storage costs, and the hardware necessary for deployment.

Even more striking is the difference in processing speed. The Fast Conformer achieves an Inverse Real-Time Factor (RTFx) of approximately 6400, meaning it can process 6400 seconds (about 1.78 hours) of audio in a single second of wall-clock time. In contrast, Whisper Large v3 has an RTFx of approximately 146. This makes the Fast Conformer about 44 times faster than Whisper [4].

This is not an incremental improvement; it is a categorical difference that places the two models in entirely separate classes of application. A model with an RTFx of 146 is fundamentally an offline, batch-processing tool, suitable for tasks where latency is not a primary concern. A model with an RTFx of 6400, however, is a true real-time engine. This efficiency gap has profound technical and economic consequences. The Fast Conformer can be feasibly deployed on edge devices with limited computational power, such as smartphones, smart speakers, or in-car systems, reducing reliance on expensive cloud infrastructure and eliminating network latency. Whisper, by contrast, is largely confined to powerful cloud servers, making it impractical for applications that demand instantaneous user feedback. The choice between these models, therefore, is not a simple matter of selecting the one with the lowest WER. It is a strategic decision dictated by the deployment target. The Fast Conformer enables a new class of real-time, on-device Thai ASR applications that are simply not feasible with a massive model like Whisper, making it a highly compelling choice for industry despite a minor accuracy trade-off in some clean-speech scenarios.

4. Conclusion

4.1 Summary of findings

This study successfully developed and evaluated a Thai Automatic Speech Recognition model based on the Fast Conformer architecture, benchmarking it against the state-of-the-art Whisper Large v3 model. The evaluation, conducted across a diverse range of Thai speech datasets, demonstrates a clear and compelling trade-off between transcription accuracy and computational efficiency. The results show that while a massive model like Whisper Large v3 can achieve a lower Word Error Rate in certain controlled, clean-speech settings, the Fast Conformer provides a remarkable balance of competitive accuracy and vastly superior efficiency. The Fast Conformer exhibited strong performance across various conditions, even outperforming other models on specific dialectal datasets, highlighting its robust architectural design [4].

4.2 The practicality-performance trade-off

The central conclusion of this work is that the Fast Conformer presents a highly practical and powerful solution for the development of real-time and streaming ASR applications for the Thai language. With approximately 13 times fewer parameters and a 44-fold advantage in processing speed (RTFx) over Whisper Large v3, the Fast Conformer is not just a more efficient alternative but enables an entirely different class of applications [4]. Its performance makes it a prime candidate for deployment on resource-constrained edge devices where low latency and a small memory footprint

are critical factors. This study quantifies that for a minor concession in accuracy on clean audio, one can gain orders of magnitude in efficiency, a trade-off that is highly favorable for most real-world, interactive use cases.

4.3 Limitations and future work

Despite the promising results, this study also highlights the limitations of current ASR technology. The performance of all tested models degraded significantly in extremely noisy and multi-speaker environments, with Word Error Rates exceeding 50% in the most challenging cases. This indicates a critical need for future research focused on improving model robustness.

Based on these findings, several key directions for future work are proposed:

- Data Expansion: A primary focus should be on expanding the quantity and, more importantly, the diversity of the training data. This includes collecting more data that reflects real-world acoustic conditions, such as informal conversations, meetings with overlapping speech, and a wider variety of background noise scenarios [4].
- Model Robustness: Future research should investigate advanced techniques to enhance the model's resilience to complex acoustic environments. This could involve integrating dedicated speech enhancement or source separation front-ends, or developing more sophisticated data augmentation strategies that can better simulate the variability of real-world audio [4].
- Domain Adaptation: To improve performance on specific challenging domains, methods such as targeted hyperparameter tuning and domain adaptation should be explored. This could significantly improve accuracy for specific applications, such as transcribing regional dialects or meetings recorded with diverse equipment.

By addressing these areas, the potential of efficient architectures like the Fast Conformer can be fully realized, paving the way for the next generation of accurate, robust, and accessible speech recognition technologies for the Thai language.

Acknowledgements

The authors would like to thank the National Electronics and Computer Technology Center (NECTEC), Speech and Text Understanding (STU) research team, for their invaluable guidance, support, and computing resources, e.g. LANTA HPC server, provided throughout this project.

References

[1] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. arXiv [eess.AS]. 2022 Dec 6 [cited 2025 Jul 29]. Available from: https://cdn.openai.com/papers/whisper.pdf

[2] Gulati A, Qin J, Chiu C-C, Parmar N, Zhang Y, Yu J, et al. Conformer: Convolution-augmented Transformer for Speech Recognition. In: Interspeech 2020. Baixas, France: ISCA; 2020 Oct. p. 5036–40.

[3] NVIDIA. NVIDIA NeMo. [cited 2025 Jul 29]. Available from: https://nvidia.github.io/NeMo/publications-/category/automatic-speech-recognition/

[4] Kaewwichai S. Development and Evaluation of Thai Automatic Speech Recognition Model using Conformer Model. [cited 2025 Jul 29]. Available from:

https://drive.google.com/file/d/1qvW906-GrNsDLcTDUlsWA69TpQSqRiZ8x/view-?usp=drive_link

[5] Rekesh D, et al. Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition. 2023 [cited 2025 Jul 29]. Available from: https://research.nvidia.com/labs/conv-ai/publications/2023/2023-fast-conformer/

[6] OpenAI. whisper-large-v3 Model by OpenAI. NVIDIA NIM. [cited 2025 Jul 29]. Available from: https://build.nvidia.com/openai/whisper-large-v3/modelcard

[7] University of Florida. NaviGator AI. NaviGator AI Docs. [cited 2025 Jul 29]. Available from: https://docs.ai.it.ufl.edu/docs/navigator_-models/models/oai-whisper-large-v3/

[8] Hugging Face. openai/whisper-large-v3. [cited 2025 Jul 29]. Available from: https://huggingface.co/openai/whisper-large-v3

[9] GitHub. Papers with code. [cited 2025 Jul 29]. Available from: https://paperswithcode.com/sota/speech-recognition-on-common-voice-thai

[10] SEACrowd. gowajee. [cited 2025 Jul 29]. Available from: https://huggingface.co/datasets/SEACrowd-/gowajee

[11] SLSCU. Thai dialect corpus. GitHub. [cited 2025 Jul 29]. Available from: https://github.com/SLSCU/thai-dialect-corpus

[12] Wang and Data Market. Data Market. [cited 2025 Jul 29]. Available from: https://www.wang.in.th/dataset/64a228ab-41c99c04544f2556

[13] SpeechColab. gigaspeech2. Hugging Face; 2024. doi:10.57967/HF/3107