

AraYoungVoices: A DIVERSE L1/L2 CORPUS OF ARABIC CHILD AND ADOLESCENT SPEECH

Shammur Absar Chowdhury, Zien Sheikh Ali, Houssam Eddine-Othman Lachemat, Hamdy Mubarak

Qatar Computing Research Institute, Qatar

ABSTRACT

State-of-the-art ASR systems primarily target native adult speech, leading to substantial performance gaps for children, adolescents, and L2 speakers. We introduce **ARAYoungVoices**, a 151.72-hour Arabic read-speech corpus from 286 speakers aged 7–18, comprising **ARAKIDS** (7–12) and **ARATEENS** (13–18). The corpus includes 146 native Arabic (L1) and 140 second-language (L2) speakers, with native speakers spanning Egyptian, Gulf, Levantine, and North African dialectal backgrounds and L2 speakers representing diverse linguistic backgrounds across the Americas, Asia, Africa, and Europe. We benchmark four pretrained ASR models under zero-shot and fine-tuned settings using unseen-speaker-&-unseen-prompt (USUP) and unseen-speaker-&-seen-prompt (USSP) evaluations. Results show that L2 speech remains substantially more challenging than L1 speech, with the largest errors observed mainly for younger L2 speakers. Age-specific fine-tuning improves the matched age group, while joint fine-tuning provides a stronger balance across populations. ASR hypotheses are also consistently closer to the standard reading prompt than to the verbatim transcription, particularly for L2 speech, suggesting partial normalization of reading deviations.

Index Terms— Arabic speech recognition, child and adolescent speech, non-native speech, speech corpus, age-aware ASR.

1. INTRODUCTION

Automatic speech recognition (ASR) has improved significantly with large pretrained multilingual speech models trained on massive amounts of speech data. However, performance remains uneven across age groups and language backgrounds. Children, adolescents, and second-language (L2) speakers remain particularly challenging because pronunciation, fluency, and acoustic characteristics differ from native adult speech and evolve with age [1, 2]. Adolescent speech, which captures the transition from child to adult speech, remains comparatively understudied.

Existing child-speech corpora cover several languages and speaking styles. Early resources such as OGI Kids, PF-STAR, CHOREC, and SingaKids supported research on child ASR [3, 4, 5, 6], while more recent corpora include MyST, BK3AT, ChildMandarin, and ChildTalk [7, 8, 9, 10]. Resources on adolescence or L2 speech are fewer comparatively [11, 12, 13]. For Arabic, available child-speech resources remain limited. Arabic Little STT, for example, contains only 355 utterances from children aged 6–13 [14]. Table 1 summarize the current resources.

This gap is particularly important for Arabic, where speech varies across major dialectal backgrounds such as Egyptian, Gulf, Levantine, and North African Arabic, while Modern Standard Arabic is

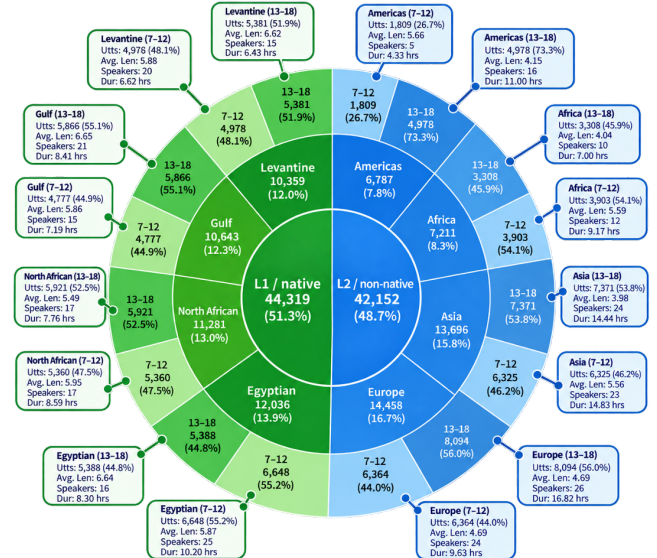


Fig. 1: Overview of ARAYoungVoices. Avg. Len.: average prompt length in words.

primarily used in formal and educational settings [15, 16]. For L2 speakers, pronunciation is further influenced by first language, proficiency, and other language exposure. Consequently, robust speech models for young Arabic speakers therefore must account for age, L1/L2 background, and dialectal variation.

To reduce the gap, we introduce ARAYoungVoices, an Arabic read-speech corpus containing **151.72 hours**, **86,471 utterances**, and **286 speakers aged 7–18**. It comprises **ARAKIDS** (7–12) and **ARATEENS** (13–18), with **146 native Arabic (L1)** and **140 non-native speakers (L2)**. L1 speakers comes from Egyptian, Gulf, Levantine, and North African dialectal backgrounds, while L2 speakers represent diverse language backgrounds across Africa, the Americas, Asia, and Europe (see Figure 1).

To understand the effectiveness of the corpus, we evaluate contemporary ASR models in zero-shot and fine-tuned settings using two complementary test conditions: USUP, with unseen speakers and unseen prompts, and USSP, with unseen speakers reading prompts observed during training. Our analysis examines how recognition varies across age, L1/L2 background, and dialect, and whether age-specific or joint fine-tuning reduces these gaps.

Our contributions are: (i) **ARAYoungVoices**, a large-scale Arabic corpus spanning children and adolescents across diverse L1 and L2 backgrounds; (ii) age- and nativity-aware ASR benchmarks with complementary speaker and prompt generalization settings; (iii) analysis of recognition gaps across age, L1/L2 status, and dialectal

The work is supported by the HBKU Flagship Research Grant (HBKU-INT-VPR-FRG-03-09).

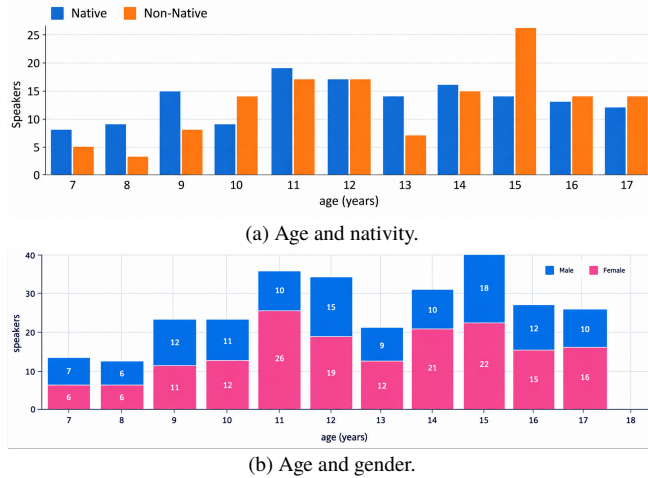
Table 1: Available Child and Adolescent speech corpora grouped by age coverage. K: kindergarten, G: grade.

Primarily Child-Focused					Child-Adolescent / Adolescent				
Corpus	Language	Age	Speakers	Hours	Corpus	Language	Age	Speakers	Hours
CHOREC [5]	NL	6–12	400	130	OGI Kids [3]	EN	K–G10	1,100	–
SingaKids [6]	ZH	7–12	255	75	PF-STAR [4]	Multi.	4–14	–	>60
ChildMandarin [9]	ZH	3–5	397	41.25	TLT-school [11]	L2 EN/DE	9–16	2,710	–
MyST [7]	EN	G3–G5	~1,300	~400	Estonian Adolescent [12]	ET	9–18	309	~70
BK3AT [8]	Multi.	G1–G3	–	122	Ph@tSessionz [13]	DE	12–20	~1,000	–
ChildTalk ^a [10]	ZH	2–8	498	112.5	ARAYoungVoices	AR	7–18	286	151.72
Arabic Little STT [14]	AR	6–13	288	–					

Table 2: Average utterance duration (s) by age and nativity.

	7–12	13–18
Native	5.36	5.12
Non-native	8.45	6.67

backgrounds; and (iv) release¹ of fine-tuned ASR models for young Arabic speech and along with annotation guidelines.

**Fig. 2:** Speaker characteristics in ARAYoungVoices.

2. ARAYoungVoices

ARAYoungVoices is an Arabic read-speech corpus designed to study variation across age and language background. It contains **151.72 hours** of speech and 86,471 utterances from **286 speakers** aged 7–18. The corpus comprises two age groups: ARAKIDS (7–12) and ARA TEENS (13–18), with 146 native Arabic (L1) and 140 non-native Arabic (L2) speakers.

The L1 speakers represent four major Arabic dialectal backgrounds – *Egyptian, Gulf, Levantine, and North African*, while the L2 speakers span diverse linguistic backgrounds across *Asia, Europe, the Americas, and Africa*. This design supports controlled analysis of age, L1/L2 background, and dialectal variation in speech and other NLP models. Figure 1 provides an overview of the corpus composition, while Figure 2 summarizes speaker distributions by age, gender, and nativity. Table 2 reports the average utterance duration, and Table 3 lists the top-10 countries by utterance count.

¹Link removed for anonymity.

Reading Material For recording, reading prompts are tailored to age, school grade, and Arabic proficiency. L1 ARAKIDS use fully diacritized Grade 1–4 material, including stories, simple sentences, and isolated words for younger readers. L1 ARA TEENS, on the other hand, use non-diacritized Grade 9–12 material with longer sentences and passages. L2 material is selected separately by age and Arabic proficiency, using simpler content for children and dedicated non-native Arabic material for adolescents.

Data Collection and Metadata Speech was recorded on mobile devices through a dedicated collection platform, with guardian consent obtained for minors. Speaker metadata includes age, gender, nationality, mother tongue, L1/L2 status, years of learning Arabic, multilingual background, and education level, together with basic recording environment information. No personally identifiable information was retained. All recordings undergo quality control, and samples with incomplete or incorrect readings, excessive noise, clipping, or corrupted audio are removed.

Data Splits We construct speaker-disjoint training, development, and test splits. The training set includes L1 and L2 speakers from both ARAKIDS and ARA TEENS, while the development set is held out from training and contains only L1 speakers. We define two test conditions: USUP (*Unseen Speaker&Unseen Prompt*), where both speakers and prompts are unseen during training and development, and USSP (*Unseen Speaker&Seen Prompt*), where unseen speakers read prompts that also occur in training. USUP contains 2,500 utterances: 2,000 from ARAKIDS using 5–7 L1 speakers per region (~250 utterances), and 500 from ARA TEENS using two speakers per region (~62 utterances). USSP uses the same speaker pool as USUP, but with recordings of prompts seen during training. Detailed statistics are given in Table 5.

Prompt–Verbatim Annotation For USUP, we have both the original reading prompt and a manually produced verbatim transcript of what the speaker actually said. Reading deviations are explicitly annotated as substitutions, insertions, and deletions, while repetitions, stuttering, fillers, unclear speech, non-speech events, and speaker overlap are also marked. Dialectal and pronunciation-specific realizations are preserved, using extended Arabic characters when needed for sounds not represented in standard MSA orthography. These paired prompt–verbatim references support both ASR evaluation and fine-grained analysis of reading, pronunciation, and dialectal variation.

Reading Deviation Analysis We then analyze the reading deviations in the manually transcribed USUP set by comparing each original prompt with its verbatim transcription. Character-level alignment is used to identify insertions, deletions, and substitutions, providing a fine-grained view of reading variation across L1/L2 speakers and

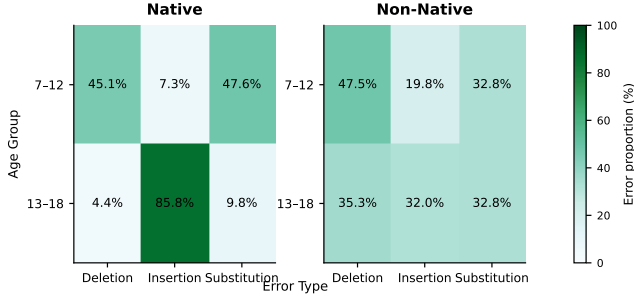


Fig. 3: Char-level reading deviations in the USUP set across age groups.

Table 3: Top-10 countries by utterances.

#	Country	Utts. (%) / Spk.	#	Country	Utts. (%) / Spk.
1	Egypt	14,315 (16.6%) / 48	6	Morocco	5,818 (6.7%) / 17
2	India	8,081 (9.3%) / 26	7	Jordan	5,787 (6.7%) / 21
3	France	7,448 (8.6%) / 25	8	Palestine	5,485 (6.3%) / 16
4	U.S.	6,351 (7.3%) / 20	9	Pakistan	3,996 (4.6%) / 14
5	Algeria	6,165 (7.1%) / 19	10	Iraq	3,462 (4.0%) / 11

age groups (see Figure 3). Reign-wise, L1 Egyptian speakers showed the highest deletion (62.7%) and substitution (58.6%) rates, whereas L2 speakers exhibited more balanced error patterns across regions. Our initial exploration of prompt-verbatim alignments reveals clear differences between L1 and L2 speakers.

Deletions are the largest error category for both groups and are dominated by short-vowel and tanwīn drops. For L1 speakers, final short-vowel deletion is frequent and largely reflects pausal pronunciation and loss of grammatical markers (*iʿrāb*) rather than true reading errors. L2 speakers, however, show substantially more gemination (*shadda*) drops than L1 speakers (120 vs. 21), consistent with difficulty in producing consonant length contrasts. They also omit long-vowel *alif* more often (65 vs. 31), effectively reducing long vowels to short ones.

Substitutions are dominated by vowel changes. Consistent with the deletion patterns above, L1 speakers frequently reduce short-vowel endings to sukūn: 371/882 (42%) of vowel changes are vowel→sukūn, consistent with reduced final *iʿrāb* realization. L2 speakers instead tend to substitute vowels toward the default fathā, with 184/534 (34%) moving toward /a/ and → being the most common pattern (104, 19.5%). The /u/→/a/ pattern is also more prominent in L2 speakers (30% vs. 14%). Overall, L1 speakers **omit**, whereas L2 speakers **regularize** short vowels. Consonant substitutions, however, provide a clearer signal of speaker background: for L1 speakers, these patterns largely reflect dialectal variation, whereas for L2 speakers, they more often indicate phonological gaps in Arabic. Table 4 summarizes these substitution patterns.

Insertions also differ across groups. For L2 speakers, *fathā* is the most frequent insertion, suggesting a default-vowel strategy when the intended vowel is uncertain. Insertions of *alif* and *yāʾ*² further suggest vowel lengthening, particularly among speakers from the Americas and Asia; the same pattern appears for L2-Teens, but less frequently. For L1-Teens, the high insertion rate is largely a diacritization artifact: inserted *fathā*, *kasra*, *ḍamma*, *sukūn*, and *shadda* occur across dialect regions because speakers restore diacritics absent from the undiacritized reference. These are scored as errors despite representing valid readings, while L1-Kids show fewer such insertions because their prompts are fully diacritized.

Table 4: Frequent consonant substitutions with linguistic interpretation; Buckwalter transliteration and frequency are shown in parentheses. ج uses extended symbol c.

Substitution	Distribution	Interpretation
ج→ح (j→c)	L1; Egypt (94); L2; All regions (4)	Dialectal variation
أ→ع (ʾ→ʿ)	L2; All regions (63)	Pharyngeal → glottal
س→ث (s→ṯ)	L2; Africa/Asia (31); L1; Egypt/North Africa (16)	Interdental → alveolar
ث→ت (ṯ→t)	L2 All reigns (23); L1; North Africa/Gulf (6)	Interdental → alveolar
ذ→د (ẓ→d)	L1; North Africa (3); L2; All reigns (10)	Dialectal/L2 variation
ض→ظ (ẓ→ḏ)	L1; North Africa (5)	Emphatic merger
ج→ع (g→c)	L2; Africa (5)	Uvular → velar
ق→ك (q→k)	L2; Africa/America/Asia (5)	Uvular → velar
ث→ت (ṯ→t)	L1; North Africa (4)	Interdental → alveolar
ز→ذ (z→ẓ)	L1; Egypt (3)	Interdental → sibilant

Table 5: Corpus statistics across splits. Per split reported utterances (speakers/hrs.).

Split	Kids-L1	Kids-L2	Teens-L1	Teens-L2	Total
Train	12,572 (48/17.89)	7,340 (24/16.41)	19,705 (59/27.51)	21,957 (68/41.13)	61,574 (199/102.94)
Dev	949 (3/1.58)	— (—/—)	762 (2/1.18)	— (—/—)	1,711 (5/2.75)
USUP	1,000 (26/1.57)	1,000 (40/2.43)	250 (8/0.40)	250 (8/0.38)	2,500 (82/4.78)
USSP	7,242 (26/11.40)	10,061 (40/24.36)	1,839 (8/2.97)	1,544 (8/2.52)	20,686 (82/41.25)
All	21,763 (77/32.43)	18,401 (64/43.20)	22,556 (69/32.07)	23,751 (76/44.03)	86,471 (286/151.72)

3. EXPERIMENT SETUP

Models We evaluate four recent Arabic and multilingual speech models: AURA-STT (347M),² COHERE TRANSCRIBE ARABIC (2B),³ QWEN2.5-OMNI-ARABIC (7B) [17], and WHISPER LARGE-V3 (1.55B) [18]. The set includes dedicated Arabic ASR models and general-purpose omni models, allowing us to compare Arabic specialization along with age and nativity.

Adaptation Strategy We first evaluate each model in a zero-shot (ZS) setting without adaptation. We then fine-tune using three configurations: ARAKIDS only (AK), ARATEENS only (AT), and the joint training data (JT). The same training splits are used across models, with speakers kept disjoint between training and evaluation.

3.1. Evaluation Measure

We evaluate all systems on USUP and USSP using WER and CER. The same Arabic normalization is applied to the reading prompt (*P*), verbatim transcription (*V*), and ASR hypothesis. Verbatim markup for non-speech events, fillers, overlap, unintelligible speech, speaker labels, and control tokens is removed, while repetitions are preserved. We then remove diacritics, tatweel, punctuation, and normalize Alef variants (أ, إ, آ, إ) (><) to ا (A). Dialectal letters such as گ, ف, ظ, چ, گ, and پ are kept, as are the distinctions ه/ه (p/h) and ی/ی (Y/y).

²<https://www.api.fanar.qa/>

³<https://huggingface.co/CohereLabs/cohere-transcribe-03-2026>

⁴Special letters are not found in Buckwalter transliteration.

Table 6: WER (%) on ARAYOUNGVOICES. USUP reports WER_V / WER_P , where V and P denote the human-corrected verbatim transcription and reading prompt and USSP reports WER_P . ZS: zero-shot; AK, AT, and JT: fine-tuning on ARAKIDS, ARATEENS, and joint data. Best results are in **bold**.

Model	FT	Kids-L1	Kids-L2	Teens-L1	Teens-L2
Useen-Speaker&Unseen-Prompt USUP (WER_V / WER_P)					
Aura-Off. 342M	ZS	10.71/8.25	39.60/33.90	11.64/8.38	36.79/28.92
	AK	5.57/2.98	21.16/13.39	10.77/7.28	26.31/16.37
	AT	9.49/7.35	31.72/25.79	12.75/9.20	23.41/13.34
	JT	8.74/6.32	24.94/17.53	13.50/10.19	22.41/11.55
Cohere-Ar 2B	ZS	6.60/4.60	42.88/37.70	9.90/6.69	39.91/32.62
	AK	6.65/3.91	20.57/12.94	12.22/8.91	25.86/16.70
	AT	7.56/5.13	29.33/22.69	10.54/6.93	22.63/14.91
	JT	6.55/3.98	20.35/12.92	10.13/6.69	20.62/11.43
Useen-Speaker&Seen-Prompt USSP (WER_P)					
Aura-Off. 342M	ZS	7.24	34.03	9.89	27.71
	AK	2.10	9.76	9.18	16.43
	AT	8.36	25.77	3.49	2.92
	JT	2.36	12.16	3.55	3.25
Cohere-Ar 2B	ZS	4.37	39.97	7.65	28.89
	AK	2.66	9.79	10.25	19.47
	AT	5.19	21.55	4.60	3.28
	JT	2.44	9.87	4.64	2.90

Model	FT	Kids-L1	Kids-L2	Teens-L1	Teens-L2
Useen-Speaker&Unseen-Prompt USUP (WER_V / WER_P)					
Qwen-O-Ar 7B	ZS	9.57/7.17	37.83/32.49	14.44/11.58	31.33/23.43
	AK	10.53/8.33	27.99/21.90	23.28/20.95	29.54/21.52
	AT	14.85/12.65	38.74/34.02	18.10/14.84	28.76/21.64
	JT	11.15/8.83	25.37/18.99	15.95/12.81	26.09/17.83
Whisper-v3 1.55B	ZS	11.81/9.76	47.15/43.00	14.32/12.22	38.35/32.29
	AK	10.31/7.99	26.61/20.06	19.97/16.88	30.55/22.42
	AT	13.29/11.21	36.28/31.27	13.56/10.48	24.08/15.02
	JT	9.98/7.78	26.86/20.55	13.39/9.95	25.08/16.37
Useen-Speaker&Seen-Prompt USSP (WER_P)					
Qwen-O-Ar 7B	ZS	7.61	34.21	15.07	26.80
	AK	2.24	12.57	23.28	31.86
	AT	13.31	34.21	3.81	1.48
	JT	2.38	11.63	5.22	1.40
Whisper-v3 1.55B	ZS	10.02	40.16	15.21	40.71
	AK	4.42	14.37	18.22	26.38
	AT	10.82	29.14	6.59	2.62
	JT	4.28	14.58	6.73	2.86

Table 7: Representative dialectal substitutions seen in ASR; Buckwalter transliteration is shown in parentheses.

Pattern	Group	Type
ع → ا (E → A)	L2	Pharyngeal
ق ↔ ك, ق → غ (q ↔ k, q → g)	L2	Uvular
ج → ح (g → j)	Gulf	Dialectal
ط → ت (T → t)	Gulf	De-emphasis
ظ/ذ/ض → د (D → Z/z/d)	North Africa	Emphatic
س/ث → ت (s ↔ t)	L1	Interdental
ة → ت (p → t)	L1/L2	Connected speech

4. RESULTS AND ANALYSIS

Table 6 compares ASR performance across age, L1/L2 background, and fine-tuning settings.

Age and L1/L2 Effects Our results show that L2 speech is consistently more difficult than L1 speech across all models. In the ZS setting, L2 children also have higher WER than L2 adolescents. Fine-tuning substantially reduces this gap, particularly for L2 speech. On USSP, age-matched fine-tuning generally favors the target group such that AK performs improves for ARAKIDS, while AT favors ARATEENS. On the more difficult USUP test set, joint fine-tuning is often competitive or better, suggesting a stronger balance across the full 7–18 age range.

Prompt vs. Verbatim Speech For every USUP setting, WER_P is lower than WER_V , with a larger difference for L2 speech. This shows that ASR outputs are often closer to the intended reading prompt than to the speaker’s verbatim transcription. This is consistent with partial normalization of reading deviations. We analyze the prompt (P), verbatim transcription (V), and ASR hypothesis (H) jointly to distinguish preserved deviations ($H \approx V$) from normalization toward the prompt ($H \approx P$).

Pronunciation and Error Patterns Manual alignment reveals systematic age-, L2-, and dialect-related variation. L2 children, for example, show more frequent deletions and insertions involving ʕ and ʔ , while several consonant substitutions reflect regional or L2

pronunciation patterns (Table 7). We also observe model-related orthographic errors, including frequent ʕ insertions and hamza-seat variations such as $\text{ʕ} \rightarrow \text{ʔ}$ ($\text{ʔ} \rightarrow \text{ʕ}$) and $\text{ʔ} \rightarrow \text{ʕ}$ ($\text{ʕ} \rightarrow \text{ʔ}$). These should be distinguished from genuine pronunciation differences when evaluating young-speaker ASR.

Key Findings Overall, L2 speech remains substantially more challenging than L1 speech, with majority of the errors coming from younger L2 speakers. Age-specific fine-tuning improve the matched age group performance, while joint fine-tuning shows a stronger balance across age-nativity groups. Moreover, WER_P is consistently lower than WER_V , particularly for L2 speakers, suggesting that ASR outputs often move toward the intended prompt rather than fully preserving the dialectal information. Error patterns further show systematic age-, L2-, and dialect-related pronunciation differences alongside model-specific orthographic and decoding errors.

5. CONCLUSIONS

We introduced ARAYOUNGVOICES, a diverse Arabic speech corpus and, to our knowledge, the first to jointly cover child and adolescent speech from ages 7–18 across both L1 and L2 speakers. Following, we benchmarked the datasets with state-of-the-art ASRs in zero-, and fine-tuned setting. Our results show substantial performance gaps for L2 speech, particularly among younger speakers, while age-specific and joint fine-tuning significantly improve recognition. Analysis of prompt–verbatim differences further reveals systematic age-, dialect-, and L2-related reading variation, as well as a tendency of ASR systems to normalize some deviations toward the intended prompt. We hope this work supports more robust and inclusive Arabic ASR for young speakers.

Limitation and Distribution: A limitation of the ARAYOUNGVOICES is that it focuses on read speech and is not always uniformly balanced across age, regional, and L1/L2 groups. Moreover, as the recordings involve minors and include demographic and linguistic metadata, the raw audio and full speaker-level metadata will not be publicly released. We will instead release de-identified aggregate statistics, evaluation protocols, and resulting ASR models, subject to the applicable consent and institutional requirements.

References

- [1] Matteo Gerosa, Diego Giuliani, Shrikanth S. Narayanan, and Alexandros Potamianos, “A review of ASR technologies for children’s speech,” in *Proceedings of the 2nd Workshop on Child, Computer and Interaction (WOCCI)*. 2009, pp. 1–8, ACM.
- [2] Jiaming Zhou, Shiyao Wang, Shiwan Zhao, Jiabei He, Haoqin Sun, Hui Wang, Cheng Liu, Aobo Kong, Yujie Guo, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin, “ChildMandarin: A comprehensive mandarin speech dataset for young children aged 3–5,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, 2025, pp. 12524–12537, Association for Computational Linguistics.
- [3] Khaldoun Shobaki, John-Paul Hosom, and Ronald A. Cole, “The OGI kids’ speech corpus and recognizers,” in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, 2000, pp. 258–261.
- [4] Anton Batliner, Mats Blomberg, Shona D’Arcy, Daniel Elenius, Diego Giuliani, Matteo Gerosa, Christian Hacker, Martin Russell, Stefan Steidl, and Michael Wong, “The PF_STAR children’s speech corpus,” in *Proceedings of Interspeech*, 2005, pp. 2761–2764.
- [5] Leen Cleuren, Jacques Duchateau, Pol Ghesquière, and Hugo Van hamme, “Children’s oral reading corpus (CHOREC): Description and assessment of annotator agreement,” in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008, European Language Resources Association (ELRA).
- [6] Nancy F. Chen, Rong Tong, Darren Wee, Peixuan Lee, Bin Ma, and Haizhou Li, “SingaKids-Mandarin: Speech corpus of singaporean children speaking mandarin chinese,” in *Proceedings of Interspeech*, 2016, pp. 1545–1549.
- [7] Sameer Pradhan, Ronald A. Cole, and Wayne H. Ward, “My science tutor (MyST)—a large corpus of children’s conversational speech,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, Torino, Italia, 2024, pp. 12040–12045, ELRA and ICCL.
- [8] Kiel D. Gonzales, Jazzmin R. Maranan, Francis Paolo D. Santelices, Edsel Jedd M. Renovalles, Nissan D. Macale, Nicole Anne A. Palafox, and Jose Marie A. Mendoza, “BK3AT: Bangsamoro k-3 children’s speech corpus for developing assessment tools in the bangsamoro languages,” in *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING*, Torino, Italia, 2024, pp. 59–65, ELRA and ICCL.
- [9] Jiaming Zhou, Shiyao Wang, Shiwan Zhao, Jiabei He, Haoqin Sun, Hui Wang, Cheng Liu, Aobo Kong, Yujie Guo, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin, “ChildMandarin: A comprehensive mandarin speech dataset for young children aged 3–5,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025, pp. 12524–12537, Association for Computational Linguistics.
- [10] Jiaming Zhou, Yujie Guo, Shiwan Zhao, Yao Lu, Jianye Wang, Haoqin Sun, Hui Wang, and Yong Qin, “ChildTalk: A multi-dialect chinese child speech corpus with full-length child-caregiver conversations for speech recognition,” in *Findings of the Association for Computational Linguistics: ACL 2026*. 2026, pp. 5103–5116, Association for Computational Linguistics.
- [11] Roberto Gretter, Marco Matassoni, Stefano Bannò, and Daniele Falavigna, “TLT-school: a corpus of non native children speech,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, 2020, pp. 378–385, European Language Resources Association.
- [12] Einar Meister and Lya Meister, “Developmental changes in acoustic characteristics of speech of estonian adolescents,” *Folklore: Electronic Journal of Folklore*, vol. 90, pp. 179–206, 2023.
- [13] Christoph Draxler, “Exploring the unknown—collecting 1000 speakers over the internet for the ph@ttsessionz database of adolescent speakers,” in *Proceedings of Interspeech*, 2006.
- [14] Mouhand Alkadri, Dania Desouki, and Khlood Al Jallad, “Arabic little STT: Arabic children speech recognition dataset,” *arXiv preprint arXiv:2510.23319*, 2025.
- [15] Ahmed Ali, Shammur Chowdhury, Mohamed Afify, Wassim El-Hajj, Hazem Hajj, Mourad Abbas, Amir Hussein, Nada Ghneim, Mohammad Abushariah, and Assal Alqudah, “Connecting arabs: bridging the gap in dialectal speech recognition,” *Commun, ACM*, vol. 64, no. 4, pp. 124–129, 2021.
- [16] Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer, “The MADAR arabic dialect corpus and lexicon,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018, European Language Resources Association (ELRA).
- [17] Hunzalah Hassan Bhatti, Firoj Alam, and Shammur Absar Chowdhury, “Multi-Task Instruction Tuning via Data Scheduling for Low-Resource Arabic SpeechLLMs,” in *Proceedings of the 5th Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing*, Hengqin, China, 2026, Accepted for publication.
- [18] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.