Advances in Artificial Intelligence and Machine Learning; Research (Ahead of Print) Received 10-05-2026; Accepted 04-09-2026; Published online 11-09-2026

Low-Resource Adaptation of Whisper and MMS for Lebanese Arabic Automatic Speech Recognition

Elias Elia 92030241@students.liu.edu.lb
Department of Computer & Communications Engineering
Lebanese International University
Beirut, Lebanon

Jennifer Takla 92030348@students.liu.edu.lb
Department of Computer & Communications Engineering
Lebanese International University
Beirut, Lebanon

Milia Habib milia.habib@liu.edu.lb
Department of Computer & Communications Engineering
Lebanese International University
Beirut, Lebanon

Mohamad Kannan mohamad.kanaan01@liu.edu.lb
Department of Electrical Engineering
Lebanese International University
Beirut, Lebanon

Rabih Rammal rabih.rammal@liu.edu.lb
Department of Electrical Engineering
Lebanese International University
Beirut, Lebanon

Zaher Merhi zaher.merhi@liu.edu.lb
Department of Computer & Communications Engineering
Lebanese International University
Beirut, Lebanon

Corresponding Author: Zaher Merhi

Abstract

Arabic automatic speech recognition has grown significantly in recent years, but performance remains limited for under-resourced dialects such as Lebanese Arabic, especially in real conversational scenarios where speakers switch between Lebanese Arabic, English, and French. This paper presents a low-resource dialect adaptation study on Lebanese speech transcription for meeting-oriented applications. Due to the lack of a clear open-source Lebanese speech corpus with aligned reference text, a custom dataset was created of approximately 1100 manually transcribed audio clips extracted from Lebanese podcast recordings. This paper investigates the effect of fine-tuning multiple pre-trained speech recognition models on this

Citation: Elias Elias, et al. Low-Resource Adaptation of Whisper and MMS for Lebanese Arabic Automatic Speech Recognition. **Advances in Artificial Intelligence and Machine Learning**. 2026. (Ahead of Print) https://dx.doi.org/10.54364/AAIML.2026.65343

dataset, focusing on Whisper Small, Whisper Medium, Whisper Large, and Massively Multilingual Speech (MMS). The study compares baseline and adapted performance to determine how model size and multilingual pretraining affect dialectal generalization. The results show that fine-tuning improved Whisper Small from 44.11% to 40.36% WER, Whisper Medium from 44.37% to 35.32% WER, and Whisper Large from 37.20% to 33.83% WER. MMS also improved from 76.92% to 66.94% WER in the same low-resource setting. These findings highlight the difficulty of low-resource dialect adaptation, particularly in the presence of code-switching and limited training data, and show that model size alone does not guarantee better adaptation. This study provides an initial Lebanese dialect speech resource and offers useful insights for developing ASR systems for Lebanese meeting transcription and meeting understanding tasks.

Keywords: Arabic ASR, Fine-tuning, Lebanese dialect, Whisper

1. INTRODUCTION

Automatic Speech Recognition (ASR) has become a critical technology in domains such as meeting transcription, subtitle generation, user interface and multilingual communication. Modern end-to-end models, such as Whisper, have improved speech recognition in a variety of languages, including Arabic [1]. However, progress in Arabic dialect remains uneven, with Modern Standard Arabic (MSA) compared to local spoken dialects [2]. Lebanese Arabic, in particular, is not represented in public ASR resources, despite its widespread use in professional and technical conversations [3]. Lebanese speech creates many challenges for ASR. It differs from MSA in pronunciation, vocabulary, and informal structure, and it is frequently combined with English and French using code-switching. Furthermore, it does not have a fully defined written form, making manual transcription and automatic evaluation more difficult. These problems decrease the efficacy of general multilingual ASR systems and can lead to transcription errors, especially in real-world meeting scenarios where mixed-languages terms are common [4]. A major limitation in this area is the lack of an available open-source Lebanese-only speech dataset with aligned reference transcripts for supervised adaptation. To fix this gap, this study uses a custom dataset derived from Lebanese podcast audio that is manually transcribed into speech-text. The dataset includes short clips that reflect typical Lebanese usage, such as informal usual speech and some language switching. This makes it suitable for testing low-resource dialect adaptation under real-world conditions [5, 6]. Motivated by recent Arabic ASR studies [7] that show multilingual models can benefit from fine-tuning on dialect data, this paper examines the adaptation of Whisper and MMS to Lebanese Arabic. This work compares the baseline and fine-tuned performance across multiple Whisper model sizes and MMS to show how model size and multilingual pretraining affect Lebanese dialect recognition. The main contribution of this paper is to evaluate the efficacy of multilingual ASR models in adapting to Lebanese Arabic using a small corpus dataset, as well as, to provide an initial baseline for future research on Lebanese speech related to transcription, translation, and meeting systems.

2. RELATED WORK

Talafha et al. (2025) [4], proposed a context-aware adaptation framework to improve Whisper's Arabic ASR without fine-tuning or changing the model architecture. Instead of updating the model weights, the authors included external context via decoder prompting and encoder prefixing. They used first-pass transcriptions from SeamlessM4T, TF-IDF to retrieve similar text, prompt reordering, and voice-cloned prefix speech to reduce speaker mismatch. The results showed a significant performance gap between MSA and dialectal speech. Whisper-large-v3 had an average WER/CER of 15.79%6.35% on MSA and 57.48% 25,58% on dialect Arabic. With context-aware decoding, the best MSA result improved to 12.27% 4.83% while dialect Arabic improved to 52,22% 20,01% using voice-clone prefixing. AlBlooki et al. (2025) [5], tested five ASR models on a 4-hour traditional Emirati Arabic dataset with limited resources, comparing zero-shot and fine-tuned performance. The baseline results show Wav2Vec2 as the strongest zero-shot model with 46.50% WER and 17.13% CER, while MMS achieved 67,21% WER and 24,56% CER. After fine-tuning, MMS showed the best performance, reducing WER/CER to 41.04%/13.34%. However, Whisper Small and Whisper Medium remained weak with 100.04%/75.53% and 88.60%/72.03%, respectively. Alkadri et al. (2025) [8], presented Arabic Little STT, a Levantine Arabic children's speech dataset consisting of 355 words recorded in the classroom from 288 children aged 6 to 13. They evaluated 8 Whisper variants without fine-tuning and reported large transcription errors. Whisper Large-v3 achieved the best result with 66% WER and 34% CER, while Whisper Large and Large-v2 achieved 82% WER and 49% CER. The study found that even strong ASR models struggle with under-represented Arabic speech conditions, especially if age, dialect and noisy classroom speech differ from adult training data.

Salhab et al. (2025) [9], proposed a weakly supervised Arabic ASR approach that involved training a Conformer-CTC model from scratch on around 15,000 hours of weakly annotated Arabic speech, including MSA and dialectal Arabic. The model was evaluated on SADA, Common Voice, MASC, Casablanca, and MGB-2, showing an average WER/CER of 28.68%/10.05%, compared to the best open-source baseline 24.74%/13.37%. The study found that large-scale weak annotation can improve Arabic ASR, but it requires far more training data and computational resources than low-resource fine-tuning methods. Mansour (2026) [6], studied low-resource Sudanese Arabic ASR using Whisper Small, Medium, and Large-v2, which included full fine-tuning, self-training, TTS augmentation and combined augmentation. Zero-shot Whisper performed poorly on Sudanese speech, with Whisper Large-v2 reaching 78.8% WER and 47.7% CER on OOOK-Eval. Full fine-tuning reduced the best WER to 62.8%. Whisper Medium, trained with clean Sudanese data, pseudo-labels, and TTS data, reached 57.9%/21.3% WER/CER on OOOK-Eval and 51.6%/26.5% on the holdout set, demonstrating the effectiveness of augmentation for low-resource dialect adaptation. Özyilmaz et al. (2025) [10], presented a fine-tuning method that tackled the problem of insufficient data in Arabic ASR across multiple dialects by fine-tuning an existing open-source whisper-small model with respect to Modern Standard Arabic (MSA) and five dialects (Gulf, Levantine, Iraqi, Egyptian, Maghrebi) based on Mozilla Common Voice and MASC corpora. The study focused on assessing the effect of the number of MSA training examples, MSA pre-training before dialectal fine-tuning, and dialect-specific versus pooled fine-tuning approaches. Their findings indicated that fine-tuning with only 20% of MSA data resulted in a drop in WER of almost 10%, allowing whisper-small to approach the MSA performance of whisper-large-v3 but achieving a speedup in latency from 0.29s to 0.12s per sample. The use of MSA pre-training provided little gain for dialectal fine-tuning due

to the low feature similarity between MSA and dialectal texts. Dialect-pooled models demonstrated comparable results compared to dialect-specific models with WER deterioration of only 8.42%.

Al Ali and Aldarmaki (2024) [11], proposed Mixat, a bilingual speech dataset for the Emirati dialect and code-switching between it and English. This dataset consists of 15 hours of publicly available Emirati podcasts, containing 5,316 utterances where Emirati-English code-switching occurs in 1,947 cases (36%), with the transcripts annotated in both the Arabic and Latin script to clearly distinguish between languages. The authors compared three pretrained ASR models on the dataset (Whisper, MMS, and ArTST), showing that all three exhibited poor results, with Whisper having the worst performance by far, obtaining 168.52% WER; MMS obtained 147.2% WER, while ArTST yielded the lowest (but nevertheless high) WER at 112.2%. It is noteworthy that multilingual models managed English segments with acceptable accuracy (Whisper: 12.06% WER) but exhibited extremely poor results when encountering Emirati Arabic words (195.98% WER).

2.1 Data collection

A custom dataset was created specifically for Lebanese dialect adaptation. Audio material was collected from publicly available Lebanese podcasts and YouTube discussion videos to capture natural spoken Lebanese Arabic speech, including informal vocabulary and conversational speech. The audio stream was downloaded, converted to a standard format, divided into shorter clips, and the resulting files were arranged into structured folders using an extraction pipeline. For every clip, metadata such as the segment index, podcast identifier, and file path were also created. The corpus has about 1100 clips, according to the most recent version of the dataset. These clips are brief conversational passages that were collected from a variety of Lebanese speakers, with a small but significant speaker diversity.

2.2 Audio Preprocessing

To meet the input requirements of Whisper-based ASR models, the collected audio segments got preprocessing. The preprocessing pipeline consists of converting the audio to mono, resampling it to 16 kHz, converting it to WAV format, trimming silence when necessary, and avoiding aggressive denoising in aim to maintaining the natural acoustic characteristics of Lebanese speech [1]. Additional signal conditions were applied in the implementation code using dynamic normalization and filtering. Next, speech intervals were identified, very short segments were eliminated, and adjacent speech segments divided by brief pauses were combined using voice detection. The resulting speech were exported as brief clips that could be evaluated and changed. The same general preprocessing approach was applied in all of the experiments involving Whisper's model sizes, and MMS.

2.3 Transcription And Annotation

Each clip went through a two-stage annotation process following preprocessing. In order to have an initial baseline prediction, a pretrained Whisper model was used to create a first-pass automatic transcription. Then, a manual reference transcription was made and examined to better represent

the spoken Lebanese dialect. As a result, the audio path, clip-level identifiers, model prediction, and manually verified reference text are all included in the final metadata sheet. A light text normalization was also used to reduce artificial variation in spelling, since Lebanese Arabic does not follow a fully standardized written form. This involved removing tatweel and diacritics, like converting "شووو" to "شو" and "أنا" to "انا", normalizing hamza and alif variants "أ", "إ" and "آ" into the unified form "ا" while preserving code-switched expressions when they appeared in the original speech. The same concept was implemented in the training code using a normalization function that eliminates punctuation and error marks, combines alif variants, and collapses repeated whitespace.

2.4 Train, Validation And Test Split

For supervised training, the final labeled dataset was cleaned to remove missing entries while retraining the main fields required for ASR, including the audio path and reference text. The dataset was then split into training, validation, and test subsets using a fixed random seed. The entire dataset was divided into 80% training and 20% portion was divided equally between validation and test sets. When possible, distribution by podcast source was used during splitting to maintain balance across the various collected recordings. The Whisper and MMS experiments used the same division logic, allowing baseline and fine-tuned results to be compared under the same conditions.

The overall dataset preparation and experimental workflow is illustrated in FIGURE 1.

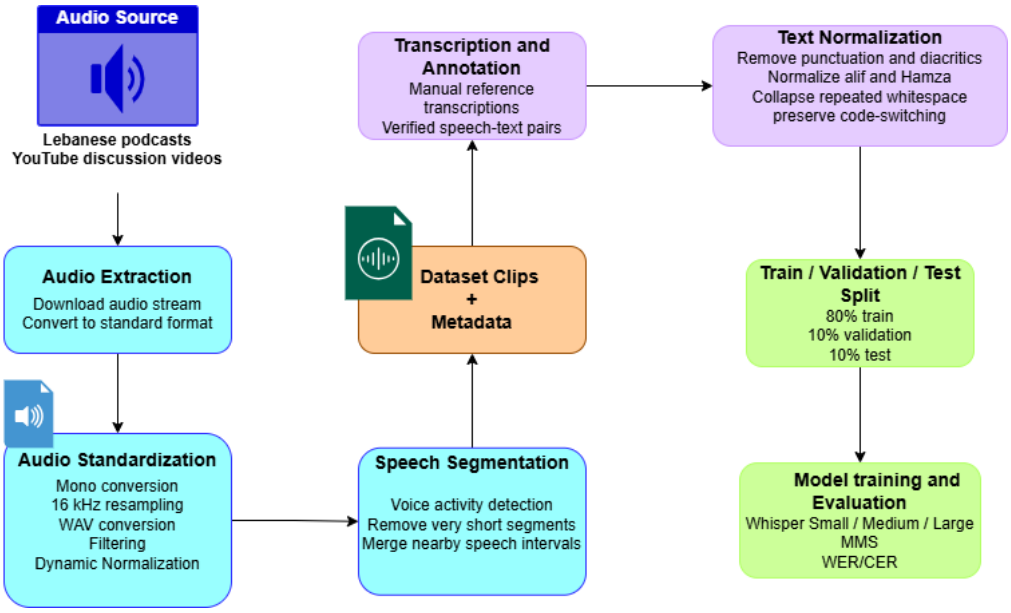

Figure 1: Pipeline of Lebanese speech dataset collection, preprocessing, annotation, splitting, and model evaluation.

3. EXPERIMENTAL SETUP

This section describes the experimental protocol used to evaluate the low-resource adaptation of multilingual ASR models to Lebanese Arabic. All models were tested under the same general conditions, each model was first evaluated in its baseline form on the test set, then fine-tuned on the Lebanese training data, validated on the validation split then re-evaluated on the same test set.

3.1 Evaluated Models

The experiments focused on a group of multilingual speech recognition models chosen to represent various parameter scales and adaptation capacities within the same Lebanese dialect setting. The Whisper family, specifically Whisper Small, Whisper Medium, and Whisper Large, because they are widely used in multilingual ASR research and provide a useful range of trade-offs between computational cost and recognition accuracy [1]. Whisper Small was included as a lightweight model that can be trained and deployed with limited hardware resources, making it suitable for low-cost adaptation scenarios. Whisper Medium was chosen as an intermediate model that may provide a better balance of model capacity and fine-tuning efficiency, especially in low-resource settings where the dataset is limited but suitable for supervised adaptation. Whisper Large was chosen as the study's highest-capacity Whisper variant, giving the highest expected baseline transcription quality and serving as a useful reference for determining how much additional benefit fine-tuning can still affect when the model is already strong. In addition to the Whisper variants, MMS was used in the study to see if broader multilingual pretraining can improve Lebanese dialect recognition in low-resource setting. Unlike the Whisper family, MMS represents an unique ASR architecture, providing a useful comparison beyond model size alone [7].

Table 1: General characteristics of the multilingual ASR models evaluated in this study. [1, 7]

Model	Approx. Params	Architecture
Whisper Small	244M	Encoder–Decoder Transformer
Whisper Medium	769M	Encoder–Decoder Transformer
Whisper Large	1.55B	Encoder–Decoder Transformer
MMS	∼1B	Wav2Vec2-based CTC model

3.2 Fine-Tuning Configuration

To ensure a fair comparison across all experiments, the same dataset split (train, validation and test) without any modification, and a single consistent preprocessing pipeline for all data is used. This involved normalizing Arabic text by removing diacritics, unifying alef variants, removing punctuation, and standardizing whitespace. Simulations were done on Google Colab using A100 GPU. The same evaluation metrics were used for all models, namely Word Error Rate (WER) and Character Error Rate (CER), which were computed after applying the same normalization to both predictions and reference transcripts. All models were trained under nearly identical conditions using the HuggingFace trainer with common settings such as learning rate, number of epochs, opti-

mizer, evaluation strategy, and optimizer. Fine-tuning was performed using HuggingFace Seq2Seq Trainer with AdamW optimizer, a learning rate of 2×10^{-6}, 50 warmup steps, 8 epochs, batch size 4, and gradient accumulation of 2. Models were evaluated and saved after each epoch, with the best checkpoint selected using WER, generation was limited to 128 tokens, and FP16 was enabled when CUDA was available [12]. Due to GPU limitations, gradient accumulation was used to simulate larger batch sizes, and FP16 mixed precision training was enabled whenever possible. Whisper models were fine-tuned using their encoder-decoder sequence-to-sequence structure through the HuggingFace Seq2SeqTrainer. For MMS, which uses a Wav2Vec2-based CTC architecture, the feature extractor was frozen, while the acoustic encoder and CTC output layers were updated.

3.3 Evaluation Metrics

To evaluate the proposed models, two standard benchmarks were used: Word Error Rate (WER) and Character Error Rate (CER) [13]. WER is the go-to metric for checking how many words the model misclassify—whether it swapped a word, missed one entirely, or added something that wasn't there. Naturally, a lower score means a more accurate transcription. However, because Arabic is so complex and spelling can vary, CER can be used to ensure a fair comparison. CER breaks words down to the character level, giving a much clearer picture of the small errors that WER might overlook. All calculations were handled using the JiWER library [14]. To make sure the comparison was fair, we applied the exact same text normalization to both, the Algorithm's predictions and the original human produced transcripts. Examining both metrics together provided an accurate assessment of the models. The below equations details WER and CER computation.

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

where S, D, and I represent substitutions, deletions, and insertions of words, respectively, and N is the total number of words in the reference transcript.

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

where S, D, and I represent character substitutions, deletions, and insertions, respectively, and N is the total number of characters in the reference transcript.

4. RESULTS AND DISCUSSION

4.1 Quantitative Results

The performance of the evaluated models was measured using Word Error Rate (WER) and Character Error Rate (CER). TABLE 2 summarizes the results for the Whisper model family, comparing their performance before and after fine-tuning on the Lebanese Arabic dataset.

TABLE 2 shows that fine-tuning reduced the WER and CER for all evaluated models. Whisper Medium achieved the largest WER reduction, decreasing WER from 44.37% to 35.32%, Whisper Large obtained the best value for WER at 33.84%, Whisper Small showed only modest improve-

Table 2: Baseline vs. Fine-tuned Performance on the Lebanese Test Set.

Model	Baseline WER	Final WER	Baseline CER	Final CER
Whisper Small	44.11%	40.36%	17.17%	16.17%
Whisper Medium	44.37%	35.32%	17.06%	14.42%
Whisper Large	37.20%	33.83%	14.20%	13.46%
MMS	76.92%	66.94%	31.16%	23.73%

ment. MMS also improved after fine-tuning, reducing WER from 76.92% to 66.94%. The CER results followed the same trend, showing that fine-tuning also reduces character-level errors.

4.2 Whisper Model Size Comparison

The experiments conducted resulted in the following differences in adaptation depending on Whisper's scale.

Whisper Medium: The best results were obtained using this model, with its WER decreasing from 44.37% to 35.32%. It can be concluded that the "Medium" scale has a good combination of model capacity required for the adaptation of a dialect in low-resource conditions such as Lebanese Arabic.

Whisper Large: This model demonstrated the best result before tuning (37.20%) and after tuning (33.83%). The percentage of its improvement was lower than that of the "Medium" scale. In other words, this model incorporates strong multilingual and Arabic knowledge, but it requires a greater amount of data to improve further.

Whisper Small: This scale demonstrated the smallest improvement of about 3.75%. Despite its advantage in efficiency, the small number of parameters makes it incapable of adapting to all the peculiarities of the considered dialect.

4.3 Discussion

The results of the current study offer valuable information about the difficulties in Lebanese ASR.

Dialect Complexity: The Lebanese Arabic dialect is very complex. As evidenced by high baselines (37%–44%), it lacks standard grammar, uses code-switching from Arabic to French or English, and relies heavily on conversational language.

Low-Resource Limitations: As discussed above, the custom dataset used in this paper is rather limited (about 1100 clips). This imposes constraints on the adaptability of the model. Fine-tuning becomes more prone to overfitting because the amount of available training data is limited.

Model Size and Fine-Tuning: Contrary to popular belief, the size of the model does not have a strictly linear effect on performance during fine-tuning. Whisper Medium demonstrates a substantial

performance gain compared to the baseline. Hence, the optimal model size for fine-tuning in this setting can be considered the medium scale.

Error Analysis: Most remaining errors were caused by Lebanese dialect complexity, spelling variations, and code-switching. The models sometimes replaced informal Lebanese words with MSA or other dialect forms, or inserted extra words during uncertain decoding. English and French terms were sometimes deleted, incorrectly translated or combined with nearby Arabic words. These errors indicate that better Lebanese ASR requires more dialect-specific data and improved handling of mixed-language conversational speech.

5. CONCLUSION AND FUTURE WORK

This paper described a low-resource adaptation study for Lebanese Arabic (ASR) recognition using Whisper and MMS models. A custom Lebanese Arabic speech dataset was created using podcast and YouTube discussion audio, which was then transcribed, normalized and split into training, validation, and test sets. The experiments compared baseline and fine-tuned performance based on WER and CER. Future work should concentrate on expanding the dataset with more speakers, longer meeting-style recordings, and more equal gender and regional coverage. Finally, the ASR output can be combined with translation and summarization models to help achieve the goal of producing accurate meeting transcripts.

6. AUTHOR CONTRIBUTION

All authors have equally contributed for this work in Conceptualization, Data collection, Data analysis, Writing – review and editing.

References

[1] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. In: International conference on machine learning. PMLR. 2023;28492-28518.

[2] Habash NY. Introduction to Arabic Natural Language Processing. Morgan & Claypool; 2010.

[3] Talafha B, Kadaoui K, Magdy SM, Habiboullah M, Chafei CM, et al. Casablanca: Data and Models for Multidialectal Arabic Speech Recognition. In: Proceedings of the 2024 conference on empirical methods in natural language processing. Stroudsburg, PA, USA: Association for Computational Linguistics. 2024;21745-21758.

[4] Talafha B, Alhassan AA, Abdul-Mageed M. Context-Aware Whisper for Arabic Asr Under Linguistic Varieties. 2025. ArXiv preprint: https://arxiv.org/pdf/2511.18774

[5] AlBlooki M, Inui K, Shehata S. ASR Models for Traditional Emirati Arabic: Challenges, Adaptations, and Performance Evaluation. In: Proceedings of the 8th international conference on natural language and speech processing (ICNLSP-2025). ACL. 2025:42-49.

[6] Mansour A. Doing More With Less: Data Augmentation for Sudanese Dialect Automatic Speech Recognition. 2026. ArXiv preprint: https://arxiv.org/pdf/2601.06802

[7] Pratap, et al. Scaling Speech Technology to 1,000+ Languages. J Mach Learn Res. 2024;25:1-52.

[8] Alkadri M, Desouki D, Jallad KA. Arabic Little STT: Arabic Children Speech Recognition Dataset. 2025. ArXiv preprint: https://arxiv.org/pdf/2510.23319.

[9] Salhab M, Elghitany M, Sait S, Ullah SS, Abusheikh M, Abusheikh H. Advancing Arabic Speech Recognition Through Large-Scale Weakly Supervised Learning. In: 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA). IEEE. 2025:1-7.

[10] Özyilmaz ÖT, Coler M, Valdenegro-Toro M. Overcoming Data Scarcity in Multidialectal Arabic Asr via Whisper Fine-Tuning. 2025. ArXiv preprint: https://arxiv.org/pdf/2506.02627

[11] Al Ali MK, Aldarmaki H. Mixat: A Data Set of Bilingual Emirati-English Speech. Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-Resourced Languages@ LREC-COLING 2024. 2024:222-226.

[12] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-Of-The-Art Natural Language Processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. Association for Computational Linguistics. 2020:38-45.

[13] Morris AC, Maier V, Green PD. From WER and RIL to Mer and WIL: Improved Evaluation Measures for Connected Speech Recognition. In: Interspeech. ISCA. 2004:2765-2768.

[14] `https://github.com/jitsi/jiwer`.