Arab Voices: Mapping Standard and Dialectal Arabic Speech Technology

Peter Sullivan[ξ] **AbdelRahim Elmadany**[ξ] **Alcides Alcoba Inciarte**[ξ] **Muhammad Abdul-Mageed**[ξ,λ]
[ξ]The University of British Columbia [λ]Canada Research Chair in NLP and ML
{prsull@student.,a.elmadany@,muhammad.mageed@,alcobaaj@mail.}ubc.ca

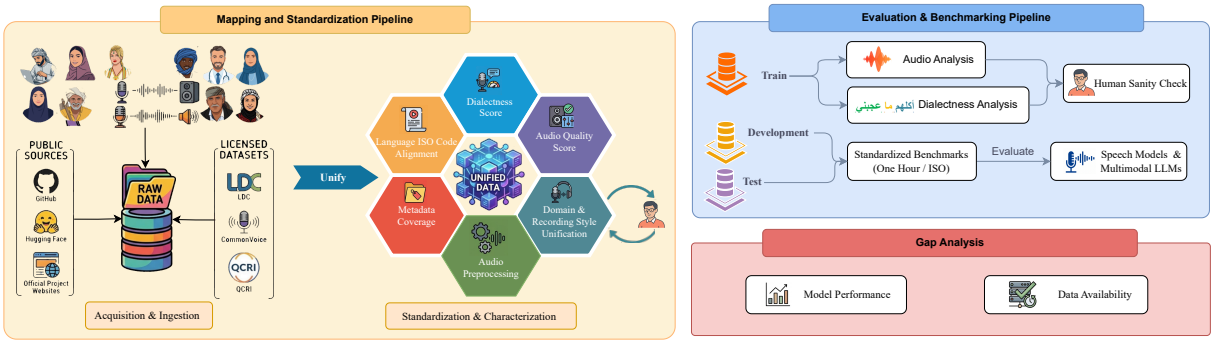


Figure 1: Mapping framework for standard and dialectal Arabic speech technology.We unify public and licensed datasets through a standardization pipeline, and then profile their demographic and quality characteristics, validating results through human-in-loop review. The unified data are then used to benchmark speech models and multimodal LLMs; the results of which alongside our dataset profiling, helps informs future data collection and modeling efforts.

Abstract

Dialectal Arabic (DA) speech data vary widely in domain coverage, dialect labeling practices, and recording conditions, complicating cross-dataset comparison and model evaluation. To characterize this landscape, we conduct a computational analysis of linguistic "dialectness" alongside objective proxies of audio quality on the training splits of widely used DA corpora. We find substantial heterogeneity both in acoustic conditions and in the strength and consistency of dialectal signals across datasets, underscoring the need for standardized characterization beyond coarse labels. To reduce fragmentation and support reproducible evaluation, we introduce Arab Voices, a standardized framework for DA ASR. Arab Voices provides unified access to 31 datasets spanning 14 dialects, with harmonized metadata and evaluation utilities. We further benchmark a range of recent ASR systems, establishing strong baselines for modern DA ASR.

1 Introduction

The landscape of Dialectal Arabic (DA) speech processing remains fragmented. While Modern Standard Arabic (MSA) has relatively robust computational support, the diverse spoken varieties of Arabic, the primary medium of daily communication for ∼450 million native of speakers, remain comparatively underserved. Recent surveys of DA speech datasets (Alqadasi et al., 2025) highlight recurring bottlenecks, including limited open-access standardization, uneven coverage of underrepresented varieties, inconsistent sub-dialect labeling, and insufficient documentation of data quality and collection conditions. A central challenge is the lack of standardized metadata that would allow practitioners to align training data with target applications. dialect labels, for instance, may be defined by geopolitical boundaries (Talafha et al., 2024; Ali et al., 2019), coarse regional groupings (Ali et al., 2017), or ISO-style codes (e.g., *apc*, *ary*) (Appen Pty Ltd, 2006a,b), which hinders principled pooling and selection of data for low-resource varieties. At the same time, pooling is not universally beneficial: variation in mutual intelligibility across the Arabic continuum can make indiscriminate mixing less effective than targeted cross-dialect transfer (Talafha et al., 2024). Richer, measurable characterization of candidate datasets can therefore help

Findings of the Association for Computational Linguistics: ACL 2026, pages 11843–11878

guide dataset selection and model design.

While existing surveys make these gaps visible, addressing them in practice requires a unified and reproducible framework. To this end, we introduce Arab Voices, a standardized ecosystem for assembling heterogeneous DA corpora into a consistent format, with harmonized metadata that supports comparison across datasets. In addition to reconciling labeling inconsistencies, our framework supports dialect organization beyond country-level tags, enabling more granular, linguistically informed representations where available. We also revisit what "quality" should mean for DA speech resources. In this setting, acoustic fidelity and linguistic authenticity do not always coincide: studio recordings can offer clean audio but may underrepresent spontaneous phenomena common in everyday speech, such as code-switching, disfluencies, and prosodic variability. Moreover, documentation of sociolinguistic variables (e.g., urban vs. rural speech, speaker demographics) is often limited or absent. We therefore complement metadata standardization with automated dataset characterization, quantifying both acoustic conditions and linguistic "dialectness" to support principled dataset selection and more robust downstream modeling.

In this paper, we build a foundation for systematically analyzing and benchmarking DA speech technology. Our contributions are: (I) **Standardized mapping.** We provide a uniform framework for aggregating and mapping heterogeneous DA datasets into a common format, including harmonized metadata and resolved dialect-label inconsistencies. (II) **Automated enrichment.** We use automated methods to characterize each dataset along two axes, i.e., linguistic "dialectness" and objective audio-quality proxies, and analyze variability across training splits. (III) **Multi-dialect benchmark.** We introduce an ASR benchmark spanning 31 datasets and 14 dialects, including varieties previously identified as underrepresented. (IV) **Broad evaluation.** We evaluate a diverse set of recent open-weight ASR systems, ranging from speech-centric architectures to audio-capable multimodal foundation models, establishing baselines for modern DA ASR.

2 A Framework for Real-World Variability in DA Speech Data

DA speech technologies are often trained and evaluated on datasets that differ substantially in linguistic coverage and recording conditions, yet these differences are not always documented in a way that supports reproducible comparison. In an ideal setting, training data would reflect the variability of real-world speech along both linguistic and acoustic axes. We therefore adapt the dataset documentation perspective of Bender and Friedman (2018) to DA speech corpora, focusing on dimensions that directly shape observable variability in audio and transcripts, and adding concrete descriptors of recording quality.[1] Therefore, we use this framework to guide what we standardize and what we quantify across DA datasets.

Linguistic variety. A primary source of variability is *regional language variation*: communities may differ in lexical choice, phonology, prosody, and morphosyntax. For practical data organization, we distinguish **dialectal variation** from **regional accent**, where the latter is more plausibly characterized by primarily phonetic differences within a shared dialectal lexicon/grammar (Wardhaugh and Fuller, 2021, p. 43). This distinction matters for dataset aggregation, since labels derived from geography or ISO codes can conflate lexical/grammatical differences with pronunciation differences, affecting both pooling decisions and the interpretation of model errors. Therefore, we harmonize dialect metadata and explicitly track dialect/region labels in a consistent schema.

Speech situation and speaker factors. Beyond region, DA corpora vary with *speech style* and social situation, which shape register and expressive content Eckert, 2016; Irvine, 2001, p. 69,74–77. Concretely, we highlight **register** (language conditioned by activity or setting) Wardhaugh and Fuller, 2021, p. 48; **affect** (emotion expressed through speech) Eckert, 2016, p. 80; and speaker-related factors such as age, gender, education, and socioeconomic background, which can correlate with systematic linguistic differences Eckert, 2016; Wardhaugh and Fuller, 2021; Lee, 2025, p. 75–76,80. In many released DA datasets, however, these sociolinguistic attributes are either missing or inconsistently reported, limiting our ability to stratify evaluation by speaker context. Therefore, we treat these factors as desiderata in our documentation

[1] Broader documentation dimensions such as curation rationale, provenance, and annotator demographics are important, but we treat them as complementary to the variability-focused axes studied here and leave their systematic treatment to future work.

schema and focus our automated characterization on measurable proxies available at scale.

Recording and channel conditions. A third major axis is the recording process itself. Device and **channel** characteristics can affect model behavior, and environmental conditions introduce reverberation and background noise that vary with room acoustics and microphone placement (Khokhlov et al., 2024; Ryu et al., 2025). Dataset-level technical choices such as **sampling rate** and **file format** can also influence some speech tasks and evaluation comparability (Ferro Filho et al., 2025). Because these factors often go under-reported, we complement metadata with objective *audio-quality proxies* computed directly from the training audio to enable consistent, cross-corpus comparison.

While this framework does not exhaustively capture all possible sources of variability, it provides a task-oriented lens for documenting and comparing DA speech resources and for identifying systematic gaps. Guided by it, we (i) standardize heterogeneous DA datasets into a unified format with harmonized dialect metadata and (ii) automatically characterize each dataset along two measurable axes (i.e., linguistic "dialectness" and acoustic conditions) to support principled dataset selection and robust benchmarking. Therefore, we next turn to mapping and analyzing the current landscape of spoken Arabic data.

3 Literature Review

'Real-World' Performance Speech recognition models often exhibit a gap between performance on curated benchmarks and performance under real-world, out-of-domain conditions (Likhomanenko et al., 2020; Radford et al., 2023; Sullivan et al., 2023). Although true deployment conditions are difficult to anticipate, cross-corpus evaluation over diverse datasets can approximate real-world variability and provide a more informative stress test of robustness (Likhomanenko et al., 2020). Common approaches for narrowing this gap include large-scale supervised pretraining (Radford et al., 2023), training objectives designed to reduce reliance on dataset-specific text style (Pratap et al., 2024), and augmentation with noise and reverberation (Likhomanenko et al., 2020). In Arabic ASR, however, evaluation is still frequently reported on a single (often in-domain) benchmark (Ali et al., 2016, 2017, 2019; Talafha et al., 2024), and comparatively fewer studies report systematic out-of-domain results, which have highlighted substantial remaining challenges for dialectal settings (Talafha et al., 2024). Therefore, our work emphasizes cross-dataset benchmarking to better reflect the variability encountered in Dialectal Arabic.

Spoken Dialectal Arabic Several systematic reviews have surveyed DA datasets (Alqadasi et al., 2025), MSA datasets (Alqadasi et al., 2023), and Arabic speech technologies including dialect identification (Elnagar et al., 2021), ASR (Alsayadi et al., 2022), and TTS (Chemnad and Othman, 2023). Collectively, these surveys summarize trends in dataset creation and intended use (Alqadasi et al., 2023, 2025), reported demographic categories (Alqadasi et al., 2023, 2025), recording environments (Alqadasi et al., 2023, 2025), and the coverage of Arabic varieties (Alqadasi et al., 2025; Chemnad and Othman, 2023; Elnagar et al., 2021; Alsayadi et al., 2022). However, such reviews typically rely on what is reported in papers and do not directly audit dataset properties; for example, while Alqadasi et al. (2023, 2025) tabulate which demographic attributes are mentioned, they do not quantify the distribution of utterance counts or durations across demographic groups.

Taxonomies of Arabic dialects are also inconsistent across prior work. Some studies adopt coarse regional groupings (e.g., Gulf, North African) (Ali et al., 2017; Abdullah et al., 2025), others use country-level labels (Ali et al., 2019; Talafha et al., 2024; Elnagar et al., 2021; Alsayadi et al., 2022; Alqadasi et al., 2025), and still others propose more fine-grained schemas (Alharbi et al., 2024; Omnilingual et al., 2025). We provide harmonized labeling and complementary, automated characterization to support more consistent comparison across DA datasets. We cover audio data quality and Speech Processing Methods in Appendix §B.

4 Standardization and Mapping

This section describes our end-to-end pipeline for (i) curating DA speech datasets, (ii) standardizing audio and text into a common representation, (iii) harmonizing dialect and domain metadata, and (iv) constructing a benchmark with a standardized training/adaptation protocol.

4.1 Standardization

Dataset curation. We identify candidate datasets through two complementary routes: (1) openly available repositories (e.g., OpenSLR and the Hug-

	Dataset	Shorthand	Duration(h)	Variety	Country	Rec. Style
ASR	ArVoice (Toyin et al., 2025)	ARVOI	6 8.3	arb	N/A	Read (*c*)
	ArzEn (Hamed et al., 2020)	ARZEN	11.4	arz	Egypt	Conv.(*d*)
	CALLHOME (Canavan et al., 1997)	CALLH	17	arz	Egypt	Conv.
	Cassablanca (Talafha et al., 2024)	CASSA	13.8	afb, apc, arq, ary, arz, mey	Various	YouTube
	Common Voice 22 (*g*) (Ardila et al., 2019)	CV22	58.1	arb	Global	Read
	Egyptian Conv (MagicData, no date(a))	EGCON	2.3	arz, acq	Egypt	Conv.
	FLEURS (Conneau et al., 2023)	FLEUR	8.2	arb	Egypt	Read
	Iraqi Telephone (Appen Pty Ltd, 2006b)	IRQTL	40.2	acm, ayp	Iraq	Conv.
	IWSLT (Abdulmumin et al., 2025)	IWSLT	2	aeb, apc	Various	Various
	GALE(*a*) (Walker et al., 2013)	GALE	972.9	arb	Various	Broadcast
	Gulf Telephone (Appen Pty Ltd, 2006a)	GULTL	97.9	afb	Various	Conv.
	L2 KSU (Alrashoudi et al., 2024)	L2KSU	7.7	arb	KSA	Read
	LINTO (Naouara et al., 2025)	LINTO	1.4	aeb	Tunisia	Various
	MASC (Al-Fetyani et al., 2023)	MASC	518.2	arb	Various	YouTube
	MediaSpeech (Kolobov et al., 2021)	MS	10	arb	Various	Broadcast
	MGB2 (Ali et al., 2016)	MGB2	1,147.5	arb	Various	Broadcast
	MGB3 (Ali et al., 2017)	MGB3	12.2	arz	Egypt	YouTube
	MGB5 (Ali et al., 2019)	MGB5	54	ary	Morocco	YouTube
	QASR (Mubarak et al., 2021)	QASR	1,914.5	arb	Various	YouTube
	Quran Speech (OpenSLR)	QURAN	795.6	arb(*b*)	Various	Recitation
	SADA2022 (Alharbi et al., 2024)	SADA	437.6	acw, ars	KSA	Broadcast
	SCC22 (SDAIA, 2022)	SCC22	4.5	UNK(*e*)	KSA	Podcast (*d*)
	TARIC-SLU (Mdhaffar et al., 2024)	TARIC	0.9	aeb	Tunisia	Various
	Tunisian MSA (OpenSLR, 2003)	TUMSA	8.7	arb	Tunisia	Read
	Yemeni Conv (MagicData, no date(b))	YEMTL	4.8	ayn, acq	Yemen	Conv.
	ZAEBUC (Habash and Palfreyman, 2022)	ZAEBC	0.4	arb, arz, afb	UAE	Discussion (*d*)
TTS	Arabic-Diacritized-TTS (Mahmoud, 2025)	ADTTS	39.2	arb	N/A	Read (*c*)
	ClArTTS (Kulkarni et al., 2023)	CATTS	11.2	arb(*b*)	Various	Read
	Iraqi TTS (Kharrufa et al., 2024)	IRTTS	5.0	acm	Iraq	Read
EMO	KSUEmotion (Alrashoudi et al., 2024)	KSUE	5.2	arb	KSA	Read
	KEDAS (Belhadj et al., 2023)	KEDAS	2.1	arb	Algeria	Read

Table 1: Transcribed Audio datasets for Arabic. We provide total hours of the filtered datasets (removing very short utterances and non-Arabic), as well as the primary varieties of Arabic spoken. For recording styles, we identify the the major kinds of recording when possible, and highlight the ambiguity of this terms with regards to recordings sourced from YouTube. We abbreviate conversational as 'Conv.' We note recording style as 'Various' when it may consist of distinct sources. (*a*) We include Parts 2, 3, and 4 of GALE. (*b*) Classical Arabic, (*c*) Partially synthetic, (*d*) Codeswitching-focused, (*e*) Underspecified dialect, (*g*) reaching out to CommonVoice directly to obtain.

ging Face Datasets hub), and (2) scholarly search for papers that introduce or document speech corpora for specific Arabic varieties, followed by retrieval of their corresponding releases when available. We fix a cutoff date of September 2025. Table 1 summarizes the resulting collection, and §C provides dataset-specific notes and provenance.

Audio and text preprocessing. To enable consistent downstream processing and evaluation, we convert all audio to a standardized representation: mono, 16 kHz sampling rate, and 16-bit PCM encoding. For corpora released as two-channel conversational audio with per-channel speakers (e.g., Appen Pty Ltd, 2006a), we split channels and align each channel to its corresponding speaker transcript when such alignment is provided. For other multichannel recordings without speaker-channel semantics, we downmix to mono by averaging channels. We retain the original transcript as a raw field and create a standardized text field for analysis and scoring. We convert Buckwalter transliteration to Arabic script when present, and we preserve diacritics and author-provided annotations in the raw field. For the standardized field, we remove segments that are entirely Latin-script (e.g., metadata-only lines) and retain mixed-script utterances. We apply light normalization intended to improve comparability across corpora while minimizing semantic changes (details in §5 and §C). For compatibility, we also provide the original 'raw' text field.

Each utterance is represented with a common **schema** that includes (when available): a unique utterance ID, audio path, duration, raw transcript, standardized transcript, dataset/source identifiers, speaker ID, recording metadata, domain labels, and dialect labels. This standardized representation is used both for dataset characterization and for benchmark construction.

4.2 Mapping and Metadata Harmonization

Dialect alignment. To harmonize dialect labels across heterogeneous annotation practices, we align dialects as best as possible using a combination of ISO 639-3 language codes plus country-region codes. This combination provides a precise

and compact description of the type of language being spoken (e.g. afb_ARE-AZ: Gulf/Khaleeji Arabic as spoken by someone in Abu Dhabi). For many datasets, the ISO 639-3 code can be inferred from country and city information, however, this provides problems for dialects from countries with multiple major dialects (e.g. Saudi Arabia), which often remain ambiguous without speaker information. In cases where the location within a country is known we use Ethnologue's language maps (Eberhard et al., 2025) to better align to ISO language description.

Domain normalization. Domain labels are inconsistently named across datasets (e.g., overlapping or near-synonymous descriptions). Across the 61 distinct domain strings observed in metadata, we manually normalize synonymous and closely related labels (e.g., *places to go* and *travel*) into 11 broad themes used for reporting and stratification. We retain both the original domain string and the normalized theme to support reproducibility.

Label sanity check. To assess gross inconsistencies in mapped labels, we conduct an initial manual review of a stratified sample of utterances ($n = 83$), focusing on dialect and domain tags where ambiguity or label drift is most likely. Findings from this check inform minor corrections to the mapping rules and highlight datasets requiring conservative treatment in the benchmark (e.g., exclusion due to unresolved ambiguity).

4.3 Automated Characterization and Analysis

Guided by the variability dimensions in Table A.1, we first scan each dataset's *training split* for available metadata.[2] We then quantify two measurable axes at scale: (i) linguistic "dialectness" and (ii) acoustic/recording conditions, which together provide a more informative characterization than coarse labels alone. We report aggregate statistics per dataset and per dialect grouping, and we analyze how these signals vary across corpora.

Text analysis (dialectness). We apply two complementary approaches. First, we use an in-house binary MSA-DA classifier to estimate whether an utterance exhibits primarily MSA-like vs. dialectal lexical/orthographic patterns. Second, we apply Arabic Level of Dialectness (ALDi) (Keleg et al., 2023), which provides a graded estimate of dialectal intensity. We aggregate these utterance-level outputs to dataset-level distributions, enabling comparisons across corpora and supporting detection of potential label mismatches (e.g., corpora labeled as DA but exhibiting strongly MSA-like transcripts).

Audio-quality analysis. To characterize recording conditions without matched clean references, we use TorchAudio-SQUIM (Kumar et al., 2023) to *predict* no-reference proxies of common speech quality/intelligibility measures. Specifically, we compute predicted PESQ (Rix et al., 2001), predicted STOI (Taal et al., 2011), predicted SI-SDR (Le Roux et al., 2019), and a predicted no-reference MOS (NMR-MOS) (Pranay Manocha and Anurag Kumar, 2022). We treat these as comparative proxies rather than ground-truth scores and analyze their distributions across datasets and dialect groupings. Figure D.1 summarizes key trends.

4.4 Human Sanity Check

Because automated proxies can fail in systematic ways, we conduct a lightweight human sanity check to contextualize the signals used in our analysis. We sample approximately 200 utterances (stratified across datasets and dialect labels where possible) for manual inspection by a native Arabic speaker. For dialectness, the annotator assigns ALDi bins following Keleg et al. (2023). For audio, the annotator provides a 1-5 MOS-style rating anchored using examples from the VoiceMOS 2022 challenge.[3] We use this study to verify that (i) extreme-score cases correspond to perceptible differences and (ii) the automated measures are directionally aligned with human judgments, while treating it as a sanity check rather than a full-scale validation.

4.5 Benchmarking and Split Release

To build our final benchmark we identify the dev and test splits of each dataset where available and ensure these materials retain their original split designation. For dialects lacking datasets with canonical splits, if there is enough data we sample one hour for test and dev, retaining the rest for train. Where we do not have enough for this, we evenly split data across train, dev, and test. For datasets with ambiguous (more than one possible ISO match e.g. 'Maghrebi') we exclude these from the final benchmark. Subsets corresponding to localities are

[2]This analysis is limited to 28 datasets with accessible training material; the remaining datasets are evaluation-only.

[3]https://zenodo.org/records/6572573

also provided, for the samples where this information is available.

Similarly to how we characterize the training data through automated tools in §5, we use the same sets of labeling tools to conduct a set of experiments on the dev set to better categorize ASR performance based on these factors, however, the test set remains untouched.

In total we arrive with a benchmark covering 14 of the ISO 639-3 labels, with each dialect provide with an adaptation split of five hours, one hour development, and a one hour test split. These splits pull from multiple datasets where possible (see Table F.1 for detailed breakdown of composition), in order to better reflect performance on diverse domains and conditions. Due to license and copyright restrictions, we do not provide the dataset itself, but instead provide a set of scripts and metadata for other researchers to generate the splits themselves [4]. The scripts take the downloaded data archives and generate corresponding parquet files with standardized metadata and ISO-aligned language codes.

5 Dataset Analysis

Metadata coverage. Across the major dimensions outlined in §2, we find that only a subset can be directly recovered from released metadata. We summarize available metadata for dialect coverage, age, gender, and domain in §D. Therefore, beyond metadata alone, we rely on automated characterization to support consistent comparison across datasets and to inform benchmark construction.

Text analysis of dialectness. Using ALDi, we observe patterns that are consistent with the intended collection conditions of several corpus types. Broadcast news datasets (MASC, QASR, MGB2) are predominantly MSA, and a similar trend holds for many read-speech datasets, including FLEUR, CATTS (Classical Arabic), and TUMSA. For MGB2, the original release estimated a lower bound of at least 70% MSA (Ali et al., 2016). Our ALDi distributions are broadly consistent with this estimate: 24.5% of utterances have ALDi $>$ 0.11 (corresponding to *little DA* in Keleg et al., 2023), while 5.6% exceed 0.44 (*mixed*) and 1% exceed 0.77 (*mostly DA*). In aggregate, the *mostly DA* bin corresponds to 11 hours of speech that could potentially be isolated for dialect-focused use.

[4] https://github.com/UBC-NLP/arab_voices

At the other end of the spectrum, conversational telephone datasets (GULTL, IRQTL, CALLH) are largely dialectal, although YEMTL is a notable exception. Compared to MSA-oriented corpora, these datasets exhibit a wider spread in degree of dialectness. A similar pattern holds for datasets collected explicitly for particular dialects, such as SADA, MGB3, and MGB5. We provide dataset-level violin plots in §D. These ALDi-based predictions are mirrored by our binary MSA/DA classifier (Figure D.5); computing Pearson correlation between the ALDi scores and the binary predictions (MSA$=0$, DA$=1$) yields $r = 0.91$. Therefore, we use dialectness distributions as a complementary signal when selecting and interpreting per-dialect adaptation and evaluation subsets in our standardized protocol.

Noise and intelligibility. We next analyze predicted recording conditions using our audio-quality proxies. Consistent with the recording setups typically used for TTS corpora (Table D.2), ADTTS, CATTS, and IRTTS exhibit high predicted quality under the objective proxies (all mean predicted PESQ $>$ 3 and mean predicted SI-SDR $>$ 20). Several other read-speech datasets, including ARVOI, TUNAI, and TUMSA, show similarly strong scores. More detailed anaylsis in Appendix §D

6 Experimental Setup

We evaluate the zero-shot performance of publicly available pretrained models on the development and test splits of our benchmark, reporting both overall results and per-dialect performance. We use Word Error Rate (WER) as the primary metric and Character Error Rate (CER) as a secondary metric for assessing ASR quality. To ensure consistent scoring across heterogeneous corpora, we apply a standardized text normalization pipeline to both model predictions and reference transcripts prior to computing WER/CER; details are provided in §F.1.

We evaluate the following models: Whisper-large-v3 (Radford et al., 2023), MMS (Pratap et al., 2024), SeamlessM4T v2 (Barrault et al., 2023), Omnilingual (Omnilingual et al., 2025), and Qwen3 (Yang et al., 2025). Model descriptions and configuration details are provided in §E.

To further understand how ASR performance is impacted by the gender, domain, dialectness, and audio quality, we also perform zero-shot benchmarking using the development set and analyze the

Dialect	Encode-Only (CTC / Acoustic Encoder)					Encode-Decoder			Multimodal (Speech + LLM Core)					
	E1	E2	E3	E4	E5	ED1	ED2	ED3	M1	M2	M3	M4	M5	M6
MSA (arb)	38.94	26.18	16.49	13.5	13.8	18.51	15.55	14.7	17.52	11.51	13.57	11.52	11.3	**10.19**
Algerian (arq)	96.19	88.26	81.65	78.64	**78.01**	175.34	116.92	124.2	109.76	82.56	83.6	79.25	78.92	86.29
Egyptian (arz)	74.32	58.87	45.7	42.28	45.59	48.16	35.69	37.61	43.25	68.6	41.7	49.38	48.76	**34.8**
Hassaniyya (mey)	96.57	91.79	89.19	88.55	88.45	204.97	135.94	93.46	106.15	83.52	88.24	84.3	84.58	**83.17**
Hijazi (Saudi Arabia) (acw)	82.03	72.77	59.08	56.73	58.05	191.21	105.94	62.14	91.53	55.93	67.79	73.33	56.43	**53.49**
Khaleeji (afb)*	89.62	81.54	74.96	73.54	71.72	213.89	105.14	61.89	80.23	**53.71**	121.57	84.05	91.34	68.39
Levantine (apc)*	51.75	38.91	29.72	26.84	26.88	36.31	32.53	26.38	28.28	23.81	27.61	25.45	25.17	**23.53**
Mesopotamian Arabic (Iraq) (acm)	94.02	86.43	80.36	81.2	80.13	271.25	119.91	106.26	87.53	**71.16**	108.36	86.76	81.93	73.86
Moroccan (ary)	113.83	84.48	79.78	**79.71**	81.77	164.21	120.43	86.2	122.79	99.33	94.68	98.85	93.19	103.8
Najdi (Saudi Arabia) (ars)	84.01	76.93	64.06	59.31	60.17	173.42	113.92	62.63	72.73	**46.9**	80.75	63.92	61.31	60.23
North Mesopotamian Arabic (Iraq) (ayp)	94.65	89.14	87.58	87.99	84.11	367.74	134.53	87.78	99.13	**75.45**	88.48	93.12	104.72	95.16
Sanaani Arabic (Yemen) (ayn)	74.99	58.25	50.36	50.49	47.81	188.3	59.06	44.25	47.67	**34.63**	52.32	53.7	46.19	40.37
Sudanese (apd)	74.82	63.97	49.1	45.71	44.12	70.9	49.19	47.26	48.71	40.06	45.49	37.68	41.33	**36.01**
Ta'izzi-Adeni Arabic (Yemen) (acq)	72.61	55.22	46.42	45.84	45.44	125.22	77.06	39.54	50.62	**28.77**	44.02	43.43	38.03	35.5
Tunisian (aeb)	92.46	86.38	78.37	80.51	79.33	225.04	111.93	80.63	85.32	247.39	80.62	84.17	77.55	**74.31**

Table 2: WER performance across varieties for 14 models spanning three distinct ASR architectural paradigms on TEST dataset. **Encode-Only:** **E1.** MMS-1B-ALL, **E2.** omniASR-CTC-300M-v2, **E3.** omniASR-CTC-1B-v2, **E4.** omniASR-CTC-3B-v2, and **E5.** omniASR-CTC-7B-v2. **Encode-Decoder:** **ED1.** Whisper-Large-v3-Turbo, **ED2.** Whisper-Large-v3, and **ED3.** SeamlessM4T-v2-Large. **Multimodal:** **M1.** Voxtral-Small-24B-2507, **M2.** Qwen3-Omni-30B-A3B-Instruct, **M3.** omniASR-LLM-300M-v2, **M4.** omniASR-LLM-1B-v2, **M5.** omniASR-LLM-3B-v2, and **M6.** omniASR-LLM-7B-v2. **Bold** refers to the best performance for each dialect. *Khaleeji (afb) and Levantine (apc) include varieties from multiple countries. *For CER performance, details are provided in Table G.2 and for subdialect breakdown see Table G.1.*

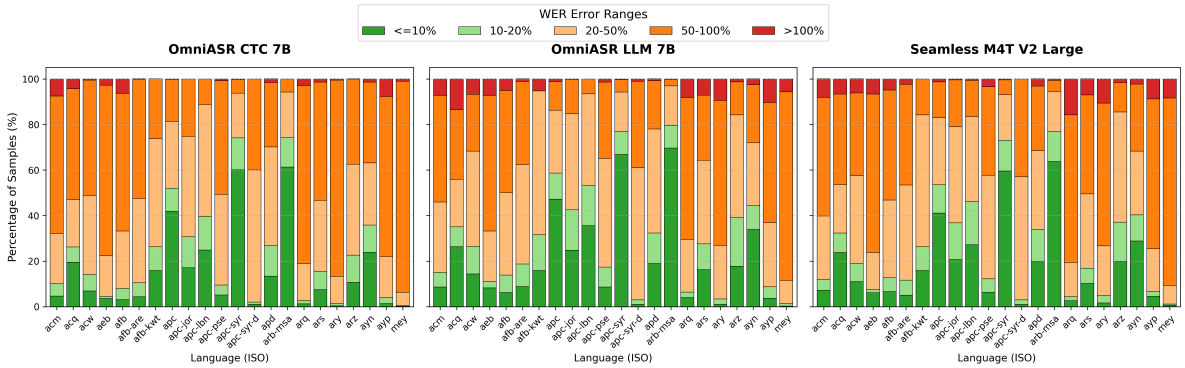


Figure 2: WER distribution across languages for the best-performing ASR models from the three architectures listed in Table 2. Stacked bar charts show the proportion of samples in different WER ranges (from $\leq$10% to >100%) across multiple languages. Colors range from green (low error rates) to red (high error rates). "*OmniASR LLM 7B*" consistently achieves a higher share of low-WER samples and more stable performance across languages, highlighting the advantages of LLM-based architectures over CTC-based approaches for multilingual ASR. *For the full analysis on WER and CER performance, details are provided in Figure G.1 and Figure G.2 Appendix G.*

results across the corresponding categories.

7 Results and Discussion

Dataset analysis. Our metadata audit highlights a substantial gap between the variability dimensions in our idealized framework (§2) and what is routinely recorded in released DA corpora. Utterance-level sociolinguistic descriptors (e.g., register or affect) are uncommon, with emotion recognition datasets being a notable exception where such labels are often integral to the task. We also observe inconsistency in how location metadata is used: in some cases it reflects speaker origin (and thus may index regional accent), while in others it reflects only the recording site, which can confound dialect mapping and downstream interpretation. Demographic fields are similarly uneven: gender is the most frequently recorded attribute, and age is sometimes available, whereas socioeconomic, education, or rural-urban indicators are rarely documented despite their potential value for understanding representativeness. Therefore, our benchmark and standardized protocol rely on harmonized dialect labels where possible and supplement missing metadata with automated dataset characterization.

Automated profiling provides more reliable signals for transcript-based characterization than for recording-condition assessment. While our predicted objective audio-quality proxies are broadly consistent with one another (Figure D.1), they do not consistently track human judgments in our manual inspection. Likewise, the conservative behavior of the non-matching reference MOS predictor may reflect domain mismatch on out-of-domain samples. In qualitative review, we encountered cases where audio that sounded clean received low predicted scores and vice versa, and the relationship between these proxies and downstream model training remains unclear. Therefore, we treat audio-quality estimates as comparative context signals rather than as definitive measures, and we retain dataset provenance in our benchmark/adaptation protocol to support analysis conditioned on recording conditions.

Human verification. To contextualize the automated profiling signals, we conduct a human-in-the-loop check comparing model predictions against native-speaker annotations (Appendix §G.1). We find that coarse binary distinctions (MSA vs. Dialect) and ALDi are both quite accurate in analyzing transcriptions. We further observe that standard quality metrics (e.g., PESQ) can penalize expressive speech (e.g., emotional or religious recordings) that human listeners rate highly, reinforcing the need to interpret predicted quality proxies with caution. Therefore, our analysis layer emphasizes robust, coarse-grained dialectness signals for dataset characterization, and we use audio-quality proxies primarily for contextualizing benchmark outcomes rather than for filtering or ranking data.

Benchmark results. Across dialects, zero-shot performance of off-the-shelf models remains challenging (Table 2). Nevertheless, we observe several stronger configurations on specific varieties: Omnilingual LLM 7B achieves WERs of 34.8 (arz), 23.53 (apc), and 36.0 (apd), while Qwen3 attains WERs of 34.63 (ayn) and 28.77 (acq). For MSA, we observe that three of four Omnilingual LLM variants and Qwen3 achieve WER below 12.0. Therefore, our standardized evaluation protocol provides a consistent basis for identifying which model families transfer more effectively in a zero-shot setting and where per-dialect adaptation is most warranted.

We further note strong performance from the Omnilingual CTC models on the low-resource Algerian (arq) and Moroccan (ary) subsets (see also CER results in Table G.2). We report results without language-model decoding; incorporating an LM could plausibly improve CTC performance, but a practical barrier is obtaining suitably in-domain text for the target variety. Therefore, our benchmark design, with fixed per-dialect adaptation, dev, and test splits, is intended to support controlled exploration of such adaptation strategies (including LM-based approaches) under comparable supervision budgets.

Finally, we observe a characteristic failure mode for encoder-decoder models: in some cases, Whisper produces WER above 200, corresponding to excessively long, low-quality outputs that appear consistent with failures to terminate generation appropriately. Therefore, our benchmark/adaptation protocol, combined with the accompanying analysis layer, facilitates systematic identification of such failure modes across dialects and recording conditions.

Categorized ASR performance. In addition to our benchmark results, we also consider how our ASR performance varies across domain, dialectness, gender, and audio quality. We find a minor imbalance of performance across gender, ranging from an absolute WER difference of 0.07 (Omnilingual CTC 7B) to 0.03 (Omnilingual LLM 7B as well as Qwen3 Omni 30B) both in favor of male speakers.

For domain Work/Professional and Cultural/Arts both provide the most difficulty for ASR systems, with no system performing under 100% for the latter and only Qwen 3 (80%) and Seamless (90%) breaking that barrier for the former. This indicates a clear need for audio recordings of more spontaneous meetings and discussions as well as more arts focused recordings covering a broader range of genre (for instance theater and poetry).

Dialectness measures appeared to provide an avenue for mining harder utterances, with WER generally increasing as ALDi score increases. The large gap between MSA and LOW dialectness, however, may indicate the benefit of simply using ALDi to filter out MSA utterances for this purpose.

Audio quality metrics similarly reflect conventional wisdom that poorer quality audio is harder for ASR systems. We find that the performance across PESQ, STOI, SI-SDR follows this trend, meaning that they are all potentially beneficial tools

Model	PS	NE	EM	HW	LH	ES	TC	WP	Gen	TT	CA	Unk
omniASR CTC 7B v2	51	48	43	43	58	29	62	101	38	72	96	68
omniASR LLM 7B v2	44	37	43	38	68	29	68	124	46	75	102	62
mms-1b-all	76	73	72	70	86	56	83	103	100	84	106	87
Qwen3-Omni	40	31	79	32	51	26	50	80	38	52	104	60
seamless-m4t-v2-large	45	43	43	41	81	37	54	90	38	68	293	73
whisper-large-v3	87	45	85	93	97	41	186	183	38	152	109	104

Table 3: WER on Arab Voices DEV broken down by domain. Domain key: People & Society (PS), Nature & Environment (NE), Entertainment & Media (EM), Health & Wellness (HW), Lifestyle & Home (LH), Education & Science (ES), Technology & Communication (TC), Work & Professional (WP), General / Other (Gen), Travel & Transport (TT), Culture & Arts (CA), Unknown (Unk). Please see Table D.1 for examples of the raw fields that were mapped to each of these categories.

for hard-mining utterances. However, we find that this trend of utterances with poor estimated quality having higher WER does not apply to the subjective non-matching reference approach, potentially limiting the usefulness of this approach.

Existing gaps. A major methodological gap is the lack of an *audio-based* analogue of ALDi (Keleg et al., 2023). While text-based dialectness models help characterize corpora with transcripts, their reliance on text prevents direct application to untranscribed speech (e.g., spoken dialect identification settings). In terms of resource coverage, we identify several dialects that remain undersupported and challenging in our benchmark, particularly North African varieties (Algerian, Hassaniyya, Moroccan, Tunisian) and Iraqi varieties (Mesopotamian and North Mesopotamian), and we are unable to identify datasets for several other dialects within our cutoff and inclusion criteria (Table C.1). We also reiterate the need for dataset creators to document collection conditions more systematically and to expand coverage toward underrepresented domains (e.g., health, arts, and nature). Therefore, Arab Voicesprovides both a reproducible evaluation protocol and a practical scaffold for prioritizing future data collection and dialect adaptation efforts where gaps are most acute.

8 Conclusion

We present a detailed analysis of a broad selection of spoken Arabic datasets through the complementary lenses of linguistic *dialectness* and recording conditions. As an infrastructure-focused approach, we standardize and map heterogeneous resources into a unified schema, and we introduce a benchmark with a standardized evaluation and training/adaptation protocol for assessing modern ASR performance on Dialectal Arabic. It is our hope that the standardization framework aids resource-constrained groups in approach the diverse landscape of Dialectal Arabic datasets. Our analysis highlights substantial variation across corpora and underscores persistent gaps in both documentation and coverage. We identify several directions for future work: (1) evaluating how predicted audio-quality measures relate to ASR performance in practice; (2) expanding benchmark data to cover additional low-resource dialects; (3) developing language models suitable for CTC decoding in Dialectal Arabic; and (4) extending benchmarks to a broader range of domains.

Limitations

Our dialect mapping uses paired ISO 639-3 labels and country-region tags, which can support more fine-grained organization than country-only labeling. However, this representation alone may not capture distinctions that cut across geography, such as rural-urban variation, and further refinement would require additional metadata that is often unavailable in released corpora.

In addition, ISO labels are an imperfect administrative proxy for a dialect continuum: they may not fully reflect sociolinguistic realities and can be revised over time, which may complicate long-term reproducibility of mappings. Our analysis and benchmark should therefore be interpreted as a best-effort snapshot under the dataset releases and labeling conventions available within our collection window.

While we present an account of ASR performance across dialects, we recognize the importance of other tasks in Arabic speech processing (TTS, Emotion classification), and note that limited size of datasets supporting these tasks. Due to the na-

ture of using existing datasets for our benchmark, we are unable to leverage or expand on the multi-reference evaluation approach used in MGB5 (Ali et al., 2019), but acknowledge the limitations in evaluating the flexibile orthography of DA when using only one reference. While we provide average WER performance across a variety of dialects, a direct error analysis and comparison between types of errors made by different models would supplement these findings, we leave this exploration to future work.

Finally, our dialectness analysis relies on text-based models such as ALDi, which can miss pronunciation differences that are not reflected orthographically (e.g., cases where written DA and MSA are indistinguishable without diacritization). Developing an audio-based analogue of ALDi ("spoken ALDi") is a promising direction to address this limitation in future work. Additionally, in checking the consistency of our ALDi labeling tool, we acknowledge the limited number of samples which were examined, however, we view this as 'sanity check,' and supplementary to the validation work done in the release of ALDi.

Ethical Considerations

Our framework emphasizes sociolinguistic and contextual factors that can matter for real-world speech technology, but richer documentation can also increase privacy risks for recording participants. In particular, inferring sensitive attributes from audio or transcripts (e.g., gender, socioeconomic status, or education) can be intrusive and may enable unwanted profiling. Accordingly, our work focuses on dialect and recording-condition characterization and reports results in aggregate, and we do not attempt to infer sensitive demographic attributes beyond what is explicitly provided in released metadata. We encourage dataset creators to balance transparency with participant privacy and to document consent, intended use, and data handling practices when possible.

Statement of Generative AI use

We used Gemini to customize initial data preprocessing scripts for each of the different datasets. These were later edited and manually checked for correct preprocessing output.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[5] and UBC ARC-Sockeye.[6]

References

Muhammad Abdul-Mageed, Abdelrahim Elmadany, and 1 others. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105.

Badr M. Abdullah, Matthew Baas, Bernd Möbius, and Dietrich Klakow. 2025. Voice Conversion Improves Cross-Domain Robustness for Spoken Arabic Dialect Identification. In *Interspeech 2025*, pages 2790–2794.

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. Findings of the IWSLT 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Maryam Al Ali and Hanan Aldarmaki. 2024. Mixat: A data set of bilingual emirati-english speech. In *SIGUL 2024: 3rd Annual Meeting of the ELRA/ISCA Special Interest Group on Under-resourced Languages, a Satellite Workshop of LREC-COLING 2024*.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2023. Masc: Massive arabic speech corpus. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1006–1013. IEEE.

Abdulkafi Albirini. 2011. The sociolinguistic functions of codeswitching between standard arabic and dialectal arabic. *Language in society*, 40(5):537–562.

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim,

[5] https://alliancecan.ca

[6] https://arc.ubc.ca/ubc-arc-sockeye

Raghad Aloraini, Raneem Alnajim, and 1 others. 2024. Sada: Saudi audio dataset for arabic. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290. IEEE.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019. The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Ammar Mohammed Ali Alqadasi, Rawad Abdulghafor, Mohd Shahrizal Sunar, and Md Sah Bin HJ Salam. 2023. Modern standard arabic speech corpora: A systematic review. *Ieee Access*, 11:55771–55796.

Ammar Mohammed Ali Alqadasi, Akram M Zeki, Mohd Shahrizal Sunar, Siti Zaiton Mohd Hashim, Md Sah hj Salam, and Rawad Abdulghafor. 2025. Arabic dialects speech corpora: A systematic review. *Speech Communication*, page 103322.

Norah Alrashoudi, Hend AlKhalifa, and Yousef Ajami Alotaibi. 2024. L2-ksu native and non-native arabic speech. https://doi.org/10.35111/n3d8-t960. LDC2024S11.

Hamzah A Alsayadi, Abdelaziz A Abdelhamid, Islam Hegazy, Bandar Alotaibi, and Zaki T Fayed. 2022. Deep investigation of the recent advances in dialectal arabic speech recognition. *IEEE access*, 10:57063–57079.

Eiman Alsharhan and Allan Ramsay. 2020. Investigating the effects of gender, dialect, and training size on the performance of arabic speech recognition. *Language Resources and Evaluation*, 54(4):975–998.

Appen Pty Ltd. 2006a. Gulf arabic conversational telephone speech. https://doi.org/10.35111/nsvg-dd69. LDC2006S43.

Appen Pty Ltd. 2006b. Iraqi arabic conversational telephone speech. https://doi.org/10.35111/2dcs-9751. LDC2006S45.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Mourad Belhadj, Ilham Bendellali, and Elalia Lakhdari. 2023. Kasdi-merbah university emotional database in arabic speech. https://doi.org/10.35111/qqer-qz15. LDC2023S10.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Alexandra Canavan, George Zipperlen, and David Graff. 1997. Callhome egyptian arabic speech. https://doi.org/10.35111/d8yb-9m13. LDC97S45.

Khansa Chemnad and Achraf Othman. 2023. Advancements in arabic text-to-speech systems: a 22-year literature review. *IEEE Access*, 11:30929–30954.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. Ethnologue: Languages of the world.

Penelope Eckert. 2016. *Variation, meaning and social change*, page 68–85. Cambridge University Press.

Ashraf Elnagar, Sane M Yagi, Ali Bou Nassif, Ismail Shahin, and Said A Salloum. 2021. Systematic literature review of dialectal arabic: identification and detection. *IEEE Access*, 9:31010–31042.

Alexandre Ferro Filho, Diogo Fernandes Costa Silva, Pedro Elias Engelberg Silva Borges, and Arlindo Rodrigues Galvão Filho. 2025. Evaluating Deep Speaker Embedding Robustness to Domain, Sampling Rate, and Codec Variations. In *Interspeech 2025*, pages 1113–1117.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2021. Investigating the Impact of Gender Representation in ASR Training Data: A Case Study on Librispeech. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 86–92. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation

benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Nizar Habash and David Palfreyman. 2022. Zaebuc: An annotated arabic-english bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88.

Nawar Halabi. 2016. *Modern standard Arabic phonetics for speech synthesis*. Ph.D. thesis, University of Southampton.

Injy Hamed, Fadhl Eryani, David Palfreyman, and Nizar Habash. 2024. ZAEBUC-spoken: A multilingual multidialectal Arabic-English speech corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17770–17782, Torino, Italia. ELRA and ICCL.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. Arzen: A speech corpus for code-switched egyptian arabic-english. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4237–4246.

Judith T. Irvine. 2001. *"Style" as distinctiveness: the culture and ideology of linguistic differentiation*, chapter 1. Cambridge University Press.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. Aldi: Quantifying the arabic level of dialectness of text. *arXiv preprint arXiv:2310.13747*.

Hayder Kharrufa, Adam Taha, and Mohammed Baraq. 2024. Training a Text-to-Speech System for Dialectal Arabic with a Focus on the Iraqi Dialect.

Yuri Khokhlov, Tatiana Prisyach, Anton Mitrofanov, Dmitry Dutov, Igor Agafonov, Tatiana Timofeeva, Aleksei Romanenko, and Maxim Korenevsky. 2024. Classification of Room Impulse Responses and its application for channel verification and diarization. In *Interspeech 2024*, pages 3250–3254.

Rostislav Kolobov, Olga Okhapkina, Andrey Platunov Olga Omelchishina, Roman Bedyakin, Vyacheslav Moshkin, Dmitry Menshikov, and Nikolay Mikhaylovskiy. 2021. Mediaspeech: Multilanguage asr benchmark and dataset. *Preprint*, arXiv:2103.16193.

Ajinkya Kulkarni, Atharva Kulkarni, Sara Abedalmon'em Mohammad Shatnawi, and Hanan Aldarmaki. 2023. Clartts: An open-source classical arabic text-to-speech corpus. In *2023 INTERSPEECH*, pages 5511–5515.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. 2023. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. 2019. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE.

Nala H Lee. 2025. *Sociolinguistic variation*, chapter 2. Routledge.

Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. 2020. Rethinking evaluation in asr: Are our models robust enough? *arXiv preprint arXiv:2010.11745*.

MagicData. no date(a). Asr-egarbcsc: An egyptian arabic conversational speech corpus.

MagicData. no date(b). Asr-egarbcsc: An egyptian arabic conversational speech corpus.

Nourhan Mahmoud. 2025. Arabic-diacritized-tts.

Salima Mdhaffar, Fethi Bougares, Renato De Mori, Salah Zaiem, Mirco Ravanelli, and Yannick Estève. 2024. Taric-slu: A tunisian benchmark dataset for spoken language understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15606–15616.

Ali Hamid Meftah, Yousef Ajami Alotaibi, and Sid-Ahmed Selouani. 2017. Ksuemotions. https://doi.org/10.35111/q1eh-6457. LDC2017S12.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. Qasr: Qcri aljazeera speech resource–a large scale annotated arabic speech corpus. *arXiv preprint arXiv:2106.13000*.

Gautham J Mysore. 2014. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges. *IEEE Signal Processing Letters*, 22(8):1006–1010.

Hedi Naouara, Jean-Pierre Lorré, and Jérôme Louradour. 2025. Linto audio and textual datasets to train and evaluate automatic speech recognition in tunisian arabic dialect. In *Workshop on Preparing Good Data for Generative AI: Challenges and Approaches*.

ASR Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, and 1 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*.

OpenSLR. Mohammed - Quran Speech to Text Dataset. https://www.openslr.org/132/.

OpenSLR. 2003. Tunisian_MSA. https://www.openslr.org/46/.

Pranay Manocha and Anurag Kumar. 2022. Speech Quality Assessment through MOS using Non-Matching References. In *Interspeech 2022*, pages 654–658.

Pranay Manocha and Zeyu Jin and Adam Finkelstein. 2022. Audio Similarity is Unreliable as a Proxy for Audio Quality. In *Interspeech 2022*, pages 3553–3557.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

Marshima Mohd Rosli, Ewan Tempero, and Andrew Luxton-Reilly. 2018. Evaluating the quality of datasets in software engineering. *Advanced Science Letters*, 24(10):7232–7239.

Myeonghoon Ryu, Hongseok Oh, Suji Lee, and Han Park. 2025. Unified Microphone Conversion: Many-to-Many Device Mapping via Feature-wise Linear Modulation. In *Interspeech 2025*, pages 1333–1337.

SDAIA. 2022. Saudilang code-switch corpus (scc). https://www.kaggle.com/datasets/sdaiancai/saudilang-code-switch-corpus-scc. CC BY-NC-SA 4.0.

Hagen Soltau, Lidia Mangu, and Fadi Biadsy. 2011. From modern standard arabic to levantine asr: Leveraging gale for dialects. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 266–271. IEEE.

Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. On the robustness of arabic speech dialect identification. In *Interspeech 2023*, pages 5326–5330.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on audio, speech, and language processing*, 19(7):2125–2136.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, and 1 others. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. *arXiv preprint arXiv:2410.04527*.

Hawau Olamide Toyin, Rufael Marew, Humaid Alblooshi, Samar M. Magdy, and Hanan Aldarmaki. 2025. Arvoice: A multi-speaker dataset for arabic speech synthesis. *Preprint*, arXiv:2505.20506.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Kevin Walker, Christopher Caruso, Kazuaki Maeda, Denise DiPersio, and Stephanie Strassel. 2013. Gale phase 2 arabic broadcast conversation speech part 1. https://doi.org/10.35111/j224-tx54. LDC2013S07.

Ronald Wardhaugh and Janet M Fuller. 2021. *An introduction to sociolinguistics*. John Wiley & Sons.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Omar Zaidan and Chris Callison-Burch. 2011. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41.

Appendices

This appendix provides supplementary material to support the main findings of this work. It is organized as follows:

- Appendix A Framework of Audio Variability
- Appendix B Literature Review
- Appendix C Mapping and Standardization
- Appendix D Dataset Analysis
- Appendix E Experimental Setup
- Appendix G Results

Key Tables

- Table A.1: A framework for characterizing aspects of variabiltiy in audio recordings.
- Table C.1: Dialect-wise breakdown of transcribed Arabic audio by train/dev/test splits (hours and utterances).
- Table D.1: Mapping of raw to normalized domain labels.
- Table D.2: Audio metrics overview.
- Table F.1: Benchmark overview.
- Table G.1: A comprehensive comparison of WER performance on TEST dataset.
- Table G.2: A comprehensive comparison of CER performance on TEST dataset.
- Table G.3: A comprehensive comparison of WER performance on DEV dataset.
- Table G.4: WER on DEV broken down by gender.
- Table G.5: WER on DEV broken down by dialectness.
- Table G.6: WER on DEV broken down by audio quality (PESQ).
- Table G.7:WER on DEV broken down by audio quality (SI-SDR).
- Table G.8: WER on DEV broken down by audio quality (STOI).
- Table G.9: WER on DEV broken down by audio quality (NMR).

Key Figures

- Figure D.1: Audio Performance Landscape
- Figure D.2: Age distribution.
- Figure D.3: Gender distribution.
- Figure D.4: DOmains distribution.
- Figure G.1: WER distribution across languages for all ASR models.
- Figure G.2: CER distribution across languages for all ASR models.

A Framework

Please see Table A.1.

B Literature Review

Dataset Quality Analysis General-purpose dataset quality frameworks (e.g., Rosli et al. 2018) are not tailored to the sources of variability that are specific to speech data, particularly recording and channel effects (Table A.1). Data documentation proposals such as Bender and Friedman (2018) provide a principled vocabulary for describing language variety, speakers, and speech situations, but they are primarily prospective standards and depend on creators to report these attributes consistently. In practice, many relevant factors are incompletely documented, and audio fidelity can be difficult to estimate at scale without matched clean references. Recent progress in non-intrusive (no-reference) quality estimation offers an alternative by providing model-based proxies intended to correlate with perceived quality (Kumar et al., 2023; Pranay Manocha and Anurag Kumar, 2022). Therefore, we complement metadata with automated audio-quality measures computed directly from released training audio.

Speech Processing Methods Modern ASR systems predominantly build on transformer (Vaswani et al., 2017) and conformer (Gulati et al., 2020) backbones, spanning encoder-decoder architectures (Radford et al., 2023; Omnilingual et al., 2025), CTC-based transformer encoders (Pratap et al., 2024; Omnilingual et al., 2025), and conformer encoders paired with transformer decoders (Barrault et al., 2023). The distinction between CTC-based and decoder-based systems is particularly relevant for dialectal settings: decoder-based

Variation Category	Dimension	Metadata Representation	Example
Region	Dialect	ISO 639-3	arz
Region	Regional Accent	ISO 3166-2	EG-C or EGY-C (Cairo)
Region	'Rural-ness'	Categorical	Urban
Stylistic	Register	Description	Informal
Stylistic	Affect	Description	Joyous
Stylistic	Occasion	Description	A family gathering
Demographic	Age	Numeric	40
Demographic	Gender Expression	Categorical	M
Demographic	Ethnic Identity	Categorical	Egyptian
Demographic	Education	Categorical	Bachelors
Demographic	Socioeconomic	Categorical	Middle Class
Demographic	Language History	Description	L1 Arabic, L2 English speaker
Demographic	Speech Disfluencies	Description	Stutter present
Recording	Channel	Description	Samsung Galaxy A16 smartphone
Recording	Environment	Description	Noisy; recording in a large room
Recording	Sampling Rate	Numeric	16kHz
Recording	Encoding	Categorical	FLAC

Table A.1: An idealized framework for analyzing sources of variability in audio recordings for speech processing, alongside potential standardized ways these dimensions could be recorded in metadata. Open ended dimensions, such as style, lend themselves towards descriptive elements, while demographic elements might be better represented by speakers providing their own labels.

models learn an explicit language model component that can bias outputs toward dominant training styles, whereas CTC systems operate via token-level alignment and may exhibit different failure modes (Pratap et al., 2024). This consideration becomes salient in large-scale multilingual training, where Arabic is often treated as a single language, potentially privileging MSA-like transcriptions, as observed in widely used models such as Whisper (Radford et al., 2023) and MMS (Pratap et al., 2024). Therefore, our evaluation compares model families with different decoding biases to better understand performance on dialectal Arabic.

C Mapping and Standardization Pipeline

We begin this appendix with descriptions of all the datasets, and also provide a summary table of the total hours of these datasets broken into dialect (Table C.1).

Monolingual Transcribed Datasets

ASR-EgArbCSC (MagicData, no date(a)): 5.5 hours of conversational Egyptian speech. While demographic information is provided (age, gender, city), and information about the conversational topic is also provided, there is little other information regarding the provenance and collection strategy used for this dataset.

ASR-YeArCSC (MagicData, no date(b)): 10.42 hours of conversational Yemeni speech. While demographic information is provided (age, gender, city), and information about the conversational topic is also provided, there is little other information regarding the provenance and collection strategy used for this dataset. This is somewhat problematic, as language information may need to be inferred based on the city information.

CALLHOME (Canavan et al., 1997): This 60 hours Egpytian Arabic corpus consists of unscripted telephone calls. A selection of each call is provided with a transcript, meaning that far fewer than the full total amount of time is available for use in training ASR systems (14.3 / 3.6 / 1.7 hr.). Additional details are provided about each recording, for instance for elderly speakers, gender of speakers, and any challenging conditions, however, little additional demographic information is provided for recipients of the calls.

Casablanca (Talafha et al., 2024): This dataset consists of 48 hours dialect Arabic of which a validation and test split of 8 hours each has been released. Eight country-level dialects are included covering: Algeria, Egypt, Jordan, Morocco, Mauritania, Palestine, UAE, and Yemen. Audio was collected from YouTube, with manual transcription and validation. Gender splits vary widely between

Dialect	Primary Countries	ISO	TRAIN		DEV		TEST	
			Dur (H)	Uttr	Dur (H)	Uttr	Dur (H)	Uttr
Levantine	JOR, LBN, PSE, SYR	apc	23.2	18,581	1.0	974	9.5	7,415
Khaleeji	KWT, QAT, ARE, SAU	afb	96.2	62,235	1.1	922	3.0	2,204
MSA	-	arb	3951.6	2,217,449	36.1	26,484	36.6	27,201
Algerian	DZA	arq	0.5	600	1.0	844	1.0	921
Egyptian	EGY	arz	427.2	22,872	7.4	5,554	7.8	4,893
Hassaniya	MRT, MLI, ESH	mey	–	–	1.0	953	0.9	953
Libyan	LBY	ayl	0.0	37	–	–	–	–
Moroccan	MAR, ESH	ary	39.5	31,266	8.5	6,985	8.9	6,559
Sudanese	SUD	apd	4.3	3,933	–	–	0.1	127
Tunisian	TUN	aeb	61.7	21,139	0.0	36	2.3	2,060
Hijazi	SAU	acw	40.2	34,833	0.6	528	1.1	809
Najdi	SAU	ars	117	90,657	3.3	2,250	2.1	1,704
Mesopotamian	IRQ	acm	32	20,848	0.4	446	2.2	1,511
North Mesopotamian	IRQ	ayp	4.7	3,032	–	–	0.7	414
Sanaani	YEM	ayn	0.70	1,115	–	–	–	–
Ta'izzi-Adeni	YEM	acq	0.8	1,224	–	–	–	–
Total	-	-	4799.6	2,529,821	60.4	45,976	76.2	56,771
Unspecified	unk	unk	1,371.4	634,411	11.8	14,278	32.80	25,753

Table C.1: Dialect-wise breakdown of transcribed Arabic audio by train/dev/test splits (hours and utterances). Asterisks (*) indicate total hours after splitting datasets without canonical splits of train, dev, and test. We do not find any datasets corresponding to the following ISO codes: Algerian Saharan Arabic (aao), Tajiki Arabic (abh), Baharna Arabic (abv), Omani (acx), Cypriot (acy), Dhofari (adf), Saidi (aec), Uzbecki (auz), Eastern Egyptian Bedawi Arabic (avl), Hadrami (ayh), Sudanese Creole Arabic (pga), Chadian Arabic (shu), and Shihhi Arabic (ssh). While we did identify small portions of Libyan (ayl), we did not have enough to constitute a credible piece of our final benchmark.

country ranging from over 92% male for Palestine to 57% male for Morocco.

Common Voice (Ardila et al., 2019): The most recent version (as of writing) of Common Voice, 21, consists of 92 hours of validated read speech. Sentences are read by volunteers and recorded on their own devices, for instance a laptop mic, with the sentences themselves submitted and verified by volunteers. The sentences selected are MSA, although Common Voice does support dialects of other languages (for Cantonese and Minnan for Chinese on top of Mandarin).

FLEURS (Conneau et al., 2023): This smaller dataset, consisting of 8 hours of MSA. To enhances multilingual speech research, the FLEURS dataset was proposed using aligned text for all languages, originally sourced from English Wikipedia as part of the FLORES-101 project (Goyal et al., 2022). FLEURS takes these sentences and recruited native speakers to record each sentence. In the case of the Arabic split of the dataset, Egyptian speakers, with varying degrees of accent, provided the voices.

GALE (Walker et al., 2013)[7]: GALE phases 2-4 produced both conversational (578 hr.) and broadcast news (632 hr.) datasets for Arabic speech. While mainly MSA, a significant amount is in DA, which has lead to additional work to identify, extract, and use the DA segments for ASR systems (Soltau et al., 2011; Alsharhan and Ramsay, 2020). In annotating parts of GALE Phase 3 (Alsharhan and Ramsay, 2020) identified 44 hours of DA, mainly consisting of Levantine dialect. Gender imbalance also persists between dialects, with the work of (Alsharhan and Ramsay, 2020) indicating, gender ratios of at most 41% female speakers with Levantine dialect, and as little as 5% with Iraqi, of the amount they annotate.

Gulf Conversational Speech (Appen) (Appen Pty Ltd, 2006a): This conversational speech datasets consists of roughly 47 hours of Gulf dialect speech. The associated transcripts provide single references with full diacritization.

Iraqi Conversational Speech (Appen) (Appen Pty Ltd, 2006b): A 50 hour telephone speech dataset consisting of spontaneous conversational

[7]For simplicity this is the first of the GALE Arabic speech series

speech. Like the other Appen dataset, this is a 8khz recording with single reference but fully diacritized transcripts.

L2-KSU (Alrashoudi et al., 2024): This small (6 hour) dataset offers a unique selection of L1 and L2 Arabic speakers recording read MSA sentences. The majority of the L2 speakers are from African language backgrounds, while L1 Arabic speakers reflect a mixture of Egyptian, Gulf, and Levantine dialect backgrounds.

MASC (Al-Fetyani et al., 2023) The Massive Arabic Speech Corpus is a 1000 hour multigenre and multidialectal dataset sourced from Arabic YouTube channels with manually uploaded transcripts. A majority (569 hrs) of the language being MSA, and the top 5 represented dialects being: Syrian (197 hrs), Egyptian (120 hrs), Jordanian (39 hours), Saudi (31 hours) and Lebanese (23 hrs). The audio was lightly filtered with manual checking to ensure the captions were in Arabic, however, alignment issues may remain. A large majority of the speakers were identified as male accounting for 74% of the main speakers (excluding mixed gender segments) and accounting for 79% of the total speaking time. Text preprocessing involved using the Maha library to perform Alef normalization, including splitting Lam Alef to separate Lam and Alef characters, and normalize Teh Marbuta to Heh. The dataset is provided in 16khz, 16-bitdepth, 1-channel recordings.

MGB-2 (Ali et al., 2016): The multigenre broadcast corpus 2 consists of over 1,200 hours of mainly MSA audio sourced from 19 different Al Jazeera programs covering a wide range of program formats. The audio is mainly political coverage, with 24% falling into other topics which the authors indicate include society, economics, media, law, and science. The original paper does not give a very precise estimate of the amount of dialectal language used (only that it accounts for no more than 30%), however (Mubarak et al., 2021) performed an analysis of the test set, and indicates 78% MSA with 22% dialectal. Gender imbalance is present in the dataset, with (Mubarak et al., 2021) indicating a ratio of 78% (male) to 11% (female) speakers in the MGB-2 test set. These ratios are not known for the training split of the MGB-2. While there is no overlap in segments from the training to the evaluation splits, no speaker linking has been performed to confirm that there is not a speaker overlap. Surface normalization of the text is applied.

MGB-3 (Ali et al., 2017): 15.4 (4.6/4.8/6.0) hour Egyptian Arabic dataset sourced from YouTube videos covering a number of different genres (balanced across seven broad categories such as "sports" and "family"). The authors note 1.4 hours include overlapping speakers with the rest of the audio single speaker recordings. Four different transcriptions are available for each recording to account for the lack of standardized orthography. The authors do not provide demographic information about this dataset. Surface normalization is applied for alef, yah, and hah.

MGB-5 (Ali et al., 2019): 13 (10.2/1.3/1.4) hour Moroccan Arabic dataset sourced from YouTube videos and following a similar strategy as MGB-3. This includes a balance of different genres, multi reference with four different annotations, and surface text normalization. The largest departure from MGB-3 is in the way that the dataset has been split, with significantly smaller development and testing sets in comparison to MGB-3. No demographic information is provided.

SADA (Alharbi et al., 2024): This corpus of 668 hours of Saudi dialect speech (of which 437.6 hours have transcribed labels), represents one the largest transcribed dialectal datasets. There is a breadth of coverage in terms of Saudi dialect and genre of program (Comedy, Drama, Cooking etc.). Surface normalization was applied to the transcripts, and standard spelling is used for MSA, however, it is unclear how annotators agreed on transcription standards for the dialectal Arabic. While most of the dataset corresponds to ISO-639-3 level language codes (e.g. Najdi Arabic), there is ambiguity with how they define Maghrebi Arabic (which could be one of a number of North African Arabic languages), as well as their use of Northern (*Shamali*) and Southern (*Janubi*).

QASR (Mubarak et al., 2021): 2041 hour dataset mainly containing MSA from multigenre Al Jazeera recordings. The majority of language is expected to be MSA, however, dialectal Arabic is present. A small sample of 6,000 utterances were identified as containing code-switching between MSA, English, and/or French. A split of 4,000 segments were annotated for Speaker and Dialect ID purposes.

Quran Speech to Text (OpenSLR) (OpenSLR): Recordings of verses of the Quran taken from https://quran.ksu.edu.sa and resampled to 16khz. As these are recited verses, they are potentially closer to sung utterances than read utterances, and may provide particular distinct domain of audio.

Tunisian MSA (OpenSLR) (OpenSLR, 2003): This 11.2 hour dataset consists of read and prompted utterances, from Tunisian speakers in MSA. No additional demographic information is provided, nor does it contain any documentation of methodology for data collection.

Codemixed Datasets

Codemixing refers to the use of two different languages within a single utterance (Albirini, 2011), and is quite common in real world dialectal Arabic conversations, with the language mixed in dependent on the country and context.

ArzEn (Hamed et al., 2020): A 12 hour dataset consisting of spontaneous speech from interviews with Egyptian university students and teaching assistants (with the addition of one university employee). The focus of this dataset is on English and Egyptian Arabic codemixing (one of the two main modes of codemixing in Egyptian society, the other being MSA and Egyptian Arabic), with the interview style setting supportive of this natural speaking style. The vast majority of utterances (89%) consist of Egyptian Arabic with English inserted.

Mixat (Al Ali and Aldarmaki, 2024): This 14.9 hour dataset leverages podcast audio of hosts that naturally codemix between Emirati (a Gulf Arabic dialect) and English in their recordings. Unlike Arzen, which consists mainly of codemixed sentences, Mixat only contains a portion (36%) which contain codemixing. The majority is monolingual Emirati Arabic.

Saudilang Code-Switch Corpus (SCC) (SDAIA, 2022): 5 hour dataset sourced from Thmanyah podcast. The datasets consists of English and Arabic code-switch dialogues designed primarily as an out-of-domain test set.

ZAEBUC-Spoken (Hamed et al., 2024): Using a fairly novel Zoom-based collection method, this dataset consists of brainstorming discussion between students. The study used a mix of English speaking facilitators (from a wide range of accents), as well as Arabic speaking facilitators (primarily speaking MSA, but occasionally MSA with Egyptian codemixing), and the students spoke in a mix of English and Gulf Arabic, as well as MSA and Gulf Arabic. The majority of the utterances are either monolingual English or Arabic, with codemixing Eng-Arabic accounting for only 16% of the total utterances. For MSA-Dialectal codemixing, only a portion of the Arabic utterances were annotated using a five-level scale of dialectness. The majority of these utterances (58%) consisted of either pure MSA or imperfect MSA, without dialectal markers.

TTS-oriented datasets

The following speech datasets indicate suggested use for training TTS systems.

ASC (Halabi, 2016): Single speaker speech corpus 3.7 hour datasets with aligned diacritized texts for the purpose of training MSA TTS systems. Detailed diacritization and phonetic information is provided.

ClArTTS (Kulkarni et al., 2023): In contrast to many of the other datasets, this 12 hour dataset uses classical Arabic . Recorded at 44.1kHz.

ArVoice (Toyin et al., 2025):
83.52 hours of which 73.5 is synthetic audio (11 voices) and 10 hours of human speakers (7 voices). The origin of the text is Tashkeela, Khaleej, and a modified version of the texts used in ASC(Halabi, 2016). All texts are written in MSA.

Iraqi TTS (Kharrufa et al., 2024): With 3.7 hours of MSA and 1 hour of dialectal Arabic speech, this dataset is designed for training TTS systems that can handle a mix of MSA and dialectal speech. Lack of description of the speakers, as well as lack of precision in describing the dialect limit its utility.

Emotion Recognition Datasets

Some datasets have additional emotion labeling for emotion recognition models.

Kasdi-Merbah University Emotional Database of Arabic Speech (Belhadj et al., 2023): This is a small, two hour, dataset aimed at emotion recognition. Notably for the small size of the dataset a large number of speakers are included including 254 female, 246 male speakers, each reading 10 sentences. The audio was recorded at 44 kHz.

KSUEmotions (Meftah et al., 2017): 5 hours of emotion-labeled speech. The speech is read newswire, however, it is unclear how the choice of news text paired with somewhat artificially selected emotions impact the real world performance of emotion classifier trained on this dataset.

D Dataset Analysis

Dialect Coverage Across the datasets we find a large difference in the quantity of audio for each dialect, with limited amounts for Algerian (arq), Yemeni dialects (acq, ayn, and ayh), Sudanese (apd), Libyan (ayl), Tunisian (aeb), and Hassaniya (mey) (see Table C.1).

Age Of the datasets, only six provide age demographics for speakers, which we provide in Figure D.2. Of the datasets which provide age information, there appears to be a relative lack of older speakers represented in the datasets, with the vast majority of speakers under 40 years old.

Gender A common challenge in speech processing dataset curation is ensuring a relatively equal gender distribution of speakers. While normative expectations to obtain balanced gender ratios may not guarantee fair gender performance in general (see (Garnerin et al., 2021)), few papers included gender information (16 out of 28), and those that did often indicated wildly imbalanced ratios that may be dependent on the collected country (Talafha et al., 2024). Existing benchmarks rarely provide gender breakdown of performance, further hindering development of equitable speech processing systems. Figure D.3 provides breakdown of the duration from a given gender, with quite a few are extremely unbalanced, with seven reporting over 80% of the duration belonging to just one gender. Fig. D.3 shows an overview of the gender distributions.

Domain We provide a full breakdown of domain per dataset in Fig. D.4 as well as a mapping from the raw metadata field to the normalized field in Table D.1. We find that the distribution of domains is not particularly diverse, with People & Society accounting for the vast majority of utterances, however, utterances with unknown domain (either a generic label, "misc," or not provided) account for the largest amount of total duration (2625 hours).

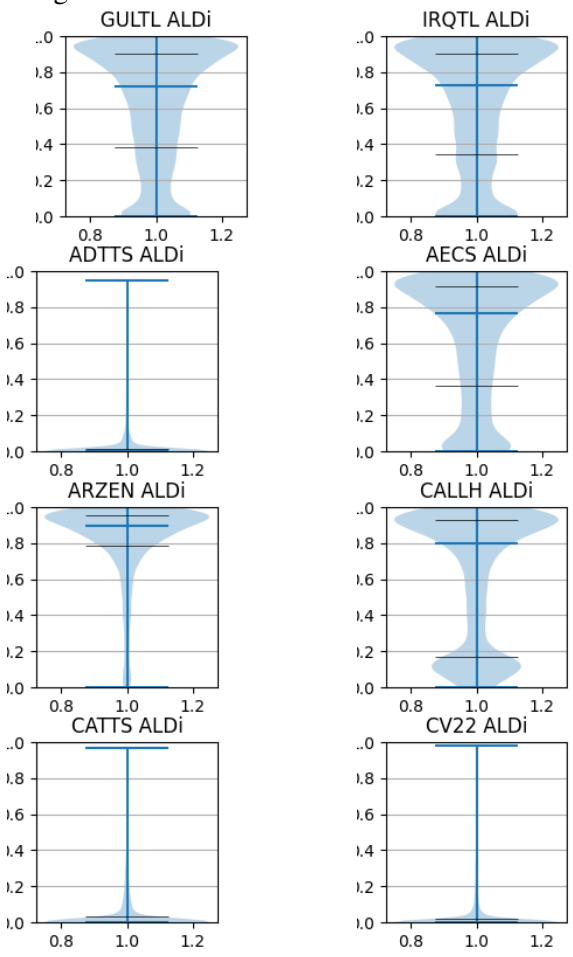
D.1 Dialectness

Binary MSA vs DA classifier : We use a finetuned version of MARBERTs (Abdul-Mageed et al., 2021), finetuned on an in-house text dataset.

ALDi regression model : Also a finetuned MARBERT (Abdul-Mageed et al., 2021), trained on the Arabic Online Commentary (AOC)-ALDi dataset, which is augmented version of the AOC dataset (Zaidan and Callison-Burch, 2011). In contrast to the binary model, which have a classification head, ALDi uses a regression approach to provide an estimate between 0 and 1 of the level of dialectness. The original ALDi-AOC dataset uses the following bins: [0, 0.11[; [0.11, 0.44[; [0.44, 0.77[; [0.77, 1[for MSA, little DA, mixed, and mostly DA respectively.

A limitation of textual approaches, however, is that they are not able to identify sentences, where the dialectal and MSA sentence would be written the same way without diacritics.

Detailed Charts For ALDi we provide violin plots showing the distribution of ALDi scores from 0 (MSA) to 1 (purely dialect). Quartiles are shown in light black.



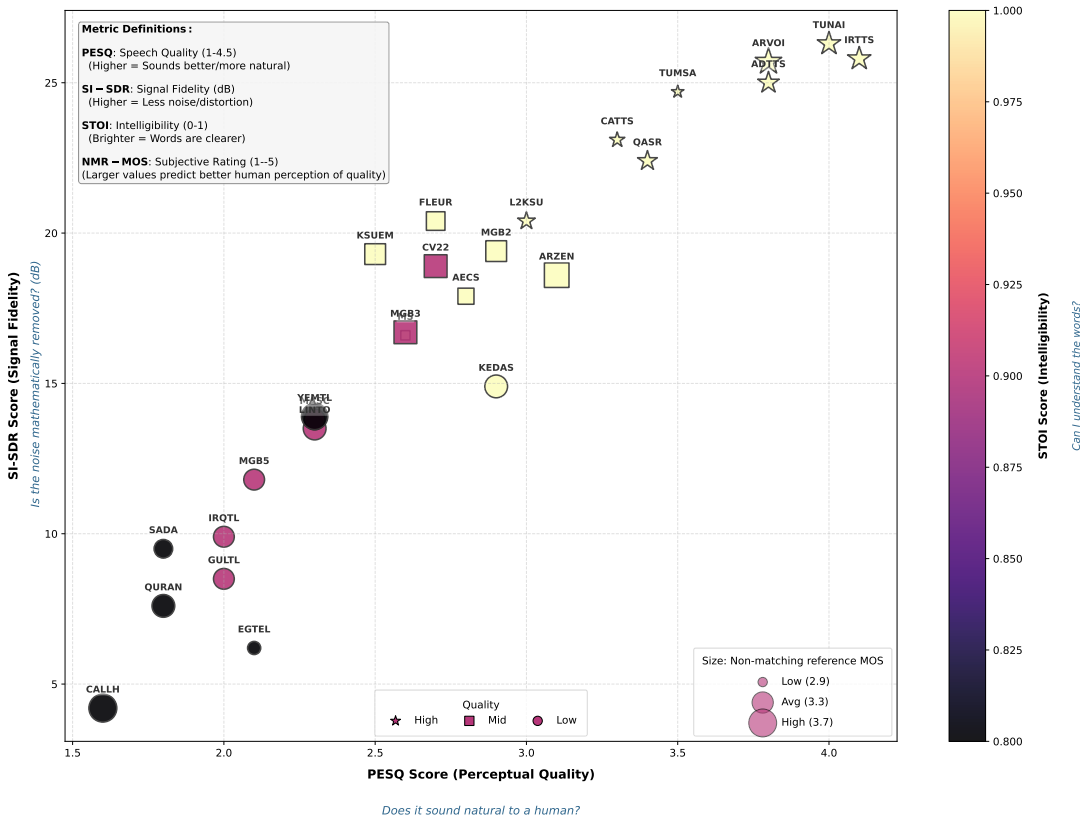


Figure D.1: We plot the relationship between the four audio quality metrics (see Appendix D for detailed descriptions of each metric).While SI-SDR, PESQ, and STOI (shown as color) largely align, we note that this is not the case with the NMR-MOS model (shown as size).

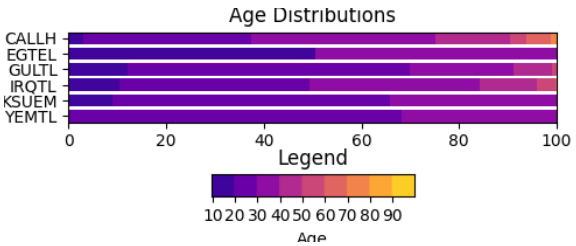
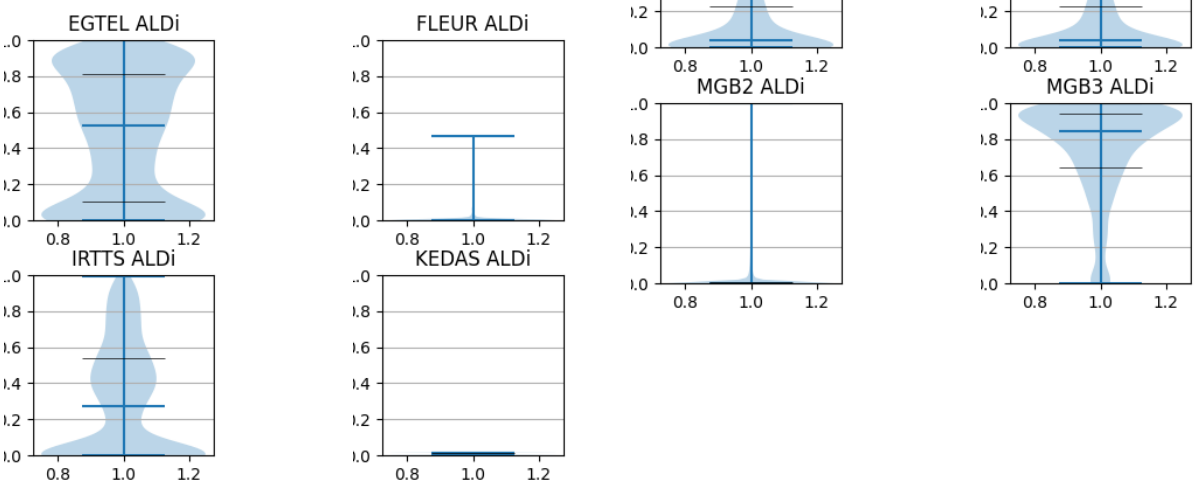


Figure D.2: Percentage distribution of age ranges in Datasets with age information of speakers. Speakers binned to 10 year age intervals.



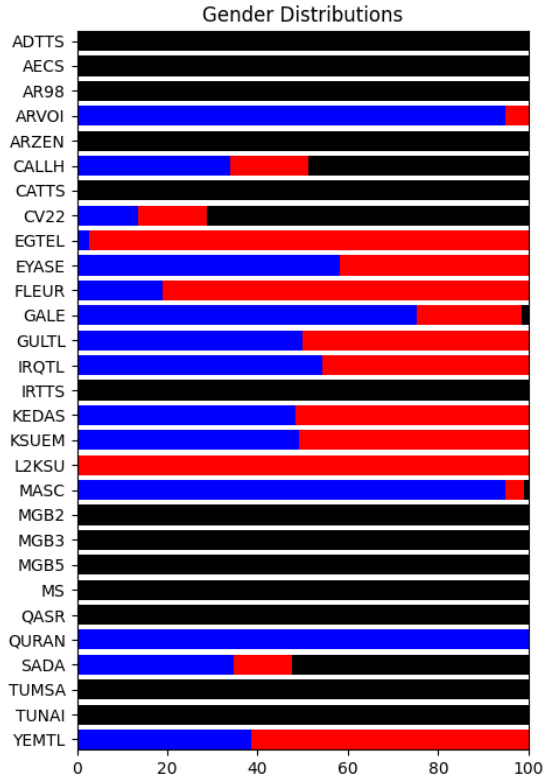
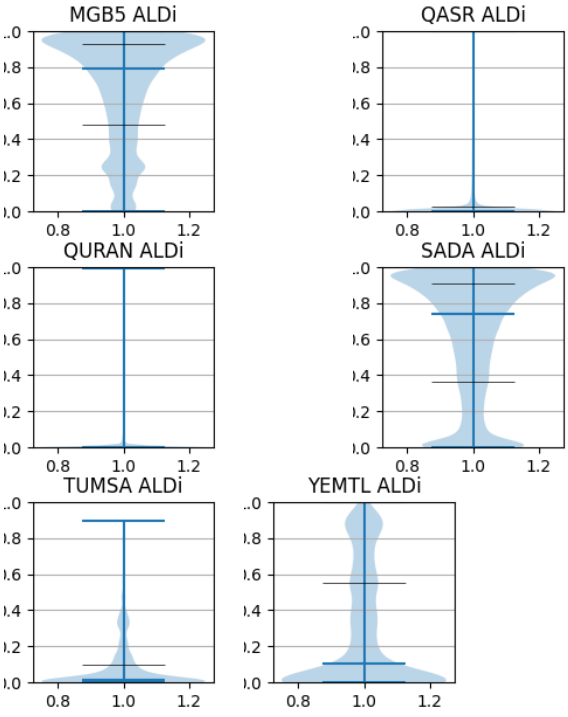


Figure D.3: Percentage of Male (blue), Female (red) and unknown (black) speakers where datasets report gender.



D.2 SQUIM

PESQ can be thought of as a prediction of subjective human perception of audio quality (albeit measured with an objective model), and includes cognitive and auditory modeling to predict a mean opinion score in the range [1,4.5] (Rix et al., 2001). SI-SDR, is an updated signal-to-noise metric that accounts for issues in the signal-to-noise formula caused by the scaling of the estimated audio (Le Roux et al., 2019), the updated SI-SDR formula accounts for this to provide a metric that is independent of amplitude scaling of the audio. Like signal-to-noise, the metric is measured in decibels, and represent how many orders of magnitude the target audio is compared to any distortions. Finally, STOI measures intelligibility of audio based on 382ms audio chunks, and measures the correlation between chunks of clean and noisy audio, with the higher alignment indicating better intelligibility of the audio file (Taal et al., 2011).

TorchAudio-Squim uses a transformer based model to simulate these metrics without the need for isolated clean audio. While these objective measures of audio quality may not correlate well with human judgment (Pranay Manocha and Zeyu Jin and Adam Finkelstein, 2022), they may still have an impact on model performance for models that are highly sensitive to domain and channel information, and by providing this information, we aim to set the stage for future evaluation of these factors. As an alternative, we also use Squim's subjective model, which is based on the non-matching reference approach in (Pranay Manocha and Anurag Kumar, 2022), this model provides a prediction correlated with human mean opinion score (MOS) ratings. We randomly sample 5 seconds of clean audio from the DAPS dataset (Mysore, 2014) to provide our non-matching reference.

As PESQ and the Squim-Subjective metric are aligned to mean opinion scores, they are the most interpretable of the metrics, higher scores should thus align to human perception of audio quality. On the other hand, SI-SDR, which reflects the relative ratio of target audio to distortion on a decibel scale; and STOI with its 0 to 1 rating of intelligibility, both lack easily interpretable meaning. However, they still provide a valuable metric when compared to other utterances on the same scale as a way to characterize the noisiness of a particular dataset.

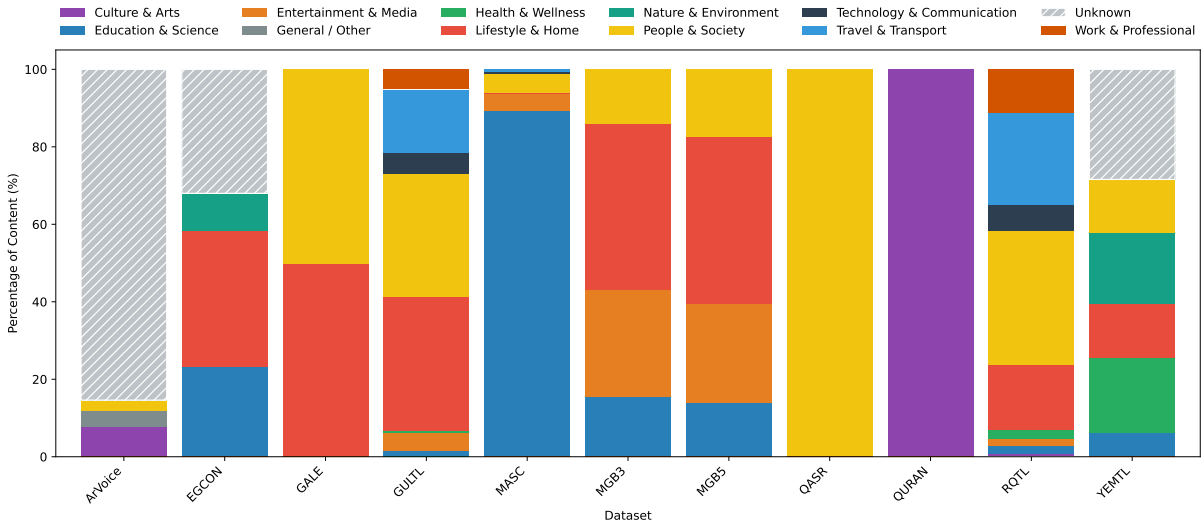


Figure D.4: Per-dataset, per-domain breakdown of utterance counts and total hours. Dataset names are shown only on the first row of each block. We omit dataset which do not provide any information about domain (11 out of the 28).

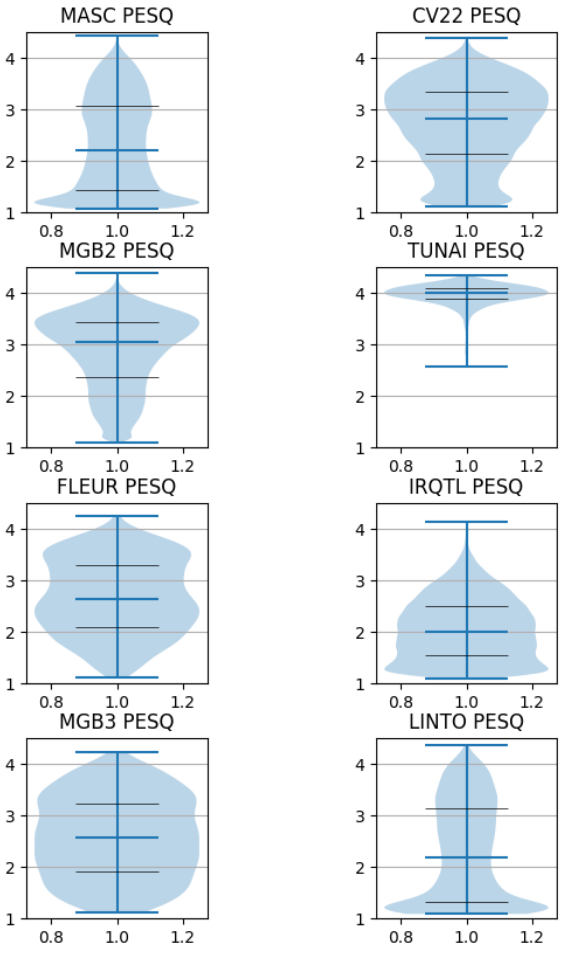
D.3 Noise and intelligibility.

We next analyze predicted recording conditions using our audio-quality proxies. Consistent with the recording setups typically used for TTS corpora (Table D.2), ADTTS, CATTS, and IRTTS exhibit high predicted quality under the objective proxies (all mean predicted PESQ > 3 and mean predicted SI-SDR > 20). Several other read-speech datasets, including ARVOI, TUNAI, and TUMSA, show similarly strong scores. Within broadcast-domain recordings, QASR scores well by these proxies and appears higher quality than the similarly situated MGB2 dataset as well as the other large-scale MSA dataset, MASC.

In contrast, mobile and telephone recordings tend to score lower on these proxies, with several datasets exhibiting mean predicted PESQ ≤ 2 and mean predicted SI-SDR < 10, including GULTL, IRQTL, and CALLH. Other notable datasets include SADA and QURAN, which have mean predicted PESQ < 2, mean predicted SI-SDR < 10, and mean predicted STOI around 0.8.

Predicted NMR-MOS yields a different ordering from the objective proxies: CALLH and ARVOI are both assigned relatively high predicted MOS (3.7), despite representing opposite ends of the predicted objective-quality spectrum (PESQ: ARVOI 3.8 vs. CALLH 1.6). §D provides detailed figures for PESQ, SI-SDR, and STOI for each dataset. Therefore, rather than collapsing quality into a single score, our benchmark/adaptation protocol retains dataset provenance and reports results in the context of these complementary audio-condition signals.

D.4 SQUIM PESQ



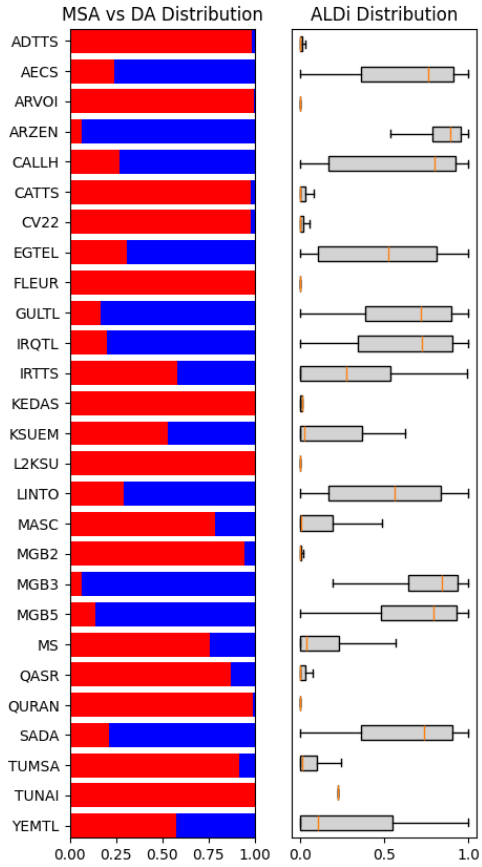
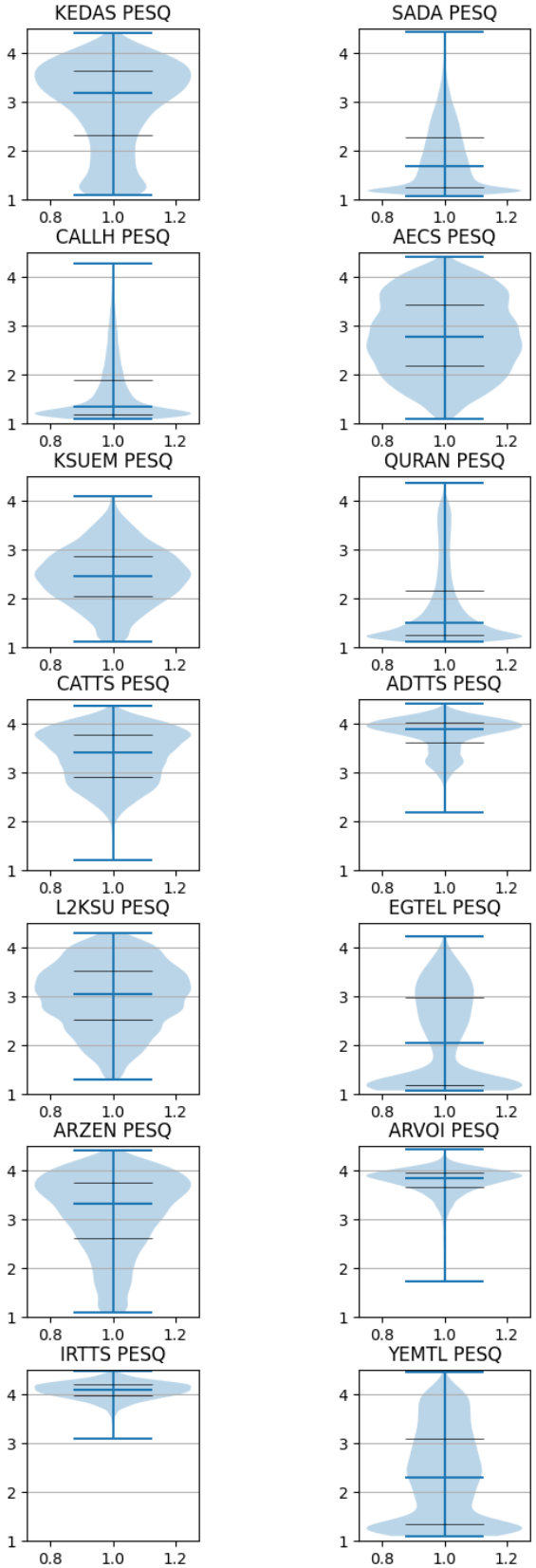
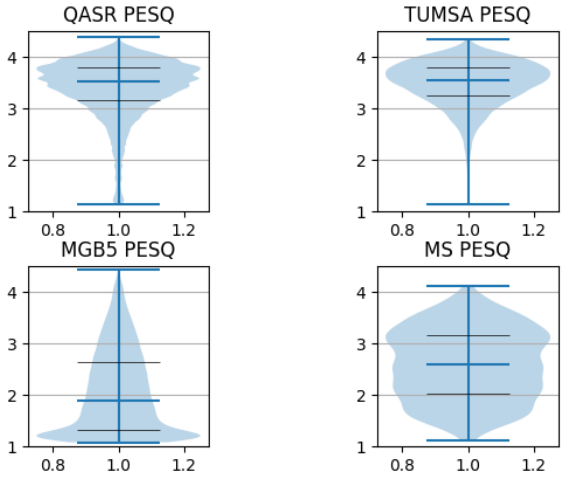
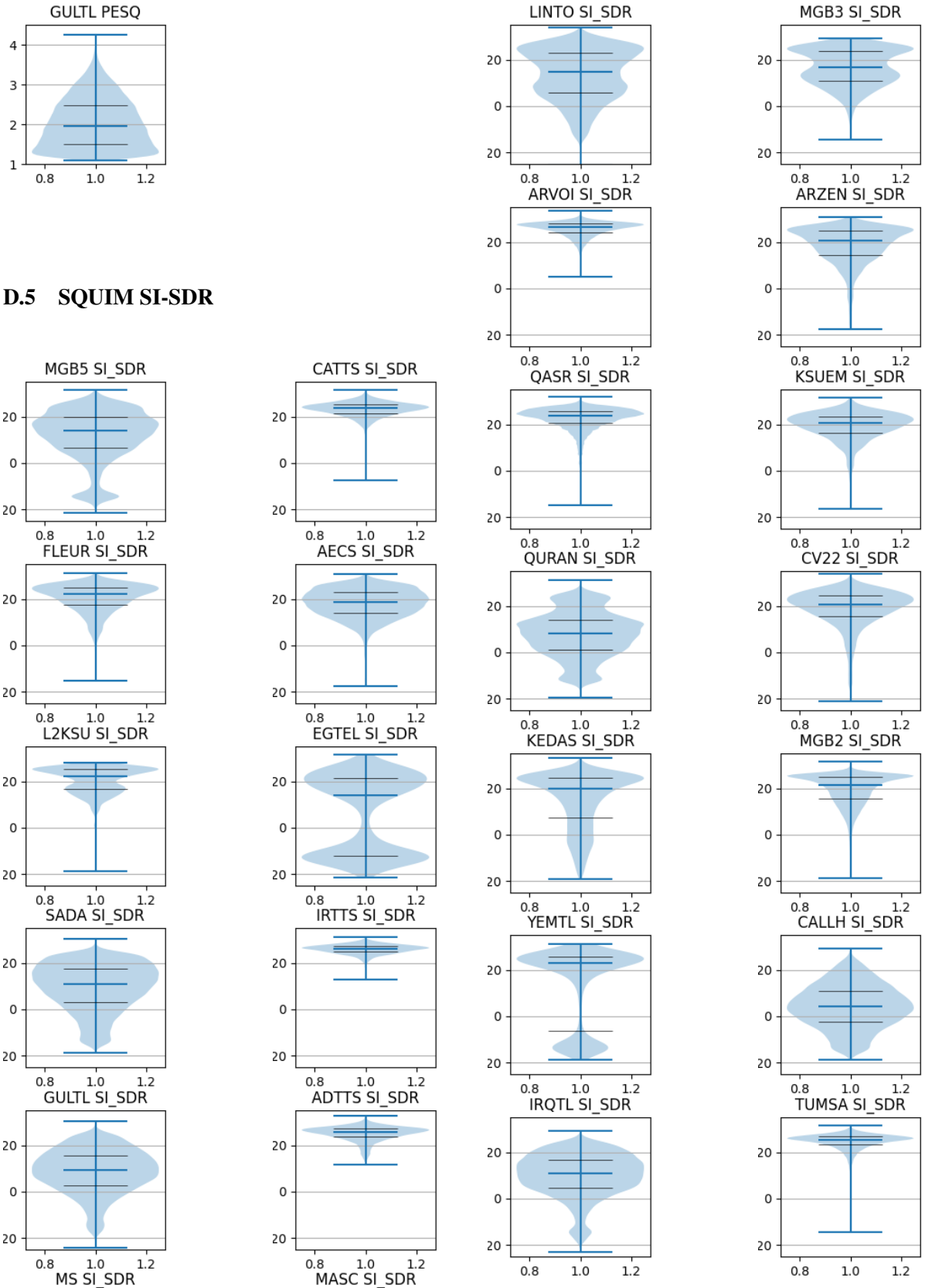


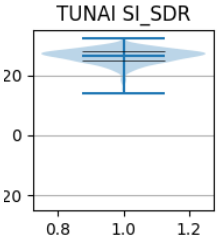
Figure D.5: (Left) Percentage of MSA (red) and DA (blue), based on predictions from our binary classifier. (Right) ALDi box plots, a higher ALDi score implies more dialectness.

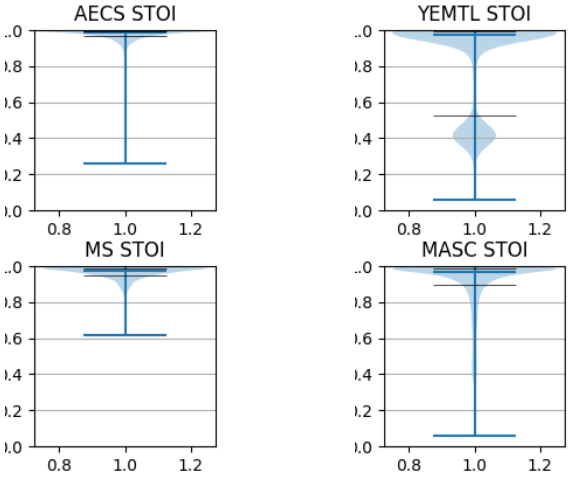




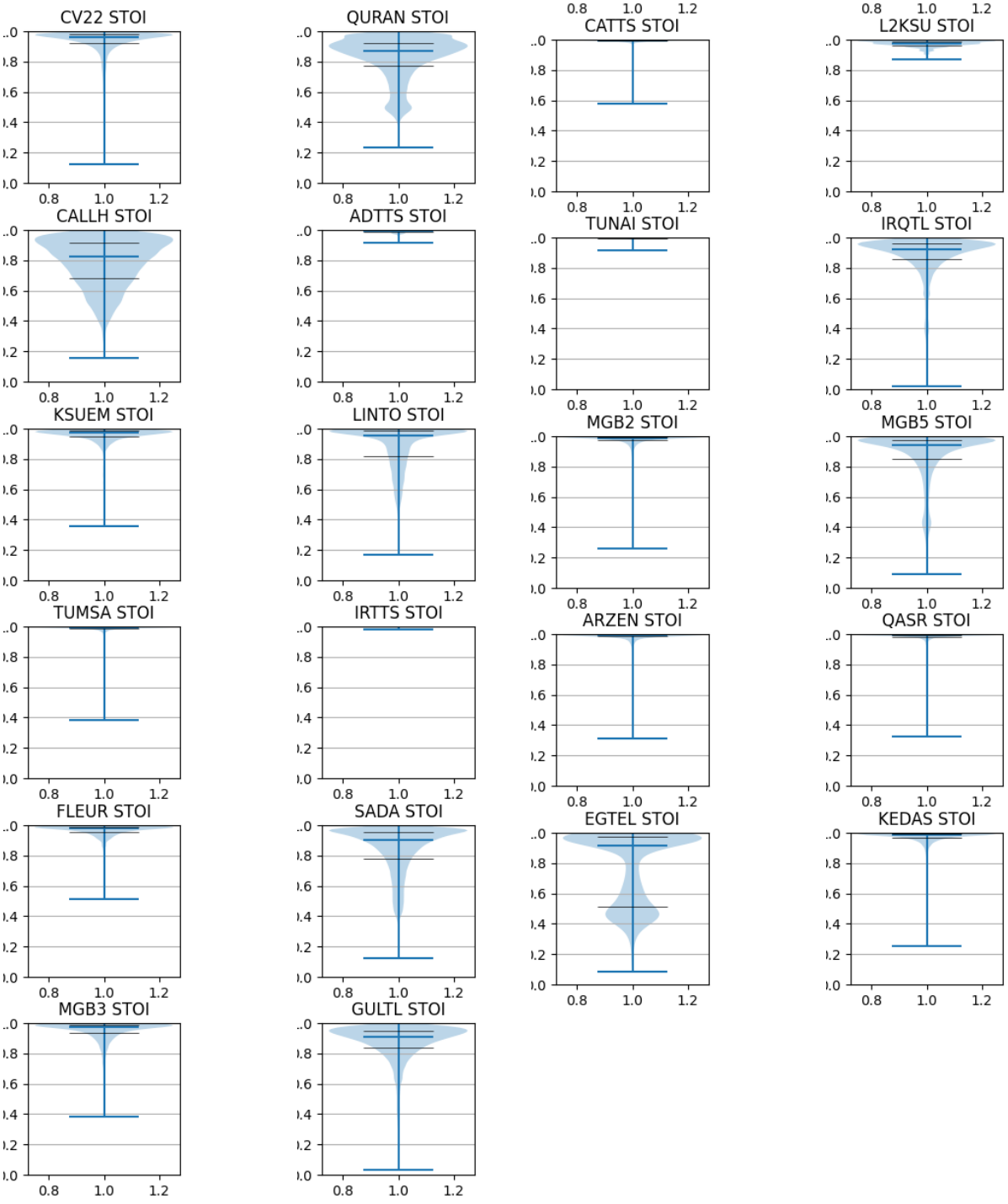


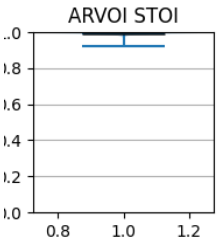
D.5 SQUIM SI-SDR

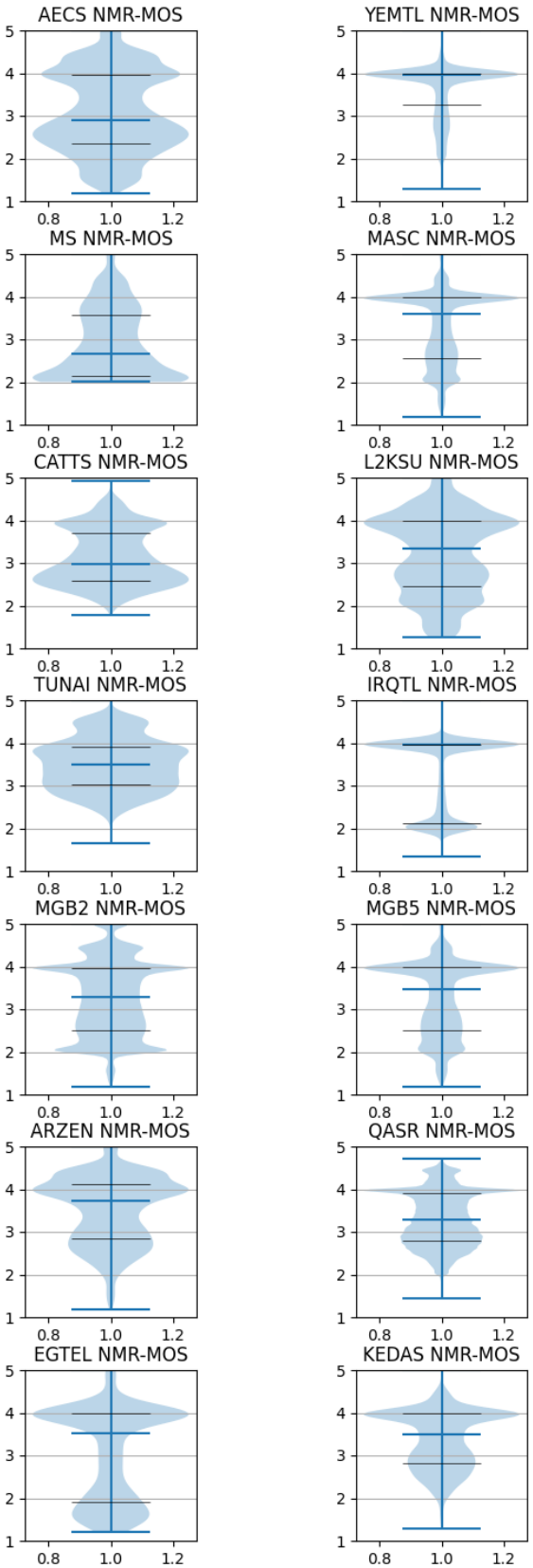

TUNAI SI_SDR
20
0
20
0.8
1.0
1.2


AECS STOI
YEMTL STOI
MS STOI
MASC STOI
0.8
1.0
1.2

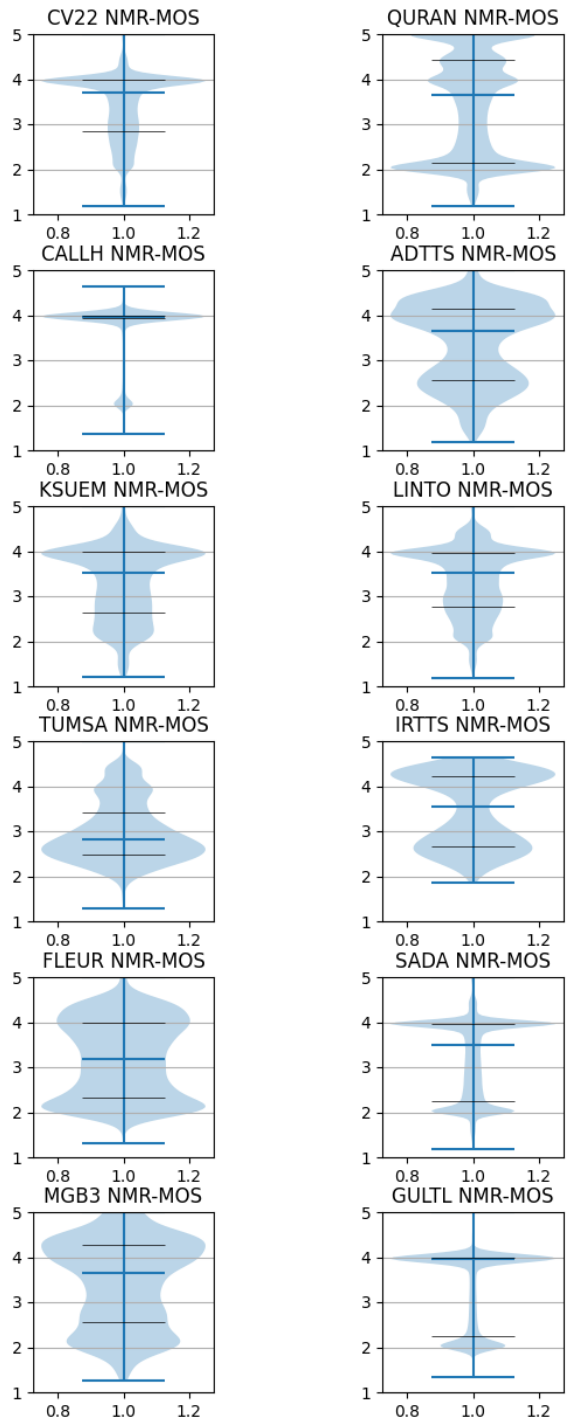
SQUIM STOI

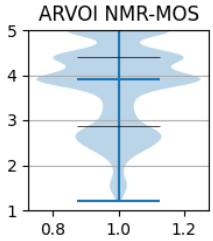

CV22 STOI
QURAN STOI
CATTS STOI
L2KSU STOI
CALLH STOI
ADTTS STOI
TUNAI STOI
IRQTL STOI
KSUEM STOI
LINTO STOI
MGB2 STOI
MGB5 STOI
TUMSA STOI
IRTTS STOI
ARZEN STOI
QASR STOI
FLEUR STOI
SADA STOI
EGTEL STOI
KEDAS STOI
MGB3 STOI
GULTL STOI
0.8
1.0
1.2





D.6 SQUIM NMR-MOS





E Experimental Setup

F Models

Whisper-large-v3 (Radford et al., 2023): This transformer-based ASR model was trained to perform long-form transcription, speech-to-text translation, and language ID using large quantities of multilingual audio sourced from internet videos with subtitles. There may be a bias towards producing MSA-like sentences, due to the decoder's language modeling capabilities combined with prevalence of writing subtitles in MSA.

MMS (Pratap et al., 2024): In contrast to Whisper's approach, MMS uses a CTC-based model trained primarily on translated religious texts. Like Whisper, however, MMS treats all Arabic dialects as belonging to one language, however, this is less likely to be problematic in converting sentences into MSA structure due to CTC not relying on language modeling.

Seamless M4T v2 (Barrault et al., 2023): While originally designed for speech-to-speech translation, this model performs ASR by leveraging only speech encoder and text decoder of the pipeline to generate target text. Aside from MSA, this model supports two other Arabic dialects: Moroccan (ary) and Egyptian (arz).

Omnilingual (Omnilingual et al., 2025): Like many of the other models, this is also a transformer-encoder based model, with options for either CTC text decoding or transformer decoder based language modeling. What sets it appart from the other models is use of a new high quality dataset focusing on extremely low-resource languages alongside existing publicly available resources. While the exact list of open source datasets that Omnilingual was trained on is not available, they also provide a newly collected set of recordings for many Arabic dialects. In total 18 dialects are provided in this new dataset, however, of these only 13 have more than 1000 utterances across all splits including: Baharna (abv), Hijazi (acw), Omani (acx), Tunisian (aeb), Saidi (aec), Gulf (afb), Sudanese (apd), Algerian (arq), Najdi (ars), Moroccan (ary), Egyptian (arz), Libyan (ayl), and North Mesopotamian (ayp).

F.1 Normalization

The normalization pipeline $\mathcal{N}(x)$ transforms an input string x through a sequential cleaning process. If the text is detected as Buckwalter transliteration, it is mapped to the Arabic script. The string then undergoes Unicode NFKD decomposition, facilitating the removal of all combining characters and Arabic diacritics (*Tashkeel*). Punctuation marks from both Latin and Arabic scripts—including symbols such as ،, ؛, and ؟—are eliminated. Orthographic unification is then applied: all Hamzated Alef forms {أ, إ, آ} are normalized to bare Alef (ا), Ta-Marbuta (ة) is converted to Ha (ه), and Alif-Maqsura (ى) is standardized to Ya (ي). Finally, all text is trimmed and consecutive whitespace is collapsed into single spaces.

G Results

Please see Tables G.1 & G.2 and Figs. G.2 & G.1 for detailed breakdown of the performance of off the shelf models.

As well as Tables G.4 (Gender) G.5 (Dialectness) and the following audio quality metrics G.6 (PESQ), G.7 (SI-SDR), G.8 (STOI), G.9 (Non-matching reference) for granular results on the DEV set.

G.1 Human Sanity Check

To ensure the reliability of our automated profiling tools, we conducted a human-in-the-loop validation process focusing on two critical dimensions: audio quality and dialect identification. A stratified sample of the datasets was reviewed by native Arabic speakers to corroborate the output of our automated models.

Audio Quality Validation We compared our automated Perceptual Evaluation of Speech Quality scores against human Mean Opinion Scores (MOS). While the automated metrics generally correlated with human perception, our analysis revealed distinct domain-specific biases. The automated models frequently underestimated the quality of datasets containing expressive prosody or distinct acoustic environments. For instance, in the `ksuemotions` and `quran-speech` datasets, automated scoring yielded low values (approx. 1.15–1.55), whereas human annotators rated the qual-

ity as high (4.0/5.0). This indicates that standard reference-free metrics may penalize the silence intervals or dynamic range inherent to these domains. Conversely, synthesized speech (e.g., `arabic-diacritized-tts`) occasionally received high model scores despite lower human ratings, highlighting the model's insensitivity to certain unnatural robotic artifacts.

Dialectness Validation We further validated the automated dialect labels using two granularities: a binary classification (MSA vs. Dialectal Arabic) and a fine-grained dialect identification (ALDI).

- **Binary Classification:** We observed near-perfect alignment between automated predictions and human annotation for the binary distinction between MSA and Dialectal Arabic, with accuracy scores approaching 100% across most datasets. This confirms that current tools are highly robust at distinguishing standard Arabic from regional varieties.
- **Level of dialectness:** The ALDi model achieved an accuracy of 91% (n=90) with 6 samples scoring low-dialectal when they should have been high, and 2 samples with segmentation errors in the transcription.

This divergence underscores the necessity of our mapping framework: while broad categories are easily automated, precise dialect mapping still benefits significantly from human verification and metadata standardization.

Raw Domain	Normalized Domain	Hours
missing	Unknown	437.6
academic	Education & Science	0.8
autos & vehicles	Travel & Transport	0.2
cars	Travel & Transport	25.1
comedy	Entertainment & Media	9.7
commands	Technology & Communication	3.6
cooking	Lifestyle & Home	18.8
daily activities	Lifestyle & Home	4.8
daily_activities	Lifestyle & Home	25.9
domesticity	Lifestyle & Home	0.7
drama	Entertainment & Media	6.7
education	Education & Science	439.9
education and health	Health & Wellness	1.2
entertainment	Entertainment & Media	14.6
environment and climate	Nature & Environment	0.7
family	People & Society	44.1
family-children	People & Society	8.4
familykids	People & Society	1.8
fashion	Lifestyle & Home	9.9
film & animation	Entertainment & Media	9.0
friends	People & Society	16.4
gaming	Entertainment & Media	0.3
general	General / Other	2.7
health	Health & Wellness	2.2
howto & style	Lifestyle & Home	4.2
human science	Education & Science	1.8
iraq news	People & Society	1.4
islamic classical books	Culture & Arts	5.4
legal matters	People & Society	0.4
marriage	People & Society	0.9
miscellaneous	General / Other	0.2
moviesdrama	Entertainment & Media	1.7
music	Entertainment & Media	5.6
news	People & Society	1,884.9
news & politics	People & Society	5.0
nonprofits & activism	People & Society	7.4
people & blogs	People & Society	23.8
personal characteristics	People & Society	0.3
pets & animals	Nature & Environment	0.2
places to go	Travel & Transport	1.7
places_to_go	Travel & Transport	6.3
poetry	Culture & Arts	0.6
ramadan	Culture & Arts	0.2
school	Education & Science	4.4
science	Education & Science	8.9
science & technology	Technology & Communication	2.1
shopping	Lifestyle & Home	15.7
sports	Lifestyle & Home	24.1
study and education	Education & Science	0.4
technology	Technology & Communication	9.3
telephone calls	Technology & Communication	1.9
telephone_calls	Technology & Communication	3.0
travel	Travel & Transport	10.2
travel & events	Travel & Transport	5.5
tv/cinema	Entertainment & Media	9.8
unknown	Unknown	1,594.6
women	People & Society	15.4
women's fashion	Lifestyle & Home	0.7
women's_fashion	Lifestyle & Home	1.1
work	Work & Professional	15.5
yesterdayactivities	Lifestyle & Home	5.8

Table D.1: Raw domain field mapping to normalized fields along with hours.

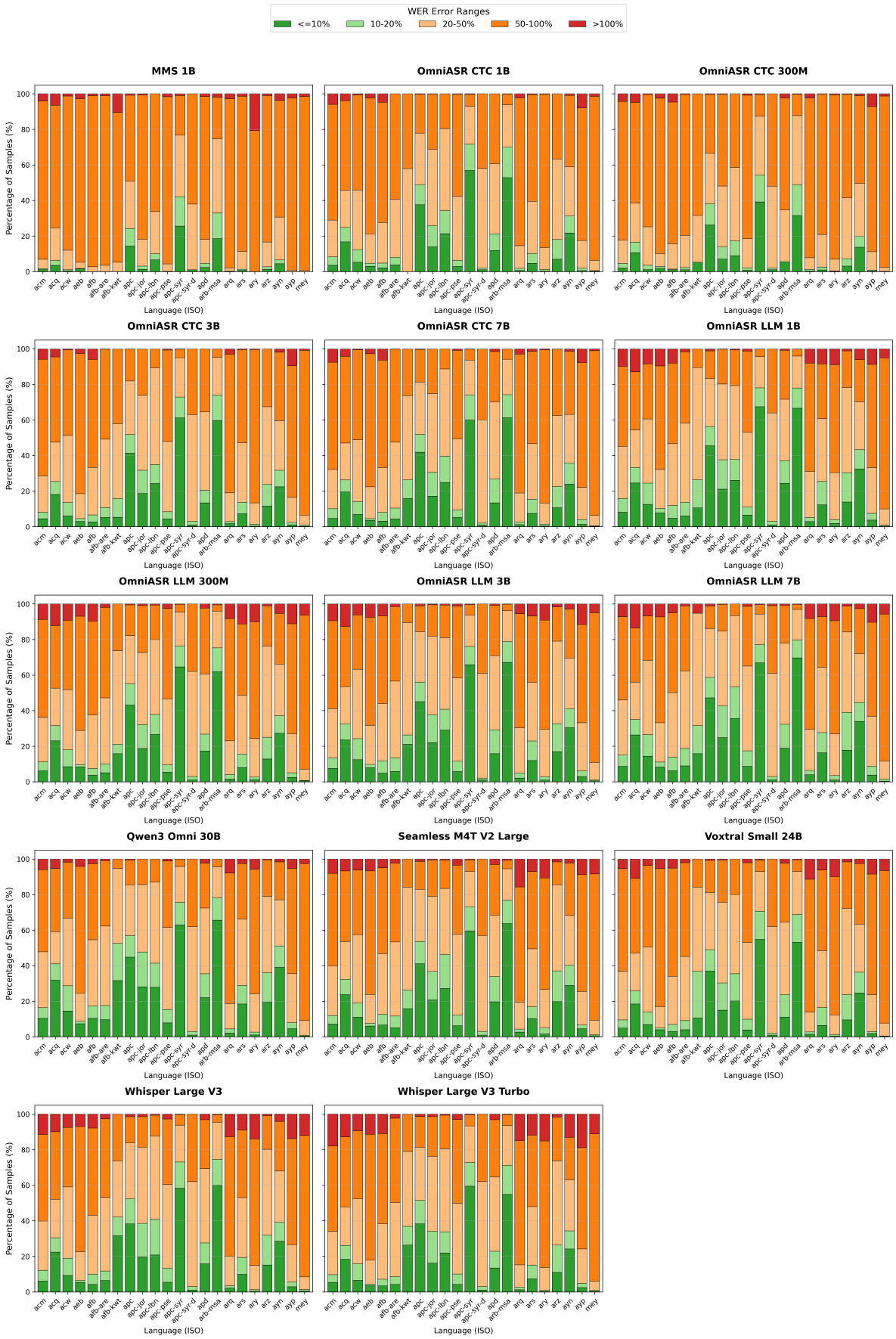


Figure G.1: WER distribution across languages for all ASR models from the three architectures listed in Table 2. Stacked bar charts show the proportion of samples in different WER ranges (from $\leq$10% to >100%) across multiple languages (ISO codes). Colors range from green (low error rates) to red (high error rates).

Figure G.2: CER distribution across languages for all ASR models from the three architectures listed in Table 2. Stacked bar charts show the proportion of samples in different CER ranges (from $\leq$10% to >100%) across multiple languages (ISO codes). Colors range from green (low error rates) to red (high error rates).

Dataset	PESQ ↑	SI-SDR (dB) ↑	STOI ↑	NMR-MOS ↑
ARVOI	3.8 (0.3)	**25.7** (3.1)	**1.0** (0.0)	**3.7** (0.9)
GULTL	2.0 (0.6)	8.5 (9.4)	0.9 (0.1)	3.3 (0.8)
IRQTL	2.0 (0.6)	9.9 (9.4)	0.9 (0.1)	3.3 (0.9)
ADTTS	3.8 (0.3)	**25.0** (3.2)	**1.0** (0.0)	3.4 (0.9)
AECS	2.8(0.8)	17.9 (6.3)	**1.0** (0.0)	3.1 (0.9)
ARZEN	3.1 (0.8)	18.6 (8.0)	**1.0** (0.1)	**3.5** (0.8)
CALLH	1.6 (0.6)	4.2 (9.3)	0.8 (0.2)	**3.7** (0.6)
CATTS	3.3 (0.5)	23.1 (3.2)	**1.0** (0.0)	3.1 (0.6)
CV22	2.7 (0.8)	18.9 (8.1)	0.9 (0.1)	3.4 (0.7)
EGTEL	2.1 (0.9)	6.2 (16.4)	0.8 (0.2)	3.0 (1.1)
FLEUR	2.7 (0.7)	20.4 (6.2)	**1.0** (0.0)	3.2 (0.9)
IRTTS	**4.1** (0.2)	**25.8** (2.0)	**1.0** (0.0)	**3.5** (0.8)
KEDAS	2.9 (0.9)	14.9 (11.8)	**1.0** (0.0)	3.4 (0.7)
KSUEM	2.5 (0.6)	19.3 (6.0)	**1.0** (0.1)	3.3 (0.8)
L2KSU	3.0 (0.7)	20.4 (5.8)	**1.0** (0.0)	3.2 (0.9)
LINTO	2.3 (1.0)	13.5 (10.9)	0.9 (0.1)	3.4 (0.7)
MASC	2.3 (0.9)	13.8 (10.7)	0.9 (0.1)	3.3 (0.8)
MS	2.6 (0.7)	16.6 (6.1)	**1.0** (0.0)	2.9 (0.8)
MGB2	2.9 (0.7)	19.4 (7.0)	**1.0** (0.1)	3.3 (0.9)
MGB3	2.6 (0.8)	16.7 (7.7)	0.9 (0.1)	3.4 (1.0)
MGB5	2.1 (0.8)	11.8 (10.8)	0.9 (0.2)	3.3 (0.9)
QASR	3.4 (0.6)	22.4 (5.3)	**1.0** (0.0)	3.3 (0.6)
QURAN	1.8 (0.8)	7.6 (9.7)	0.8 (0.1)	3.4 (1.1)
SADA	1.8 (0.6)	9.5 (10.1)	0.8 (0.2)	3.2 (0.9)
TUNAI	**4.0** (0.2)	**26.3** (2.5)	**1.0** (0.0)	**3.5** (0.6)
TUMSA	3.5 (0.4)	**24.7** (3.2)	**1.0** (0.0)	3.0 (0.7)
YEMTL	2.3 (0.9)	13.9 (16.4)	0.8 (0.2)	**3.6** (0.6)

Table D.2: Audio metrics overview. Values are reported as Mean (Standard Deviation). For bounded metrics the ranges are PESQ [1, 4.5], STOI [0, 1], SUB [1, 5]; SI-SDR scores are in decibels (dB). Notable low scores are highlighted in red, and higher quality scores in **dark green**.

Variety	Dev		Test	
	Dur (h)	Dataset	Dur (h)	Dataset
acm	0.4	(IRQTL 0.7, MASC 0.3)	1.0	(MASC 0.1, IRQTL 0.9,)
acq	0.9	(YEMTL 0.9)	0.9	(YEMTL 0.9)
acw	0.6	(SADA 0.6)	1.0	(SADA 1.0)
aeb	0.0	(MASC 0.0)	1.0	(TARIC 0.4, LINTO 0.6,)
afb	1.0	(MASC 0.0, CASA 0.5, GULTL 0.5)	1.0	(CASA 0.3, MASC 0.0, GULTL 0.7,)
apc	1.0	(MASC 0.7, IWSLT 0.0, SADA 0.0, CASA 0.3,)	0.9	(TUNAI 0.0, MASC 0.6, IWSLT 0.0, CASA 0.3,)
arb	1.0	(TUMSA 0.0, MASC 0.9, FLEUR 0.1, SADA 0.0,)	1.0	(TUMSA 0.0, MASC 0.8, FLEUR 0.1, SADA 0.0, ARVOI 0.1,)
arq	1.0	(CASA 1.0)	1.0	(MASC 0.1, CASA 0.9,)
ars	1.0	(SADA 1.0)	1.0	(SADA 1.0)
ary	1.0	(MGB5 0.9, MASC 0.0, CASA 0.1,)	1.0	(MGB5 0.9, MASC 0.0, CASA 0.1,)
arz	1.0	(MGB3 0.5, MASC 0.3, SADA 0.0, CASA 0.1,)	1.0	(MGB3 0.6, MASC 0.2, SADA 0.0, CASA 0.1,)
ayn	0.7	(YEMTL 0.7)	0.7	(YEMTL 0.7)
Total	**9.6**		**11.5**	

Table F.1: Benchmark Overview. Time in hours. Subsets with 0 hours indicate < 5 minutes of duration

Dialect	Encode-Only (CTC / Acoustic Encoder)					Encode-Decoder			Multimodal (Speech + LLM Core)					
	E1	E2	E3	E4	E5	ED1	ED2	ED3	M1	M2	M3	M4	M5	M6
MSA (arb)	38.94	26.18	16.49	13.5	13.8	18.51	15.55	14.7	17.52	11.51	13.57	11.52	11.3	**10.19**
Khaleeji (afb)*	89.62	81.54	74.96	73.54	71.72	213.89	105.14	61.89	80.23	**53.71**	121.57	84.05	91.34	68.39
Khaleeji (Kuwait) (afb-kwt)	86.47	59.6	58.43	48.37	45.81	35.17	30.85	33.65	34.53	**19.45**	39.59	30.92	30.71	29.48
Khaleeji (UAE) (afb-are)	84.27	70.13	58.37	53.92	56.07	68.09	59.48	54.7	60.69	46.22	55.53	50.23	50.08	**45.71**
Najdi (Saudi Arabia) (ars)	84.01	76.93	64.06	59.31	60.17	173.42	113.92	62.63	72.73	**46.9**	80.75	63.92	61.31	60.23
Hijazi (Saudi Arabia) (acw)	82.03	72.77	59.08	56.73	58.05	191.21	105.94	62.14	91.53	55.93	67.79	73.33	56.43	**53.49**
Sanaani Arabic (Yemen) (ayn)	74.99	58.25	50.36	50.49	47.81	188.3	59.06	44.25	47.67	**34.63**	52.32	53.7	46.19	40.37
Ta'izzi-Adeni Arabic (Yemen) (acq)	72.61	55.22	46.42	45.84	45.44	125.22	77.06	39.54	50.62	**28.77**	44.02	43.43	38.03	35.5
North Mesopotamian Arabic (Iraq) (ayp)	94.65	89.14	87.58	87.99	84.11	367.74	134.53	87.78	99.13	**75.45**	88.48	93.12	104.72	95.16
Mesopotamian Arabic (Iraq) (acm)	94.02	86.43	80.36	81.2	80.13	271.25	119.91	106.26	87.53	**71.16**	108.36	86.76	81.93	73.86
Levantine (apc)*	51.75	38.91	29.72	26.84	26.88	36.31	32.53	26.38	28.28	23.81	27.61	25.45	25.17	**23.53**
Levantine (Jordan) (apc-jor)	72.89	54.07	40.51	36.08	36.61	46.78	41.26	33.21	37.46	**27.2**	37.29	32.37	32.53	29.31
Levantine (Lebanon) (apc-lbn)	62.63	46.49	32.61	29.68	28.33	32.58	29.19	29.14	32.05	27.1	32.58	32.18	30.12	**21.56**
Levantine (Palestine) (apc-pse)	82.74	70.23	57.58	54.74	54.28	73.17	53.82	50.61	55.29	47.2	55.58	51.7	50.53	**45.63**
Levantine (Syria) (apc-syr)	34.17	23.64	15.11	13.59	13.81	19.15	14.33	14.54	16.34	12.02	12.28	**11.23**	12.43	12.3
Levantine (Syrian Damascus) (apc-syr-d)	55.97	52.04	49.36	48.63	49.16	48.65	48.74	49.34	49.15	48.69	48.68	**47.87**	49.1	49.03
Egyptian (arz)	74.32	58.87	45.7	42.28	45.59	48.16	35.69	37.61	43.25	68.6	41.7	49.38	48.76	**34.8**
Sudanese (apd)	74.82	63.97	49.1	45.71	44.12	70.9	49.19	47.26	48.71	40.06	45.49	37.68	41.33	**36.01**
Algerian (arq)	96.19	88.26	81.65	78.64	**78.01**	175.34	116.92	124.2	109.76	82.56	83.6	79.25	78.92	86.29
Hassaniyya (mey)	96.57	91.79	89.19	88.55	88.45	204.97	135.94	93.46	106.15	83.52	88.24	84.3	84.58	**83.17**
Moroccan (ary)	113.83	84.48	79.78	**79.71**	81.77	164.21	120.43	86.2	122.79	99.33	94.68	98.85	93.19	103.8
Tunisian (aeb)	92.46	86.38	78.37	80.51	79.33	225.04	111.93	80.63	85.32	247.39	80.62	84.17	77.55	**74.31**

Table G.1: A comprehensive comparison of WER performance across Arabic dialect varieties for 14 models spanning three distinct ASR architectural paradigms on TEST dataset. **Encode-Only:** **E1.** MMS-1B-ALL, **E2.** omniASR-CTC-300M-v2, **E3.** omniASR-CTC-1B-v2, **E4.** omniASR-CTC-3B-v2, and **E5.** omniASR-CTC-7B-v2. **Encode-Decoder:** **ED1.** Whisper-Large-v3-Turbo, **ED2.** Whisper-Large-v3, and **ED3.** SeamlessM4T-v2-Large. **Multimodal:** **M1.** Voxtral-Small-24B-2507, **M2.** Qwen3-Omni-30B-A3B-Instruct, **M3.** omniASR-LLM-300M-v2, **M4.** omniASR-LLM-1B-v2, **M5.** omniASR-LLM-3B-v2, and **M6.** omniASR-LLM-7B-v2. **Bold** refers to the best performance for each dialect. *Khaleeji (afb) and Levantine (apc) include varieties from multiple countries. *For CER performance, details are provided in Table G.2 (§ G).*

Dialect	Encode-Only (CTC / Acoustic Encoder)					Encode-Decoder			Multimodal (Speech + LLM Core)					
	E1	E2	E3	E4	E5	ED1	ED2	ED3	M1	M2	M3	M4	M5	M6
MSA (arb)	11.12	8.24	5.64	4.66	5.11	12.35	9.38	7.04	6.64	**4.18**	5.26	4.43	4.29	4.2
Khaleeji (afb)*	49.85	45.03	42.93	43.3	40.94	344.06	176.63	**33.42**	59.15	63.7	134.19	59.73	80.09	46.23
Khaleeji (Kuwait) (afb-kwt)	27.97	21.45	20.34	18.39	17.55	118.03	11.74	11.91	13.95	**4.79**	13.09	11.17	8.51	8.77
Khaleeji (UAE) (afb-are)	37	26.76	20.76	19.3	21.45	40.36	33.09	22.6	28.34	16.6	21.3	17.56	20.41	**15.7**
Hijazi (Saudi Arabia) (acw)	43.13	37.85	31.29	**29.67**	32.27	203.55	133.56	41.81	61.28	34.98	43.87	52.56	44.51	35.43
Najdi (Saudi Arabia) (ars)	46.25	42.98	35.81	32.51	33.59	213.69	148.74	35.73	63.07	**26.73**	54.22	44.12	38.91	44.2
Sanaani Arabic (Yemen) (ayn)	40.01	33.54	28.84	30.73	31.77	421.29	1071.03	32.35	584.01	**20.56**	79.7	55.51	50.19	96.5
Ta'izzi-Adeni Arabic (Yemen) (acq)	31.87	25.71	**22.2**	22.4	23.26	185.68	216.12	24.08	64.58	32.23	50.56	50.34	41.98	80.95
Mesopotamian Arabic (Iraq) (acm)	67.35	65.13	**62.33**	66.98	64.05	479.24	230.88	100.67	69.87	70.19	180.22	96.86	83.67	136.4
North Mesopotamian Arabic (Iraq) (ayp)	60.59	60.03	58.99	61.9	57.16	546.65	296.25	61.76	113.44	**53.94**	81.48	105.31	94.37	104.4
Levantine (apc)*	19.14	14.28	11.27	10.19	10.28	21.26	17.07	13	10.88	**8.99**	11.95	9.78	9.9	9.3
Levantine (Jordan) (apc-jor)	27.66	18.53	13.16	11.75	12.1	23.96	21.41	11.79	13.69	**8.89**	12.44	9.88	10.17	9.26
Levantine (Lebanon) (apc-lbn)	19.69	13.24	8.42	7.31	7.43	9.2	7.57	9.82	9.48	7.34	8.56	7.76	7.55	**5.69**
Levantine (Palestine) (apc-pse)	34.26	26.45	20.89	19.55	20.3	42.53	32.05	19.93	22.85	17.86	25.32	18.09	18.61	**15.81**
Levantine (Syria) (apc-syr)	10.16	7.1	4.95	4.86	4.75	11.05	5.12	6.18	5.97	**3.82**	4.08	4.38	4.65	5.38
Levantine (Syrian Damascus) (apc-syr-d)	15.77	14.11	13.93	13.93	13.89	13.54	13.77	13.35	13.71	13.48	**13.17**	13.24	13.43	13.57
Egyptian (arz)	33.9	25.43	20.58	**19.38**	22.42	46.96	31.61	23.88	23.28	43.07	30.1	43.96	40.32	27.72
Sudanese (apd)	34.23	27.99	21.45	19.18	19.87	51.6	25.05	24.41	23.69	1630.3	20.18	15.17	17.76	**14.94**
Algerian (arq)	48.5	43.51	38.99	**37.22**	39.16	139.05	112.26	69.75	79.5	632.82	51.55	37.43	45.21	44.96
Hassaniyya (mey)	54.26	49.52	49.79	50.2	51.84	161.86	109.66	54.03	71.11	44.39	46.84	**42.21**	42.46	43.85
Tunisian (aeb)	46.91	46.03	40.49	42.99	43.8	246.38	150.85	45.25	47.16	1244.96	48.03	52.23	47.3	**39.27**
Moroccan (ary)	69.71	**45.64**	46.43	47.22	52.7	138.48	85.35	51.93	87.56	63.29	60.89	66.58	61.64	85.6

Table G.2: A comprehensive comparison of CER performance across Arabic dialect varieties for 14 models spanning three distinct ASR architectural paradigms on TEST dataset. **Encode-Only:** **E1.** MMS-1B-ALL, **E2.** omniASR-CTC-300M-v2, **E3.** omniASR-CTC-1B-v2, **E4.** omniASR-CTC-3B-v2, and **E5.** omniASR-CTC-7B-v2. **Encode-Decoder:** **ED1.** Whisper-Large-v3-Turbo, **ED2.** Whisper-Large-v3, and **ED3.** SeamlessM4T-v2-Large. **Multimodal:** **M1.** Voxtral-Small-24B-2507, **M2.** Qwen3-Omni-30B-A3B-Instruct, **M3.** omniASR-LLM-300M-v2, **M4.** omniASR-LLM-1B-v2, **M5.** omniASR-LLM-3B-v2, and **M6.** omniASR-LLM-7B-v2. **Bold** refers to the best performance for each dialect. *Khaleeji (afb) and Levantine (apc) include varieties from multiple countries.

Dialect	Encode-Only (CTC / Acoustic Encoder)					Encode-Decoder			Multimodal (Speech + LLM Core)					
	E1	E2	E3	E4	E5	ED1	ED2	ED3	M1	M2	M3	M4	M5	M6
MSA (arb)	40.31	25.85	16.6	14.17	14.08	26.15	17.86	13.07	17.07	16.52	13.22	11.48	11.33	**10.13**
Khaleeji (afb)*	89.09	79.41	70.57	68.74	66.81	150.95	86.83	62.98	72.49	**52.53**	128.76	86.49	70.22	62.97
Khaleeji (Kuwait) (afb-kwt)	83.55	74.97	64.79	56.75	57.39	44.33	42.83	36.55	54.32	**33.74**	53.4	37.46	38.79	34.83
Khaleeji (UAE) (afb-are)	82.29	68.24	55.97	52.4	53.86	70.55	52.77	50.49	55.49	44.7	52.97	46.68	47.1	**43.1**
Najdi (Saudi Arabia) (ars)	78.63	72.12	57.95	55.11	53.78	124.95	90.84	58.51	60.37	**40.58**	69.34	65.38	59.65	46.51
Hijazi (Saudi Arabia) (acw)	85.31	76.52	63.79	59.9	58.49	169.05	82.93	57.85	81.76	**48.02**	74.38	68.81	57	51.47
Sanaani Arabic (Yemen) (ayn)	73.08	59.38	52.59	51.52	49.86	166.52	74.14	44.82	49.34	**33.94**	51.35	54.62	45.82	39.38
Ta'izzi-Adeni Arabic (Yemen) (acq)	70.12	53.66	44.97	43.64	41.09	106.86	71.94	35.75	47.04	**27.23**	42.54	44.38	38.88	33.1
North Mesopotamian Arabic (Iraq) (ayp)	96.21	90.71	88.26	87.66	86.77	325.71	195.41	85.67	100.19	**69.45**	81.45	88.88	95.12	98.45
Mesopotamian Arabic (Iraq) (acm)	83.85	75.27	68.37	67.74	67.46	255.35	139.37	61.58	74.09	**55.21**	63.68	77.97	73.88	58.2
Levantine (apc)	55.54	38.09	28.08	26.14	26.91	29.24	26.53	34.23	28.79	**22.46**	25.42	23.71	23.82	24.16
Levantine (Jordan) (apc-jor)	69.22	51.68	38.36	35.94	35.7	39.65	35.81	33.88	36	**26.79**	36.07	32.44	33.12	28.74
Levantine (Lebanon) (apc-lbn)	65.13	42.44	29.95	26.38	28.43	29.78	25.96	60.9	31.31	**21.73**	26.2	23.42	23.21	33.11
Levantine (Palestine) (apc-pse)	81.09	69.02	56.57	52.31	54.94	61.39	50.93	49.88	52.52	45.78	54.21	49.36	48.21	**44.78**
Levantine (Syria) (apc-syr)	35.83	19.27	11.86	11.52	10.57	13.73	10.56	10.2	12.84	10.2	8.44	**7.59**	8.38	12.5
Egyptian (arz)	73.39	54.89	41.65	38.25	38.92	43.38	52.88	31.25	44.54	66.03	37	32.77	31.47	**28.84**
Algerian (arq)	93.92	86.89	80.81	79.95	79.11	138.36	111.81	107.73	90.22	79.77	86.15	**73.7**	75.22	79.43
Hassaniyya (mey)	99.79	94.27	91.66	91.22	90.62	203.92	165.2	96.01	100.48	88.59	92.21	90.99	88.77	**86.97**
Moroccan (ary)	111.94	88.36	85.3	**85.01**	85.96	159.77	136.55	123.77	109.74	88.73	104.17	108.49	95.71	107.28
Tunisian (aeb)	60.19	53.76	34.93	27.81	27.77	25.33	19.79	**17.35**	25.54	20.27	29.48	21.29	20.77	23.24

Table G.3: A comprehensive comparison of WER performance across Arabic dialect varieties for 14 models spanning three distinct ASR architectural paradigms on the development (DEV) dataset. **Encode-Only:** **E1.** MMS-1B-ALL, **E2.** omniASR-CTC-300M-v2, **E3.** omniASR-CTC-1B-v2, **E4.** omniASR-CTC-3B-v2, and **E5.** omniASR-CTC-7B-v2. **Encode-Decoder:** **ED1.** Whisper-Large-v3-Turbo, **ED2.** Whisper-Large-v3, and **ED3.** SeamlessM4T-v2-Large. **Multimodal:** **M1.** Voxtral-Small-24B-2507, **M2.** Qwen3-Omni-30B-A3B-Instruct, **M3.** omniASR-LLM-300M-v2, **M4.** omniASR-LLM-1B-v2, **M5.** omniASR-LLM-3B-v2, and **M6.** omniASR-LLM-7B-v2. **Bold** refers to the best performance for each dialect. *Khaleeji (afb) and Levantine (apc) include varieties from multiple countries.

Model	M	F	Gap %
omniASR CTC 7B v2	53	60	7
omniASR LLM 7B v2	48	51	**3**
mms-1b-all	75	81	6
Qwen3-Omni-30B-A3B-Instruct	45	48	**3**
seamless-m4t-v2-large	54	59	5
whisper-large-v3	87	91	4

Table G.4: WER (%) on Arab Voices DEV broken down by gender. Absolute % gap provided between male and female performance.

Model	MSA	LOW	MIXED	MOST
omniASR CTC 7B v2	30	56	65	74
omniASR LLM 7B v2	36	61	56	**63**
mms-1b-all	59	78	86	92
Qwen3-Omni-30B-A3B-Instruct	**26**	**49**	**55**	69
seamless-m4t-v2-large	40	63	66	73
whisper-large-v3	61	92	88	117

Table G.5: WER (%) on Arab Voices DEV broken down by dialectness.

Model	>4	3.5-4	3-3.5	2.5-3	2-3.5	1.5-2
omniASR CTC 7B v2	28	33	43	49	55	75
omniASR LLM 7B v2	24	27	37	40	48	85
mms-1b-all	58	61	69	73	78	95
Qwen3-Omni-30B-A3B-Instruct	21	27	34	38	44	83
seamless-m4t-v2-large	43	33	44	47	52	96
whisper-large-v3	28	41	62	67	84	153

Table G.6: WER on Arab Voices DEV broken down by PESQ quality bin.

Model	>20	15-20	10-15	5-10	0-5	<0
omniASR CTC 7B v2	43	57	62	68	73	88
omniASR LLM 7B v2	35	48	56	58	71	143
mms-1b-all	69	79	82	87	93	109
Qwen3-Omni-30B-A3B-Instruct	32	47	54	61	68	129
seamless-m4t-v2-large	42	56	59	68	80	151
whisper-large-v3	60	96	92	96	152	199

Table G.7: WER on Arab Voices DEV broken down by SI-SDR quality bin.

Model	>.9	.8-.9	.7-.8	<.7
omniASR CTC 7B v2	50	73	79	87
omniASR LLM 7B v2	42	63	74	155
mms-1b-all	74	91	94	112
Qwen3-Omni-30B-A3B-Instruct	40	64	72	138
seamless-m4t-v2-large	48	72	85	163
whisper-large-v3	73	119	155	215

Table G.8: WER on Arab Voices DEV broken down by STOI quality bin.

Model	>4	3.5-4	3-3.5	2.5-3	2-3.5	1.5-2
omniASR CTC 7B v2	50	61	50	49	61	50
omniASR LLM 7B v2	42	61	45	48	54	49
mms-1b-all	73	83	76	74	81	75
Qwen3-Omni-30B-A3B-Instruct	41	57	43	43	52	49
seamless-m4t-v2-large	47	62	54	52	77	59
whisper-large-v3	73	107	73	73	95	56

Table G.9: WER on Arab Voices DEV broken down by subjective non-matching reference quality bin.