## Arabic Little STT: Arabic Children Speech Recognition Dataset

Mouhand Alkadri, Dania Desouki, Khloud Al Jallad

Department of Information and Communication Engineering,

Arab International University, Daraa, Syria.

\* Corresponding author. E-mail(s): <u>k-aljallad@aiu.edu.sy</u>, <u>201820115@aiu.edu.sy</u>, <u>202020003@aiu.edu.sy</u>

#### **Abstract**

The performance of Artificial Intelligence (AI) systems fundamentally depends on high-quality training data. However, low-resource languages like Arabic suffer from severe data scarcity. Moreover, the absence of child-specific speech corpora is an essential gap that poses significant challenges. To address this gap, we present our created dataset, Arabic Little STT, a dataset of Levantine Arabic child speech recorded in classrooms, containing 355 utterances from 288 children (ages 6–13). We further conduct a systematic assessment of Whisper, a state-of-the-art automatic speech recognition (ASR) model, on this dataset and compare its performance with adult Arabic benchmarks. Our evaluation across eight Whisper variants reveals that even the best-performing model (Large-v3) struggles significantly, achieving a 66% word error rate (WER) on child speech – starkly contrasting with its sub-20% WER on adult datasets. These results align with other research on English speech. Results highlight the critical need for dedicated child speech benchmarks and inclusive training data in ASR development. Emphasizing that such data must be governed by strict ethical and privacy frameworks to protect sensitive child information We hope that this study provides an initial step for future work on equitable speech technologies for Arabic-speaking children, enriching the children's demographic representation in ASR datasets.

**Keywords**: Automatic Speech Recognition (ASR), Arabic ASR, Child speech recognition, Arabic dialects, Whisper model, ASR evaluation, low-resource languages.

### 1. Introduction

Automatic Speech Recognition (ASR) has been a hot research topic in NLP in the last several years. Beyond its importance in research, its capabilities extend from powering virtual assistants, voice-controlled devices to enabling hands-free data entry and improving accessibility for kids and for adults with disabilities. Notably, ASR is also playing an increasingly transformative role in the online education landscape. ASR is heavily used in content transcription, which does not only result in a more streamlined education process, but also creates equal opportunities for content comprehension for students with hearing disabilities. Furthermore, ASR systems offer significant advantages for language learning applications, especially those that evaluate pronunciation and provide feedback. These important applications of ASR in EdTech highlight a profound opportunity to enhance the online educational process for children, who naturally tend to interact with the world through speech rather than text. This preference for speech-based interaction highlights the critical need for accurate ASR tailored to children for an effective learning experience. ASR task has large-scale datasets, such as TED-LIUM (Hernandez et al., 2018; Rousseau et al., 2012, 2014), LibriSpeech (Librispeech: An ASR Corpus Based on Public Domain Audio Books / IEEE Conference Publication | IEEE Xplore, n.d.) and CommonVoice (Ardila et al., 2020), as well as sophisticated models, such as FastConformer-Hybrid PCD ([2507.13977] Open Automatic Speech Recognition Models for Classical and Modern Standard Arabic, n.d.), Whisper (Radford et al., n.d.), and Wav2Vec (Schneider et al., 2019).

Although ASR has large-scale datasets and sophisticated models, most of them target adult speech. Thus, most publicly available datasets lack children's voices. This gap increases in low-resources languages such as Arabic, where dialectal diversity, morphological complexity, and the absence of child-specific speech corpora is an essential gap that poses significant challenges. To overcome this gap, we created a dataset for children's voices and named it Arabic Litte STT<sup>1</sup>. Moreover, we evaluated the Whisper (Radford et al., n.d.)

-

<sup>&</sup>lt;sup>1</sup> Dataset available at this link https://huggingface.co/datasets/little-stt/little-stt-dataset

models on our newly developed dataset of Arabic-speaking children's voices, comparing its performance with performance on adult speech datasets.

### 2. Related Works

## 2.1. Speech Datasets

Automatic Speech Recognition (ASR) models rely heavily on high-quality datasets. For widely represented languages like English, numerous datasets exist, such as LibriSpeech (*Librispeech: An ASR Corpus Based on Public Domain Audio Books | IEEE Conference Publication | IEEE Xplore*, n.d.) and the three versions of TED-LIUM (Hernandez et al., 2018; Rousseau et al., 2012, 2014).

There are many multilingual datasets, such as Common Voice (Ardila et al., 2020) and FLEURS (Conneau et al., 2022) include underrepresented languages like Arabic.

Moreover, there are dedicated Arabic datasets like Casablanca dataset (Talafha et al., 2024) which represents different Arabic dialects.

Table 1 shows a comparison between some available ASR datasets.

Dataset	Size	Audio Source	Arabic Representation
LibriSpeech	1K hours	Read speech (Audiobooks)	No
TED-LIUM (Release 1)	118 hours	TED Talks	No
TED-LIUM (Release 2)	207 hours	TED Talks	No
TED-LIUM (Release 3)	452 hours	TED Talks	No
Common Voice V17 <sup>2</sup>	31175 hours total, 20408 validated hours	Crowdsourced read speech (various sources like blogs, books, movies)	157 Hours
FLEURS	~12 hours per language	Read speech (based on n-	~12 hours

<sup>&</sup>lt;sup>2</sup> https://huggingface.co/datasets/mozilla-foundation/common\_voice\_17\_0

\_

	(102 language)	way parallel sentences from	
		FLoRes-101 MT	
		benchmark)	
Casablanca	~48 hours	YouTube episodes from TV	~48 hours (across 8
		series	dialects)

Table 1. Comparison between some available ASR datasets

However, none of the previously mentioned datasets take children's speech into consideration, a gap that disproportionately affects underrepresented languages and leaves ASR systems ill-equipped to handle the acoustic and linguistic nuances of younger speakers.

## 2.2. Children's Speech Dataset

There are few available resources for children's speech, such as the CMU Kids Corpus (*The CMU Kids Corpus - Linguistic Data Consortium*, n.d.) 5,180 utterances from 76 English-speaking children, PF\_STAR (*(PDF) The PF\_STAR Children's Speech Corpus*, n.d.) 60 hours of non-native English from European children, and MyST (Pradhan et al., 2023) 400 hours of conversational English.

Moreover, these datasets are exclusively English-centric.

Although it is a linguistically diverse language, Arabic has no dedicated resources for children's speech.

Table 2 shows a comparison between some available children ASR datasets.

Dataset	Size	Audio Source
CMU Kids Corpus	9 hours	Read-aloud sentences by children
PF_STAR	60 hours	Recorded as part of the EU FP5 PF STAR project
MyST	400 hours	Recorded sessions as part of My Science Tutor project

Table 2. Comparison between available children specific ASR datasets

#### 2.3. Child-Specific ASR Adaptations

Efforts to optimize ASR systems for children's speech have focused on English. Jain et al. adapted Whisper to English children speech (Jain et al., 2023). Shi et al. applied Test-Time adaptation (TTA) methods—A process allows adult targeted ASR models to continuously adapt to each child speaker at test time—to improve ASR system on children's voices (Shi et al., 2024). These advances highlight the potential for adaptation but remain inaccessible to low-resource languages due to data scarcity.

#### 2.4. ASR State-of-the-Art Models

Significant advancements in Automatic Speech Recognition (ASR) have been largely driven by open-source models. Notable examples include FastConformer-Hybrid PCD ([2507.13977] Open Automatic Speech Recognition Models for Classical and Modern Standard Arabic, n.d.), Whisper (Radford et al., n.d.) a transformer-based multilingual model, Whisper-Medusa (Segal-Feldman et al., 2024) an optimized variant of Whisper to accelerate inference, Canary (Puvvada et al., 2024) an ASR model that Surpasses Whisper in low-resource languages and Wav2Vec (Schneider et al., 2019) a pioneer in unsupervised pretraining for speech recognition.

Table 3 shows a comparison between the state-of-the-art ASR models.

Model	Architecture	Supports Arabic	
FastConformer-Hybrid PCD	Conformer Based	Yes	
Whisper	Transformer Based	Yes (Multilingual Variant)	
Whisper-Medusa	Transformer Based	No	
Canary	Transformer Based	No	
Wav2Vec	Convolution Based	Yes (XLS-R Variant)	

Table 3. Comparison between the state-of-the-art ASR models

Comparing these models on the Open Universal Arabic ASR Leaderboard (Wang et al., 2024) which benchmarks open-source models for Arabic ASR tasks, reveals that Whisper large v3 achieved an average WER of 36.86. As of August 16, 2025, this placed it third on the leaderboard. NVIDIA's FastConformer

models occupied the top two positions, achieving the best scores using pure greedy decoding and language model (LM) integration, respectively.

#### 3. Our Created Dataset (Arabic Little STT)

Collecting child speech data for ASR poses unique challenges, as it requires navigating parental consent and institutional approvals. These factors contribute to the scarcity of Arabic child speech corpora. To address this scarcity and manage Arabic's inherent linguistic complexity, we refined our scope to focus specifically on the Levantine (Shami) dialect, a widely used dialect in the region of Levant, and more specifically we covered the Syrian variant of the Levantine dialect. Our resulting dataset comprises 355 utterances recorded from 288 children (157 male, 131 female), aged 6 to 13 years.

Aiming to reflect a realistic ASR use case, all audio was captured in classroom environments using standard smartphone microphones. This setting naturally introduced moderate levels of ambient noise, including sounds such as keyboard clicks and peer conversations. Recordings were stored in the AAC format, and later converted to WAV. The average duration of individual audio clips is approximately 10 seconds, accompanied by transcriptions averaging 13 words per utterance (observed range: 3 to 62 words).

Figure 1 is a stacked bar chart of the age and gender of the participating children, where it shows that most children were of ages 8 and 9. The minimum age group was 12 as we have only 6 children. Participating kids were more males than females in most ages.

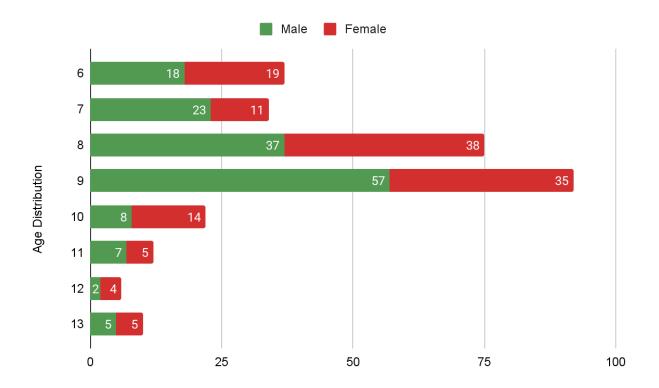


Figure 1. A stacked bar chart of the age and gender of the participated children

Transcription content was deliberately centered on themes of programming, robotics, and artificial intelligence, directly mirroring the Information Technology (IT) classroom context in which the recordings were conducted.

Table 4. Samples from transcripts in our dataset.

يعني ممكن يصير في روبوتات او اغراض الكترونية تساعدك بالأدوات المنزلية
بدي اخترع روبوت بنضف وبساوي اكل مشان ما نتعب حالنا
مثلا سيارة تيسلا بتسوق لحالها بس تحددلها الموقع بتروح لحالها بتبطىء السرعة وبتزود
أنو الناس يلي برمجو نسيو شغلة او حطو شغلة زيادة
انو البيت بالمستقبل بكون مثل الآلي بنضف البيت لبين ما روح واجي وبساوي الأكل لبين ما روح واجي

Table 4. Samples from transcripts in our dataset

#### 3.1. Audio Preprocessing

Following the initial collection phase, each record was manually verified for audibility (e.g., sufficient volume, minimal background noise) and discarded if it was not audible for a human. The resulting filtered set was manually transcribed following Whisper's non-English conventions (Radford et al., n.d.) All transcribers were native Levantine (Shami) Arabic speakers ensuring accuracy and language context awareness.

To further standardize the evaluation, Arabic-specific normalization was applied simplifying the Arabic writing into a commonly used form, while preserving the meaning for the human reader:

- **Diacritic Removal** Stripped Harakat (Fatha: فتحة / \, Damma: ضمة / \, Kasra: كسرة / \, Sukun: كسرة / \) and Shadda (- شدة \)
- **Tatweel Elimination**: Removed elongated characters (e.g., –) to align with modern Arabic text standards.
- **Alef Normalization** Unified [1,1], into 1, reducing orthographic variability.

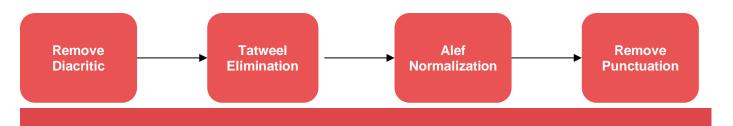


Figure 2. An illustration of the pipeline used to normalize Arabic transcription

# 4. Evaluation Experiment

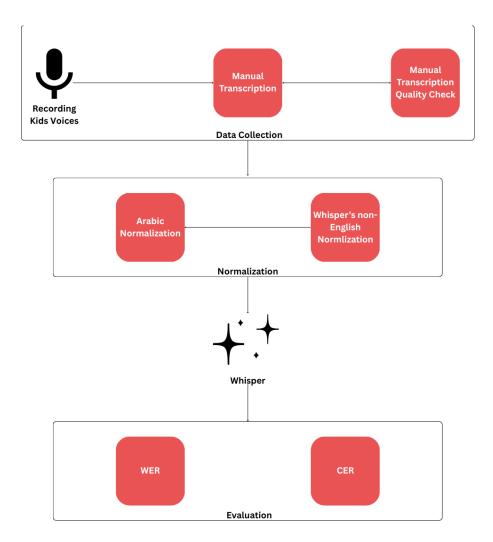


Figure 3. A block diagram highlighting the full pipeline

After exploring different state-of-the-art Automatic Speech Recognition (ASR) models, we chose the Whisper models family for testing our dataset. The decision was made for two reasons:

- Proven Strong Performance in Arabic: Whisper models perform very well with Arabic speech.

  As supported by its high ranking (third place) on the Open Universal Arabic ASR Leaderboard

  (Wang et al., 2024)
- Availability of Similar Published Results for Children's Speech: Researchers (Jain et al.) have already tested Whisper's performance on datasets of English-speaking children. As this evaluation exists, we wanted to use Whisper on our Arabic children's dataset. This allows us to see if the

patterns and challenges found when recognizing English children's speech also appear when recognizing Arabic children's speech.

## 4.1 Whisper

Whisper (Radford et al., n.d.) is a Transformer based encoder-decoder model, also referred to as a sequence-to-sequence model. The Whisper models are trained on over 680,000 hours of audio and the corresponding transcripts collected from the internet, 65% of this data (or 438,218 hours) represents English-language audio and matched English transcripts, roughly 18% (or 125,739 hours) represents non-English audio and English transcripts, while the final 17% (or 117,113 hours) represents non-English audio and the corresponding transcript. This non-English data represents 98 different languages, Arabic is one of which with 739 hours of transcribed audio.

Multiple variants of the Whisper model were released with sizes ranging from tiny (39M parameters) to Large (1.5B parameters), some of which are released into two different settings, Multilingual and English only. The multilingual models were trained on both speech recognition and speech translation. During training, tasks are distinguished using special tokens such as <|transcribe|> for speech recognition and <|translate|> for translation into English. Table 5 shows a comparison between whisper variants

Variant	Number of Parameters	English-only Model Available	Release Date
Tiny	39 M	Yes (tiny.en)	September 2022
Base	74 M	Yes (base.en)	September 2022
Small	244 M	Yes (small.en)	September 2022
Medium	769 M	Yes (medium.en)	September 2022
Large	1550 M	No	September 2022
Large v2	1550 M	No	December 2022
Large v3	1550 M	No	November 2023
Large v3 Turbo	809 M	No	November 2023

Table 5. Comparison between Whisper variants

#### 4.2. Experiment Setup

Eight Whisper model variants—Tiny³, Base⁴, Small⁵, Medium⁶, Large⁻, Large-v2⁶, Large-v3⁶, and Large-v3 Turbo¹⁰— were evaluated spanning parameter sizes from 39M (Tiny) to 1.5B (Large-v3). This coverage reflects diverse deployment scenarios: smaller models (e.g., Tiny) suit resource-constrained edge devices, while larger variants (e.g., Large-v3) target high-accuracy cloud-based applications given that ASR systems operate in diverse environments with varying computational limits (e.g., mobile app, servers). By testing Whisper's full spectrum of sizes, we provide actionable insights for developers balancing speed, memory, and accuracy trade-offs in child-centric applications.

Regarding methodology, as Whisper models are multi-task models capable of transcription and translation (Radford et al., n.d.), the experiment was conducted using transcription mode, more specifically language detection and transcription as no language prompt was provided to the model, only relying on the model's built-in language detection. Furthermore, to ensure a fair comparison, the model's output was normalized using the same steps described earlier for our dataset's transcription.

Finally, evaluation was measured using two standard ASR metrics: Word Error Rate (WER) and Character Error Rate (CER) (Srivastav et al., 2023; Wang et al., 2024)

#### 4.3. Results

All evaluated Whisper variants demonstrated strong language detection accuracy, particularly the Large variant, which achieved near-perfect detection (>99%) for Arabic (Figure 5).

<sup>&</sup>lt;sup>3</sup> https://huggingface.co/openai/whisper-tiny

<sup>&</sup>lt;sup>4</sup> https://huggingface.co/openai/whisper-base

<sup>&</sup>lt;sup>5</sup> https://huggingface.co/openai/whisper-small

<sup>&</sup>lt;sup>6</sup> https://huggingface.co/openai/whisper-medium

<sup>&</sup>lt;sup>7</sup> https://huggingface.co/openai/whisper-large

<sup>&</sup>lt;sup>8</sup> https://huggingface.co/openai/whisper-large-v2

<sup>&</sup>lt;sup>9</sup> https://huggingface.co/openai/whisper-large-v3

https://huggingface.co/openai/whisper-large-v3-turbo

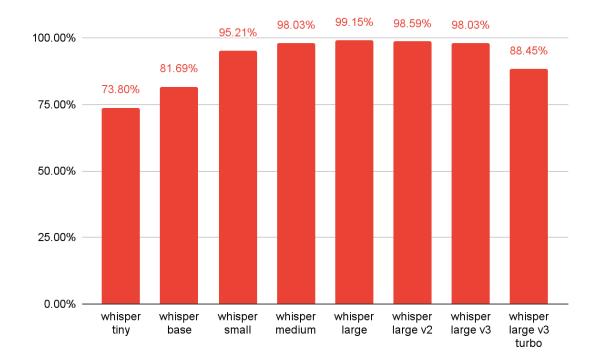


Figure 4. The language detection accuracy for each of the evaluated Whisper models

Despite this high language detection accuracy, all models exhibited significant transcription errors on the Arabic Little STT dataset. Detailed CER and WER results are presented in Table 6.

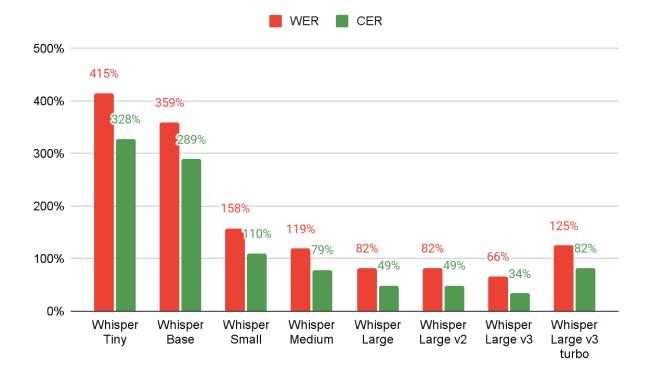


Figure 5. CER and WER results of the evaluated Whisper models on our dataset

Model	WER	CER
Whisper Tiny	415%	328%
Whisper Base	359%	289%
Whisper Small	158%	110%
Whisper Medium	119%	79%
Whisper Large	82%	49%
Whisper Large v2	82%	49%

Whisper Large v3	66%	34%
Whisper Large v3 turbo	125%	82%

Table 6. WER results of the evaluated Whisper models on our dataset

A substantial performance drop on child speech can be observed compared to results reported on adultcentric Arabic datasets (Common Voice 9 - Arabic and FLEURS - Arabic) (Radford et al., n.d.). For example, Whisper Large-v3 achieved a WER of 66% on Little STT-4.1× higher than its 16.0% WER on FLEURS (Table 7).

Model	Our Dataset	Common Voice 9 - Arabic	FLEURS - Arabic
Whisper Tiny <sup>11</sup>	415%	90.9%	63.4%
Whisper Base <sup>12</sup>	359%	84.4%	48.8%
Whisper Small <sup>13</sup>	158%	66.4%	30.6%
Whisper Medium <sup>14</sup>	119%	60.3%	20.4%
Whisper Large <sup>15</sup>	82%	56.0%	18.1%
Whisper Large v2 <sup>16</sup>	82%	53.8%	16.0%

Table 7. Comparing the WER result for the evaluated Whisper results on our dataset and the results reported in Whisper Paper (Radford et al., n.d.)

https://huggingface.co/openai/whisper-tiny
 https://huggingface.co/openai/whisper-base

https://huggingface.co/openai/whisper-small

<sup>14</sup> https://huggingface.co/openai/whisper-medium

<sup>15</sup> https://huggingface.co/openai/whisper-large

<sup>&</sup>lt;sup>16</sup> https://huggingface.co/openai/whisper-large-v2

This degradation aligns with Jain et al findings in English child speech evaluations, where Large-v2 performed 20% worse in WER for children compared to adults —Taking the best results for both cases—. The significant increase in transcription errors observed in Arabic, potentially compounded by its linguistic complexity, supports a consistent pattern. Collectively, these results indicate a systemic limitation: ASR models trained primarily on adult data struggle to generalize effectively to children's speech, irrespective of language or model size.

#### 5. Conclusion

Arabic children's speech remains a critically underrepresented demographic and linguistic domain in ASR research, hindering advancements in child-centric voice technologies. To overcome this gap, we created a dataset for Arabic children's voices, specifically in Levantine dialect, comprising 355 utterances from 288 children (ages 6–12). We conducted several experiments on our created data by evaluating the performance of Whisper models. Results show a significant performance gap: Whisper's best-performing model (Large v3) achieved a WER of 66% on our dataset, 4.1× higher than its reported WER on adult Arabic benchmarks like FLEUR. This mirrors findings in English child ASR, where Whisper's errors increase by 20% for children compared to adults, suggesting a systemic limitation of current ASR systems. Our work highlights an urgent priority for creating child-inclusive ASR datasets that capture developmental variability. We do hope that our created dataset will help the development of children Arabic ASR. Our future efforts could explore increasing the size of the dataset, or expanding it to cover other dialects to address the challenges in child speech recognition.

# 7. Acknowledgements

We gratefully acknowledge the Genius Planet team for their collaborative spirit and essential support in the collection and curation of the children's speech dataset, a cornerstone of this work.

### References

- [2507.13977] Open Automatic Speech Recognition Models for Classical and Modern Standard Arabic.

  (n.d.). Retrieved August 16, 2025, from https://www.arxiv.org/abs/2507.13977
- Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). *Common Voice: A Massively-Multilingual Speech Corpus* (No. arXiv:1912.06670). arXiv. https://doi.org/10.48550/arXiv.1912.06670
- Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., & Bapna, A. (2022). FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech (No. arXiv:2205.12446). arXiv. https://doi.org/10.48550/arXiv.2205.12446
- Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., & Estève, Y. (2018). *TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation* (Vol. 11096, pp. 198–208). https://doi.org/10.1007/978-3-319-99579-3\_21
- Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., & Cucu, H. (2023). Adaptation of Whisper models to child speech recognition (No. arXiv:2307.13008). arXiv. https://doi.org/10.48550/arXiv.2307.13008
- Librispeech: An ASR corpus based on public domain audio books / IEEE Conference Publication / IEEE Xplore. (n.d.). Retrieved May 4, 2025, from https://ieeexplore.ieee.org/document/7178964
- (PDF) The PF\_STAR children's speech corpus. (n.d.). Retrieved May 3, 2025, from https://www.researchgate.net/publication/221491189\_The\_PF\_STAR\_children's\_speech\_corpus
- Pradhan, S. S., Cole, R. A., & Ward, W. H. (2023). My Science Tutor (MyST)—A Large Corpus of Children's Conversational Speech (No. arXiv:2309.13347). arXiv. https://doi.org/10.48550/arXiv.2309.13347
- Puvvada, K. C., Żelasko, P., Huang, H., Hrinchuk, O., Koluguri, N. R., Dhawan, K., Majumdar, S., Rastorgueva, E., Chen, Z., Lavrukhin, V., Balam, J., & Ginsburg, B. (2024). Less is More:

  Accurate Speech Recognition & Translation without Web-Scale Data (No. arXiv:2406.19674).

- arXiv. https://doi.org/10.48550/arXiv.2406.19674
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (n.d.). *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv.Org. Retrieved May 3, 2025, from https://arxiv.org/abs/2212.04356v1
- Rousseau, A., Deléglise, P., & Estève, Y. (2012). TED-LIUM: An Automatic Speech Recognition dedicated corpus. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC`12)* (pp. 125–129). European Language Resources Association (ELRA). https://aclanthology.org/L12-1405/
- Rousseau, A., Deléglise, P., & Estève, Y. (2014). Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC`14)* (pp. 3935–3939). European Language Resources Association (ELRA). https://aclanthology.org/L14-1079/
- Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-training for Speech Recognition (No. arXiv:1904.05862). arXiv. https://doi.org/10.48550/arXiv.1904.05862
- Segal-Feldman, Y., Shamsian, A., Navon, A., Hetz, G., & Keshet, J. (2024). Whisper in Medusa's Ear:

  Multi-head Efficient Decoding for Transformer-based ASR (No. arXiv:2409.15869). arXiv.

  https://doi.org/10.48550/arXiv.2409.15869
- Shi, Z., Srivastava, H., Shi, X., Narayanan, S., & Matarić, M. J. (2024). Personalized Speech Recognition for Children with Test-Time Adaptation (No. arXiv:2409.13095). arXiv. https://doi.org/10.48550/arXiv.2409.13095
- Srivastav, V., Majumdar, S., Koluguri, N., Moumen, A., Gandhi, S., & others. (2023). *Open Automatic Speech Recognition Leaderboard*. Hugging Face. https://huggingface.co/spaces/hf-audio/open\_asr\_leaderboard
- Talafha, B., Kadaoui, K., Magdy, S. M., Habiboullah, M., Chafei, C. M., El-Shangiti, A. O., Zayed, H.,

- tourad, M. cheikh, Alhamouri, R., Assi, R., Alraeesi, A., Mohamed, H., Alwajih, F., Mohamed, A., Mekki, A. E., Nagoudi, E. M. B., Saadia, B. D. M., Alsayadi, H. A., Al-Dhabyani, W., ... Abdul-Mageed, M. (2024). *Casablanca: Data and Models for Multidialectal Arabic Speech Recognition* (No. arXiv:2410.04527). arXiv. https://doi.org/10.48550/arXiv.2410.04527
- The CMU Kids Corpus—Linguistic Data Consortium. (n.d.). Retrieved May 3, 2025, from https://catalog.ldc.upenn.edu/LDC97S63
- Wang, Y., Alhmoud, A., & Alqurishi, M. (2024). Open Universal Arabic ASR Leaderboard. *arXiv Preprint arXiv:2412.13788*.