

Sada-ASR: A Production Multidialectal Arabic Speech Recognition System

Mahmoud Khaled and Mohammed Abd-Elnaser and Mahmoud Mohsen

EchoNova AI, Dubai, United Arab Emirates

{mahmoud, mabdelnasser, mmohsen}@echonova.ai



Code



Dataset



echonova.ai

Abstract

Callers to Arabic customer-service lines speak their own dialect, while speech recognizers are built and evaluated mainly on Modern Standard Arabic (MSA). A recognizer that mishears the dialect breaks the call before anything else can help. We present Sada-ASR, the speech recognizer of EchoNova’s production Arabic voice agent. On a 120-hour test set of MSA and eight dialects, it has the lowest word and character error rates of twelve systems, against eleven from Microsoft, ElevenLabs, Google, OpenAI, Speechmatics, IFM, Cohere, NVIDIA and Audar: 11.69% word error rate against 13.51% for the strongest, Microsoft’s MAI-Transcribe-2 (a lead of 0.74 points on clips unaffected by that system’s fixed number and script conventions), and 15.52% against 16.96% averaged over varieties. Its largest margins are on Omani (4.30 points), Egyptian (3.78) and Saudi (2.19) speech; it is level on Syrian and loses on Lebanese (24.96% against 23.05%), mostly because it writes Lebanese words in MSA forms and is less verbatim. On six public test sets its average word error rate under the leaderboard’s current scoring is 17.17%, below every published figure we found, among them Falcon-ASR (20.92%), Jais-ASR (22.85%) and Audar-ASR-V1-Turbo (23.17%), though none of these states that it used that scoring, so the comparison is indicative. We describe the data pipeline and methodology behind Sada-ASR and release a 20-hour subset of the test set with our scoring code.

1 Introduction

Arabic is diglossic [1]: Modern Standard Arabic (MSA) is the language of writing, news and official documents, while everyday speech happens in regional dialects that differ from MSA and from each other in vocabulary, pronunciation and morphology. A caller to an Arabic customer-service line speaks their dialect, not MSA. Dialects have no standard orthography [2], and speakers mix in

English words [3]. Speech technology built and evaluated mainly on MSA meets this speech at its weakest point.

Recognition is where that mismatch does its damage first: a voice agent mishearing a caller’s dialect is wrong before the rest of the system has a chance to be right, and the caller cannot see the transcript to correct it.

We present Sada-ASR,¹ the speech recognizer of EchoNova Sada, a production Arabic voice agent for customer service, and evaluate it against eleven commercial and open-weight systems on a 120-hour private test set of MSA and dialects from the Gulf, Egypt and the Levant. Sada-ASR has the lowest word and character error rates, with its largest margins over the strongest external system, Microsoft’s MAI-Transcribe-2, on Omani, Egyptian and Saudi speech. It loses on Lebanese to MAI-Transcribe-2, and we analyse why. On six public test sets, under the leaderboard’s current scoring, its average error rate is also below every published figure we found, including those for Falcon-ASR and Jais-ASR; none of those figures states that it used this scoring, so the comparison is indicative.

Our contributions are:

- **The data pipeline and methodology behind Sada-ASR** (Sections 4 and 5).
- **A head-to-head comparison of twelve systems** on one private 120-hour multi-dialect test set, all measured by us (Section 6).
- **Per-variety results**, including the one dialect where Sada-ASR loses and an analysis of why it loses there (Sections 6.3 and 7).
- **A comparison on six public test sets** with published results for 24 systems (Section 6.4).
- **A release**: a 20-hour subset of the test set covering all nine varieties, with its reference transcripts, and our scoring code.

¹Sada (صدي) means ‘echo’. The name is unrelated to the SADA corpus [4], one of the public test sets we report on.

2 Related Work

Arabic speech recognition and its benchmarks.

The MGB challenges evaluated Arabic ASR first on largely MSA broadcast speech [5] and then on dialectal web video [6, 7], and QASR added about 2 000 hours of multi-dialect broadcast speech [8]. Hussein et al. [9] benchmark end-to-end and modular recognizers on the MGB-2, MGB-3 and MGB-5 test sets and compare them with human transcribers on a new one, and Talafha et al. [10] benchmark Whisper across Arabic varieties. Crowd-sourced and web-scale corpora followed [11, 12], as did corpora for single regions such as the SADA corpus [4], multi-dialect collections such as Casablanca [13], and corpora of Arabic–English code-switching [3], which multilingual recognizers also target [14]. The Open Universal Arabic ASR Leaderboard now evaluates six of these test sets [15], and NADI 2025 added a multidialectal ASR track [16]. Among industrial systems, Salhab et al. [17] train on automatically labelled speech and report leaderboard results against open-weight models and commercial services, and Salhab et al. [18] extend the approach to the NADI dialects; multilingual models such as Whisper [19] serve as common baselines. Our evaluation differs in what it compares and how: current commercial APIs and open-weight models measured by us on one private test set, with results for each variety.

Orthography and scoring. Dialectal Arabic has no standard orthography; CODA and its unified successor propose conventions for annotating it [2, 20]. Single-reference WER therefore penalizes valid spelling variants, which Ali et al. [21] address by scoring against multiple references.

3 Deployment Context

Sada-ASR is deployed inside EchoNova Sada, a cascaded voice agent for customer service: a speech recognizer, a hosted large language model (LLM) with tools, and a text-to-speech (TTS) model, connected through a real-time media server (Figure 1). A cascade was chosen over an end-to-end speech model because each speech stage can then be specialized to Arabic dialects on its own, the LLM can be replaced without retraining either speech model, and every intermediate result is text that can be checked before it is spoken.

The TTS model produces 24 kHz speech and was trained on 20 000 hours of speech. It provides

voices in Emirati, Kuwaiti, Egyptian, Saudi and Modern Standard Arabic, and in English; each agent is configured with a subset of them. The agent and its TTS model are the setting Sada-ASR is built for; neither is evaluated in this report.

4 Data Pipeline

Verbatim transcripts of dialectal Arabic are scarce. Each dialect needs annotators who speak it, and dialects have no standard spelling to transcribe against. Dialectal audio is easier to find than its transcripts, but the labels that come with web audio are weak. Sada-ASR is therefore trained on two streams of data, merged into one training set (Figure 2): audio transcribed on our own annotation platform, and weakly labelled audio collected from the web. The corpus also contains English speech, about a third of all training hours.

4.1 In-House Annotation Platform

We built an annotation platform with three layers. More than 100 in-house annotators from across the Arab world transcribe the audio. Reviewers continuously review and audit the annotators’ submissions. Experts, who are professional writers and domain specialists, handle the hardest cases.

The platform measures annotators and reviewers with control samples. The same evaluation samples are placed in every annotator’s queue, which gives each annotator’s performance a common baseline. Duplicate samples are inserted without the annotators’ knowledge, which measures how consistently different annotators transcribe the same audio. All control samples are removed before the dataset is compiled.

4.2 Weakly Labelled Web Data

To add volume and real-world variety, we collected thousands of hours of weakly labelled audio from the web.

4.3 Merging and Filtering

The two streams differ in label quality and in format, so they are merged in three steps:

- **Deduplication.** Repeated audio is removed across the two streams.
- **Cleanup.** Corrupted records, mismatches between audio and label, and audio artifacts are filtered out.

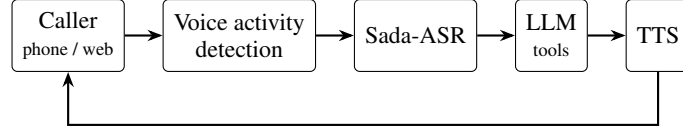


Figure 1: Where Sada-ASR sits in the voice agent at call time.

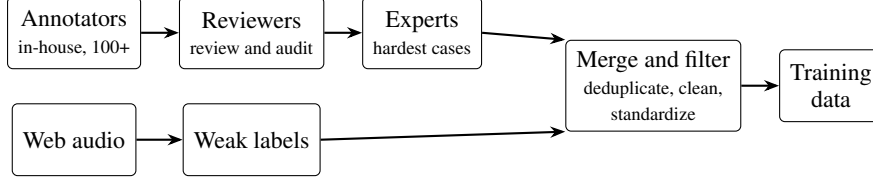


Figure 2: The two data streams behind Sada-ASR: audio transcribed on our annotation platform, and weakly labelled web audio, merged into one training set.

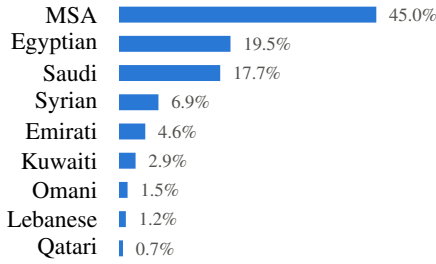


Figure 3: Relative share of Sada-ASR’s dialect-labelled Arabic training audio across the nine varieties of our test set.

- **Standardization.** Audio formats and meta-data are made uniform across both sources, which yields one training set.

Figure 3 shows how the training audio divides across the varieties of our test set. Within these nine varieties, MSA is just under half of the training audio, Egyptian about a fifth and Saudi about a sixth; Omani, Lebanese and Qatari are each under 2%.

5 Methodology

5.1 Architecture

Sada-ASR is an attention-based encoder–decoder recognizer that writes the transcript token by token.

5.2 Training

We train Sada-ASR on the corpus of Section 4. Each epoch draws half of its clips from English, repeating English audio and subsampling Arabic; within Arabic, clips are drawn in proportion to each variety’s hours (Figure 3). We apply no data augmentation: no added noise, speed perturbation, codec simulation, reverberation or SpecAugment. Audio is only resampled to 16 kHz mono.

The objective is next-token cross-entropy on the transcript and its end-of-sequence token, with no label smoothing and no auxiliary CTC loss. Arabic targets are undiacritized, keep the hamza forms of alif, write numbers as words, and keep code-switched words in Latin script, lowercased; English targets are cased and keep digits. Punctuation and casing in the targets come from an automatic punctuation-restoration model.

We select checkpoints by word error rate on a held-out validation subset of 1 482 utterances (at most 100 per variety, English included), decoded with plain greedy search. The subset is disjoint from the training clips and from every test set we report.

5.3 Decoding

Every test set, private and public, is decoded with the same configuration: plain greedy search, with no beam search, repetition penalty or n-gram constraint, and no decode-time tuning. A guard stops repetition loops, allowing at most three repeats of any 4- to 8-token unit, and each window’s output is capped at 14 tokens per second of audio plus 16 (at least 20). Audio longer than 30 s is decoded in 30 s windows with 2 s overlap and the window texts are joined. Clips are decoded one at a time.

6 Speech Recognition Evaluation

We compare Sada-ASR with eleven external recognizers—five commercial APIs and six self-hosted open-weight models—on a private test set (Sections 6.1–6.3). The test set was unreleased when the systems were measured, the APIs of Microsoft, Google and OpenAI do not train on submitted audio by default, and the open-weight models

Variety	Test set			Released	
	Hours	Clips	Mean (s)	Hours	Clips
MSA	49.8	20 452	8.8	8.27	3 400
Egyptian	20.2	9 049	8.0	3.37	1 508
Emirati	15.0	7 119	7.6	2.50	1 187
Saudi	10.0	5 492	6.6	1.67	916
Kuwaiti	5.0	2 535	7.1	0.84	423
Qatari	5.0	2 378	7.6	0.83	396
Omani	4.8	2 226	7.8	0.81	371
Syrian	5.0	2 166	8.3	0.83	361
Lebanese	5.0	2 180	8.3	0.83	363
Total	119.9	53 597	8.1	19.95	8 925

Table 1: Composition of the private test set and of its released subset. Mean is the mean clip length; totals are computed before rounding.

ran on our own hardware. No clip of the private test set, and no part of the six public test splits, was used to train or select Sada-ASR. Six public test sets connect our numbers to published results for 24 systems (Section 6.4).

6.1 Private Multi-Dialect Test Set

The test set contains 120 hours of speech in 53 597 clips covering nine varieties: Modern Standard Arabic (MSA) and eight regional dialects (Table 1). It is stratified by duration—each variety contributes a fixed number of hours—so clip counts vary with speaking rate and segmentation. Every clip is at most 30 s long, so no system needed long-form segmentation and all received identical audio. The clips are cut from 14 686 source recordings; clips from one recording share speaker and channel, so our confidence intervals resample recordings.

MSA is 41.6% of the audio and our aggregate is a corpus-level micro-average, so the overall figure is dominated by MSA: a system that is strong on MSA and weak on dialects still looks good in aggregate. We therefore treat the per-variety results as primary.

6.2 Systems, Configuration and Scoring

Table 2 lists the systems. All systems were measured on 13 October 2026, except Jais-ASR and Speechmatics, measured on 15–16 October. Every API was set to Arabic, and to verbatim output with formatting options disabled wherever the API exposes them; no system—including ours—used contextual biasing, custom vocabulary or any other adaptation. Microsoft’s service was given the locale `ar`; its API reports the locale `ar` whatever is sent, and two dialect locales gave identical

transcripts on 48 clips from six varieties. We selected Microsoft’s model explicitly, ran ElevenLabs Scribe v2 with audio-event tagging and diarization off, and ran Speechmatics with its enhanced model, which keeps disfluencies and writes numbers as digits. OpenAI’s and Google’s models were reached through OpenRouter at temperature 0; Google’s verbatim mode, which OpenRouter does not expose, is its default. The open-weight models ran at their published or default settings.²

WER and CER are corpus micro-averages: total edit distance divided by total reference length, in words or characters. On the private test set, hypotheses and references are normalized identically: punctuation and diacritics are removed; the hamza and madda forms of alif are folded into bare alif, hamza on waw and ya is folded into the bare letter, and standalone hamza is deleted; the extended letters for /p/ and /v/ are mapped to **ب** and **ف**; alif maqsurah is not folded into ya; Eastern Arabic digits are mapped to Western digits; and **ة** is folded into **ه**, a common Arabic normalization step [22]. Numbers are not converted between digits and words. For every system, any one- to three-word unit repeated more than ten times in a row is cut to ten copies (no reference repeats one more than seven times), and whitespace is canonicalized. The crop changes four systems’ WER (Gemini from 19.95% to 16.13%, Whisper from 28.94% to 26.57%, Jais-ASR from 16.53% to 16.41% and Cohere Transcribe from 29.12% to 29.11%). An API output with more than five times the reference’s word count and more than 50 words was requested again, up to three times, and the first normal output kept. Whisper, Audar, Nemotron and Jais-ASR were decoded in batches, which is not bit-identical to decoding one clip at a time; we did not measure the effect. For every system and clip we store the raw output before normalization, so the reported error rates can be recomputed without re-running a recognizer.

²Open-weight models by Hugging Face ID: IFM/Jais-ASR, audarai/Audar-ASR-V1-Turbo, nvidia/stt_ar_fastconformer_hybrid_large_pc_v1.0, openai/whisper-large-v3, nvidia/nemotron-3.5-asr-streaming-0.6b and CohereLabs/cohere-transcribe-arabic-07-2026. Whisper was decoded greedily with Arabic forced, Nemotron in streaming mode at its 1120ms setting, FastConformer with its default decoder, Jais-ASR with its model card’s decoding (greedy, repetition penalty 1.05) and Arabic set, Audar with Arabic set, and Cohere Transcribe greedily with Arabic set.

System	Type	WER	Macro	CER	Empty
Sada-ASR (ours)	self-hosted	11.69	15.52	4.76	0.00
Microsoft MAI-Transcribe-2	API	13.51	16.96	6.37	0.00
ElevenLabs Scribe v2	API	16.12	20.67	7.54	0.08
Google Gemini 3.5 Transcribe	API	16.13	20.84	8.23	0.02
OpenAI gpt-transcribe	API	17.03	21.69	8.59	0.01
Speechmatics Enhanced	API	18.08	24.37	8.73	0.00
IFM Jais-ASR	self-hosted	16.41	21.00	6.89	0.00
Audar-ASR-V1-Turbo	self-hosted	20.86	25.17	9.97	0.00
NVIDIA FastConformer (Arabic)	self-hosted	25.97	32.93	12.31	0.01
OpenAI Whisper large-v3 [19]	self-hosted	26.57	32.56	15.08	0.00
NVIDIA Nemotron-3.5 streaming 0.6B	self-hosted	28.11	36.35	11.43	0.10
Cohere Transcribe Arabic	self-hosted	29.11	37.94	19.41	0.05

Table 2: Private test set. WER and CER in %, corpus micro-averaged; Macro is the unweighted mean of the nine per-variety WERs in Table 3; Empty is the % of clips with an empty transcript.

6.3 Results

Overall. Sada-ASR has the lowest WER (11.69%) and CER (4.76%) of the twelve systems (Table 2). Against the strongest external system, MAI-Transcribe-2 (13.51%), this is a 13.5% relative WER reduction: a difference of 1.83 points (95% CI 1.72–1.93, bootstrap over source recordings). Averaged over the nine varieties without weighting by duration, the WERs are 15.52% and 16.96%. MAI-Transcribe-2’s API writes numbers as digits and English words in Latin script, and neither can be configured off; 9.23% of its clips contain digits where the reference has none, against none of Sada-ASR’s. On the clips with no digit, number word or Latin script in the reference or in either system’s output (the controlled clips), the margin is 0.74 points (95% CI 0.65–0.83), and 0.49 when hesitation fillers are also removed.

By variety. Sada-ASR has the lowest WER on seven of the nine varieties (Table 3). On Syrian it is level with MAI-Transcribe-2 (20.16% against 19.98%; the CI of the difference includes zero), and on Lebanese it is behind by 1.91 points, a loss analysed in Section 7. Its margins over MAI-Transcribe-2 are largest on Omani (4.30 points, 21.1% relative), Egyptian (3.78) and Saudi (2.19), followed by Emirati (1.65), MSA (1.38) and Qatari (1.20); on Kuwaiti the margin is a narrow 0.53. On the controlled clips, every margin keeps its sign: the largest are Omani (4.55), Egyptian (1.70) and Saudi (1.57), Syrian stays level, and the Lebanese loss widens to 3.27 points (Δ ctrl., Table 3). By CER, Sada-ASR is the lowest of the twelve systems on all nine varieties, Lebanese included; on Syrian the CI of its CER difference from MAI-Transcribe-2 includes zero (+0.03, 95% CI −0.41 to 0.45).

We do not claim that Sada-ASR degrades less than other systems from MSA to dialect: the unweighted mean of its eight dialect WERs is 2.99 times its MSA WER, against 2.61 for MAI-Transcribe-2.

Empty and runaway output. An empty transcript is scored as a full deletion, so it is already inside WER; we report it separately because it is a distinct failure in a live call. Sada-ASR returned an empty transcript for 2 of the 53 597 clips; among the other systems the rate reaches 0.10%. None of its transcripts is a runaway output—more than five times the reference’s word count and more than 50 words—against 131 of Whisper’s.

6.4 Public Test Sets

Six public test sets from the Open Universal Arabic ASR Leaderboard [15] connect our results to published ones: the SADA corpus [4], Common Voice 18.0 [11], the clean and noisy test sets of MASC [12], MGB-2 [5] and Casablanca [13]. Table 4 compares Sada-ASR with the published figures for these sets: fourteen systems reported by Salhab et al. [17], among them their own model and three commercial services; eight reported by IFM [25], among them its own Jais-ASR and the commercial Speechmatics Melia 1 and ElevenLabs Scribe v2; Falcon-ASR [23]; and Audar-ASR-V1-Turbo, the leaderboard’s best entry on 10 October 2026 [24]. We score the six sets with the leaderboard’s own evaluation code; Sada-ASR was measured on 13–14 October 2026. The leaderboard revised its scoring on 27 July 2026, fixing punctuation removal, attaching a standalone َ to the next word and collapsing extra spaces, and its earlier rows were not re-scored. Under the current scoring, Sada-ASR’s WER averaged over the six sets is 17.17% over 46 757 clips, against 20.92% reported

Variety	API						Self-hosted						Δ MAI-2 (95% CI)	Δ ctrl.
	Sada-ASR	MAI-2	Scribe v2	Gemini	gpt-tr.	Speechm.	Jais	Audar	FastConf.	Whisper	Nemotron	Cohere		
MSA	5.61	6.99	7.80	8.03	7.63	7.82	7.83	11.04	12.05	14.66	12.35	15.33	+1.38 [1.25, 1.51]	+0.23
Egyptian	9.48	13.26	15.19	14.32	17.78	17.56	16.55	25.45	28.65	28.11	31.57	25.33	+3.78 [3.59, 3.98]	+1.70
Saudi	14.74	16.93	21.36	20.68	21.18	20.94	21.06	23.95	31.72	31.11	34.40	31.32	+2.19 [1.81, 2.63]	+1.57
Emirati	18.57	20.22	24.40	24.85	26.30	26.78	23.86	26.86	37.22	37.24	37.95	45.31	+1.65 [1.37, 1.91]	+0.77
Kuwaiti	14.29	14.83	18.24	19.21	19.55	33.84	19.84	23.22	35.48	31.02	35.71	41.54	+0.53 [0.17, 0.87]	+0.58
Qatari	15.82	17.02	21.72	20.73	22.15	22.32	20.32	22.96	35.33	31.95	33.89	44.02	+1.20 [0.48, 1.95]	+0.89
Omani	16.07	20.37	23.05	22.35	23.24	22.81	22.65	26.45	34.55	33.93	36.23	40.09	+4.30 [3.82, 4.80]	+4.55
Syrian	20.16	19.98	24.90	25.70	26.93	32.89	26.54	31.63	38.55	39.60	49.74	41.85	-0.17 [†] [-0.72, 0.40]	-0.46
Lebanese	24.96	23.05	29.37	31.73	30.43	34.37	30.35	34.93	42.84	45.41	55.32	56.67	-1.91 [-2.47, -1.35]	-3.27

Table 3: WER (%) by variety on the private test set. MAI-2: Microsoft MAI-Transcribe-2; Scribe v2: ElevenLabs Scribe v2; Gemini: Google Gemini 3.5 Transcribe; gpt-tr.: OpenAI gpt-transcribe; Speechm.: Speechmatics Enhanced; Jais: IFM Jais-ASR; FastConf.: NVIDIA FastConformer (Arabic); Nemotron: NVIDIA Nemotron-3.5 streaming 0.6B; Cohere: Cohere Transcribe Arabic. Δ MAI-2: Sada-ASR’s margin over MAI-Transcribe-2 in points (positive = Sada-ASR lower), with a 95% CI from 2 000 bootstrap resamples of source recordings; [†]the only CI including zero. Δ ctrl.: the same margin on clips with no digit, number word or Latin script in the reference or in either system’s output. Differences are computed before rounding. CER: Appendix A.1.

for Falcon-ASR, 22.85% for Jais-ASR, 23.17% for Audar-ASR-V1-Turbo on the leaderboard, 23.51% for Speechmatics Melia 1 and 24.54% for ElevenLabs Scribe v2 as measured by IFM, and 34.74% for the strongest open-weight model and 40.05% for the strongest of the three commercial services in Salhab et al. [17]. None of these figures states that it used the current scoring, and how they treated punctuation is not stated, so the comparison is indicative.

7 Error Analysis: Where Sada-ASR Loses

Lebanese is a clear loss (Table 3), and a broad one: Sada-ASR makes more errors than MAI-Transcribe-2 on 55 of the 71 Lebanese source recordings, and the gap holds across clip lengths and speaking rates. The two systems make almost the same share of substitutions; Sada-ASR deletes more words and MAI-Transcribe-2 inserts more (Table 5). A word-by-word comparison with the reference shows two habits of Sada-ASR.

	Lebanese		Syrian	
	Sada-ASR	MAI-2	Sada-ASR	MAI-2
Substitutions	16.71	16.56	15.06	15.36
Deletions	7.06	4.45	3.61	2.29
Insertions	1.19	2.04	1.49	2.34
WER	24.96	23.05	20.16	19.98
CER	10.86	11.76	7.79	7.81

Table 5: Levantine errors (%). Substitutions, deletions and insertions are shares of the reference words and sum to WER before rounding.

First, Sada-ASR writes Lebanese words in MSA

forms. The references spell words shared with MSA as they are pronounced, as in **كثير** ‘a lot’ and **اكثر** ‘more’; Sada-ASR writes them with their MSA consonants, **كثير** and **اكثر**. Net of MAI-Transcribe-2’s errors of the same kind, this costs 1.18 points, 0.88 of them from the **كثير** family alone (**كثير**, **اكثر**, **بكثر**). Sada-ASR also replaces Lebanese function words with their MSA counterparts, as in **هيدا** → **هذا** ‘this’ and **نحن** → **نحن** ‘we’. It also drops the **ي** of the verb prefix, as in **بيصير** → **بصير**, which is not an MSA form but changes the person of the verb, from ‘he becomes’ to ‘I become’. The MSA forms are its own (Table 6): MAI-Transcribe-2 almost never writes **كثير** as **كثير**. Others, such as **مثل** → **مثل** ‘like’ and **انت** → **انت** ‘you’, are shared conventions that cost both systems equally and do not move the gap. As a diagnostic, counting **ت** and **ث**, **د** and **ذ**, **ظ** and **ز**, and **ا** and **ق** as equal for both systems lowers the gap from 1.91 to 0.51 points (95% CI -0.11 to 1.12), which is no longer significant.

Second, Sada-ASR is less verbatim. It omits false starts, repeated words and short words that the references keep, such as **اي** ‘yes’ and **لا** ‘no’, which costs 1.35 points net. MAI-Transcribe-2 gives back 1.46 points through more insertions, among them hesitation fillers where the reference has none, and through other substitutions, such as English words written in Latin script. Fillers do not explain the loss: both systems pay for them, since Sada-ASR drops them and MAI-Transcribe-2 often writes them with a different number of letters, and removing them from both sides widens the gap to 2.50 points. Substitutions of word form cost a

System	SADA	CV 18.0	MASC-c	MASC-n	Casab.	MGB-2	Avg
Sada-ASR (ours)	25.76	3.95	5.35	16.70	41.72	9.56	17.17
Falcon-ASR [23]	—	—	—	—	—	—	20.92
Audar-ASR-V1-Turbo, leaderboard [24]	28.92	8.09	16.73	27.19	47.02	11.08	23.17
Munsit [17]	27.71	10.42	21.74	28.08	60.04	12.10	26.68
<i>As reported by IFM [25]</i>							
Jais-ASR	30.85	8.56	16.12	25.44	46.56	9.59	22.85
Speechmatics Melia 1	34.74	5.19	16.23	25.89	46.41	12.61	23.51
ElevenLabs Scribe v2	33.96	8.71	16.91	27.15	50.20	10.30	24.54
Cohere Transcribe Arabic	37.47	5.82	19.60	27.07	49.71	15.54	25.87
omniASR-LLM-7B	41.61	8.75	19.69	29.29	56.46	14.13	28.32
Qwen3-Omni-30B-A3B	44.82	11.46	21.47	30.85	62.55	13.09	30.71
Qwen3-ASR-1.7B	45.53	16.90	24.37	34.29	64.47	16.57	33.36*
Voxtral-Small-24B	50.82	15.25	23.96	34.43	66.30	16.03	34.47
<i>Commercial, as reported by Salhab et al. [17]</i>							
ElevenLabs Scribe	49.44	28.27	31.93	41.23	63.77	25.68	40.05
OpenAI GPT-4o-transcribe	66.47	28.19	31.53	43.29	70.72	29.62	44.97
Microsoft Azure STT [‡]	58.50	33.77	40.66	45.64	64.84	30.91	45.72
<i>Open-weight, as reported by Salhab et al. [17]</i>							
NVIDIA Conformer-CTC large	47.26	10.60	24.12	35.64	71.13	19.69	34.74
Whisper large-v3	55.96	17.83	24.66	34.63	71.81	16.26	36.86
SeamlessM4T large-v2 [26]	62.52	21.70	25.04	33.24	66.25	20.23	38.16
Whisper large-v3-turbo	60.36	25.73	25.51	37.16	73.79	17.75	40.05
Whisper large-v2	57.46	21.77	27.25	38.55	71.01	25.17	40.20
Whisper large	63.24	26.04	28.89	40.79	72.18	24.28	42.57
Whisper medium	67.71	28.07	29.99	42.91	75.44	29.32	45.57
Parakeet CTC 1.1B (concat)	70.70	26.34	30.49	45.95	80.80	24.94	46.54
Parakeet CTC 1.1B (universal)	73.58	40.01	36.16	50.03	81.30	30.68	51.96
Whisper small	78.02	24.18	35.93	56.36	87.64	48.64	55.13

Table 4: WER (%) on the six public test sets; Avg is their unweighted mean. Rows other than ours are published figures: Audar-ASR-V1-Turbo is the leaderboard’s best entry (10 October 2026), and Falcon-ASR reports only an average over the same six test lists. IFM measured Jais-ASR with the leaderboard’s evaluation script, without stating its version, measured Melia 1 and Scribe v2 itself, and lists the other five from the leaderboard. ElevenLabs Scribe, GPT-4o-transcribe and Azure STT in the Salhab rows are earlier models than those we measured. [‡]Mean of Azure’s Saudi and Jordanian models. *As published by IFM; its six per-set values average 33.69. Our row uses the leaderboard’s current scoring, revised on 27 July 2026; most published rows predate that revision, so the comparison is indicative (CER: Appendix A.2).

whole word in WER but little in CER, which helps explain why Sada-ASR still has the lower Lebanese CER.

The same habits appear on Syrian, more weakly (Table 5). MSA spellings of shared words cost 0.74 points there, and MSA function words 0.06, since words such as **هيدا** and **نحنا** are characteristically Lebanese. Fewer one-letter errors and fewer insertions than MAI-Transcribe-2 offset them, and the two systems are level on Syrian. We plan to address the Lebanese loss by training on more verbatim Levantine transcripts with a consistent dialect spelling.

8 Conclusion

We presented Sada-ASR, the Arabic recognizer of a production voice agent, with the data pipeline and methodology behind it, and evaluated it against

eleven external systems on a 120-hour test set of MSA and dialects from the Gulf, Egypt and the Levant. Sada-ASR has the lowest word and character error rates, with its largest margins on Omani, Egyptian and Saudi speech, and loses on Lebanese to MAI-Transcribe-2, mostly because it writes Lebanese words in MSA forms and is less verbatim. On six public test sets, under the leaderboard’s current scoring, its average word error rate is below every published figure we found, including those for Falcon-ASR and Jais-ASR, though none of those states that it used this scoring, so the comparison is indicative. The recognizer runs inside a production voice agent, which we describe as context rather than evaluate.

We plan to test output on non-speech input, measure turn latency, run a listening test of the TTS model and evaluate the full agent.

Reference → output	Sada-ASR	MAI-2
<i>Mostly Sada-ASR</i>		
كثير → كثير 'a lot'	265	0
هذا → هذا 'this'	126	28
نحن → نحن 'we'	121	24
أكثر → أكثر 'more'	96	1
<i>Shared</i>		
مثل → مثل 'like'	94	93
أنت → أنت 'you'	70	69
رح → رح 'will'	35	41
<i>Mostly MAI-Transcribe-2</i>		
سو → So 'so'	0	44
اممم → اممم (filler)	6	66

Table 6: Frequent Lebanese substitutions and how often each system makes them.

Limitations

The private test set covers MSA and eight dialects, and MSA is 41.6% of its audio; Maghrebi, Iraqi, Sudanese and Yemeni Arabic are not represented.

Ethics Statement

Third-party services. Evaluating hosted systems meant sending test audio to their providers’ APIs, directly or through OpenRouter.

Release. A 19.95-hour subset of the private test set, 8925 clips from all nine varieties (Table 1), is released with its reference transcripts at <https://huggingface.co/datasets/echonova-ai/arabic-dialect-asr-benchmark>. Its transcripts, segmentation and metadata are licensed under CC BY 4.0; the audio is for research and evaluation only, with no commercial use or redistribution, and access requires accepting these terms. Our normalization and scoring code is released at <https://github.com/echonova-ai-solutions/arabic-asr-benchmark>.

Conflict of interest. The authors are employed by the company that develops Sada-ASR, and we ran every system on the private test set ourselves. We mitigate this by reporting where Sada-ASR loses, dating every external measurement, storing raw outputs so the reported error rates can be re-computed, and releasing the scoring code.

References

- [1] Charles A. Ferguson. 1959. *Diglossia*. *Word*, 15(2):325–340.
- [2] Nizar Habash, Mona Diab, and Owen Rambow. 2012. *Conventional orthography for dialectal Arabic*. In

Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), pages 711–718, Istanbul, Turkey. European Language Resources Association.

- [3] 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, Marseille, France. European Language Resources Association.
- [4] Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, et al. 2024. *SADA: Saudi audio dataset for Arabic*. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290.
- [5] Ahmed Ali, Peter Bell, James Glass, et al. 2016. *The MGB-2 challenge: Arabic multi-dialect broadcast media recognition*. In *Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284.
- [6] Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. *Speech recognition challenge in the wild: Arabic MGB-3*. In *Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322.
- [7] Ahmed Ali, Suwon Shon, Younes Samih, et al. 2019. *The MGB-5 challenge: Recognition and dialect identification of dialectal Arabic speech*. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033.
- [8] Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. *QASR: QCRI Aljazeera speech resource – a large scale annotated Arabic speech corpus*. 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 2274–2285, Online. Association for Computational Linguistics.
- [9] Amir Hussein, Shinji Watanabe, and Ahmed Ali. 2022. *Arabic speech recognition by end-to-end, modular systems and human*. *Computer Speech & Language*, 71:101272.
- [10] Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. *N-shot benchmarking of Whisper on diverse Arabic speech recognition*. In *Proceedings of Interspeech 2023*, pages 5092–5096.
- [11] Rosana Ardila, Megan Branson, Kelly Davis, et al. 2020. *Common Voice: A massively-multilingual speech corpus*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- [12] Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2023. *MASC: Massive Arabic speech corpus*.

- In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1006–1013.
- [13] Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, et al. 2024. [Casablanca: Data and models for multidialectal Arabic speech recognition](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758, Miami, Florida, USA. Association for Computational Linguistics.
 - [14] Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. [Towards one model to rule all: Multilingual strategy for dialectal code-switching Arabic ASR](#). In *Proceedings of Interspeech 2021*, pages 2466–2470.
 - [15] Yingzhi Wang, Anas Alhmoud, and Muhammad Alqurishi. 2025. [Open Universal Arabic ASR Leaderboard](#). In *Proceedings of Interspeech 2025*, pages 1168–1172.
 - [16] Bashar Talafha, Hawau Olamide Toyin, Peter Sullivan, et al. 2025. [NADI 2025: The first multidialectal Arabic speech processing shared task](#). In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 720–733, Suzhou, China. Association for Computational Linguistics.
 - [17] Mahmoud Salhab, Marwan Elghitany, Shameed Sait, et al. 2025. [Advancing Arabic speech recognition through large-scale weakly supervised learning](#). *Preprint*, arXiv:2504.12254v2.
 - [18] Mahmoud Salhab, Shameed Sait, Mohammad Abusheikh, and Hasan Abusheikh. 2025. [Munsit at NADI 2025 shared task 2: Pushing the boundaries of multidialectal Arabic ASR with weakly supervised pretraining and continual supervised fine-tuning](#). In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 734–739, Suzhou, China. Association for Computational Linguistics.
 - [19] Alec Radford, Jong Wook Kim, Tao Xu, et al. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518.
 - [20] Nizar Habash, Fadhl Eryani, Salam Khalifa, et al. 2018. [Unified guidelines and resources for Arabic dialect orthography](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.
 - [21] Ahmed Ali, Walid Magdy, Peter Bell, and Steve Renals. 2015. [Multi-reference WER for evaluating ASR for languages with no orthographic rules](#). In *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 576–580.
 - [22] Ossama Obeid, Nasser Zalmout, Salam Khalifa, et al. 2020. [CAMEL Tools: An open source python toolkit for Arabic natural language processing](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.
 - [23] Abdul Muneer, Rishabh Saraf, Ramesh Gundluru, et al. 2026. [Introducing Falcon ASR](#). Hugging Face blog, 7 October 2026.
 - [24] Elm Research Center. 2026. [Open Universal Arabic ASR Leaderboard](#). Hugging Face Space, accessed 10 October 2026.
 - [25] IFM Team. 2026. [Introducing Jais-ASR: Arabic speech recognition for every dialect](#). Hugging Face model card, accessed 16 October 2026.
 - [26] Seamless Communication, Loïc Barrault, Yu-An Chung, et al. 2023. [Seamless: Multilingual expressive and streaming speech translation](#). *Preprint*, arXiv:2312.05187.

A Speech Recognition: Supplementary Results

A.1 CER by Variety

Variety	Sada-ASR	API					Self-hosted					
		MAI-2	Scribe v2	Gemini	gpt-tr.	Speechm.	Jais	Audar	FastConf.	Whisper	Nemotron	Cohere
MSA	2.08	3.58	3.72	4.17	3.63	3.57	3.21	6.22	6.33	9.68	4.72	11.36
Egyptian	3.73	6.56	7.33	7.31	8.79	8.17	6.67	11.62	11.92	14.49	12.49	15.52
Saudi	5.61	7.53	9.72	9.79	9.82	9.79	8.57	10.17	14.24	16.30	13.52	19.25
Emirati	8.54	9.44	11.60	13.56	14.83	13.97	10.83	13.24	18.83	22.02	16.73	31.22
Kuwaiti	4.79	5.43	6.69	8.03	7.95	17.56	7.20	9.11	15.91	15.46	12.72	27.13
Qatari	7.56	8.56	10.85	11.48	12.78	11.76	9.65	11.53	18.74	18.38	15.14	31.13
Omani	7.77	10.02	11.55	11.96	13.11	11.19	11.15	13.57	18.35	20.37	17.22	28.08
Syrian	7.79	7.81	10.43	11.92	12.49	15.51	10.64	13.64	17.53	21.95	20.84	26.08
Lebanese	10.86	11.76	15.52	18.20	17.46	18.35	13.82	16.88	22.01	25.79	24.81	39.10

Table 7: CER (%) by variety for all twelve systems; systems as in Table 2.

A.2 Public Test Sets: CER

Table 8 gives Sada-ASR’s WER and CER on each of the six public test sets under the leaderboard’s current scoring (Section 6.4).

	SADA	CV 18.0	MASC-c	MASC-n	Casab.	MGB-2	Avg
WER	25.76	3.95	5.35	16.70	41.72	9.56	17.17
CER	12.07	1.19	2.13	6.65	15.41	4.82	7.05

Table 8: Sada-ASR on the six public test sets under the leaderboard’s current scoring (%); Avg is the unweighted mean.