SEBHA UNIVERSITY CONFERENCE PROCEEDINGS, VOL.5 NO. 4 2026
DOI: 10.51984/v08aa731

وقائع مؤتمرات جامعة سبها
Sebha University Conference Proceedings

Conference Proceeding homepage: http://www.sebhau.edu.ly/journal/CAS

A Zero-Shot Benchmark of Arabic Dialect Identification Models on Libyan Speech

Nour Zohdy*, Mansour Essgaer

Faculty of Information Technology, Sebha University, Sebha, Libya

Keywords:
Arabic dialect identification
Libyan Arabic
zero-shot evaluation
speech benchmark
low-resource languages
model generalisation

A B S T R A C T

Spoken Arabic dialect identification (DID) has advanced quickly, yet Libyan Arabic remains almost invisible in public benchmarks: it is either absent from the label set or folded into a coarse "Maghrebi" class, so a model can appear to cover Libyan speech while never recognising it. No published study reports how existing spoken Arabic DID systems perform on Libyan speech. This study tests whether publicly available Arabic DID models recognise Libyan speech when they encounter it. We evaluate three off-the-shelf models spanning the three label granularities available in public checkpoints: a 5-class Massively Multilingual Speech (MMS) based regional identifier (badrex), an 18-class HuBERT classifier, and the 20-class ADI-20 system that won the spoken-DID subtask of the 2025 Nuanced Arabic Dialect Identification (NADI) evaluation, on 79,395 segments (135.4 hours, 664 speakers, 163 sources) of region-labelled Libyan Arabic, with no fine-tuning. Every evaluated speaker was reviewed by an annotator raised in Libya, and speakers judged to be speaking Modern Standard Arabic or a non-Libyan variety were removed before evaluation. Because every segment carries a single true label (Libyan), we report recall rather than accuracy and organise it hierarchically (exact-country, Maghrebi-region, and top-3). ADI-20 assigns the correct Libyan label to only 14.1% of segments (95% confidence interval 10.2–19.2%), routing 53.1% of Libyan speech to Tunisian and 18.1% to Modern Standard Arabic. The HuBERT model reaches 48.1% Libyan recall (43.9–52.7%) and 73.3% top-3 recall, exceeding ADI-20 by 34.0 points on exact Libyan recall (95% CI 31.2–37.2 points, McNemar $p < 0.001$). ADI-20 shows a significant regional collapse, dropping from 15.6% recall on the coastal regions to 6.3% on southern Fezzan, while HuBERT is stable across regions. Recall rises modestly with segment duration, and the differences we observe between male and female speakers are reported as descriptive, since they are confounded with source and speaker composition. An episode-level provenance audit traces 1.10% of the evaluation segments to the ADI-17 catalogue. On that audio the two catalogue-trained models reach about 0.93 recall, against 0.13 and 0.48 on the rest, while badrex, which is not built on ADI-17, shows no gain. Removing it shifts every headline figure by less than 0.01. Provenance is established for 57.7% of segments, so further overlap cannot be excluded. The results show that "Maghrebi" coverage does not imply Libyan coverage, and that reporting a single macro-accuracy hides where a model actually fails.

تقييم نماذج تحديد اللهجات العربية على الكلام الليبي في بيئة صفريّة التدريب

نور زهدي*، منصور الصغير

كلية تقنية المعلومات، جامعة سبها، سبها، ليبيا

الكلمات المفتاحية:
تحديد اللهجات العربية
اللهجة الليبية
التقييم دون تدريب
قياس أداء الكلام
اللغات محدودة الموارد
تعميم النماذج.

الملخص

شهد تحديد اللهجات العربية المنطوقة (DID) تقدّمًا سريعًا، غير أنّ اللهجة الليبية تظلّ شبه غائبة عن معايير القياس العامة: فهي إمّا غير موجودة في مجموعة التصنيفات أو مُدمَجة ضمن فئة «مغاربية» واسعة، فيبدو النموذج مُغطّيًا للكلام الليبي دون أن يتعرّف عليه فعليًا. ولا توجد دراسة منشورة تُبيّن أداء أنظمة تحديد اللهجات العربية المنطوقة على الكلام الليبي. نطرح سؤالًا مباشرًا: هل تتعرّف نماذج تحديد اللهجات العربية المتاحة للعموم على الكلام الليبي عند مواجهته؟ نُقيّم ثلاثة نماذج جاهزة تُغطّي مستويات التصنيف الثلاثة المتاحة في النماذج المنشورة: مُعرّف إقليمي من خمس فئات مبني على MMS (badrex)، ومُصنّف HuBERT من ثماني عشرة فئة، ونظام ADI-20 من عشرين فئة الذي فاز في مهمّة الكلام ضمن تقييم NADI 2025، على 79،395 مقطعًا (135.4 ساعة، 664 متحدّثًا، 163 مصدرًا) من الكلام الليبي الموسوم إقليميًّا، دون أيّ

*Corresponding author:
E-mail addresses: nour.tammous@idp.sebhau.edu.ly ,(M. Essgaer)man.essgaer@sebhau.edu.ly
Article History : Received 30 May 2026 - Received in revised form 30 September 2026 - Accepted 11 November 2026

تدريب إضافي. وقد رُوجِع كلّ متحدّث في مجموعة التقييم من قِبَل مُحكِّم نشأ في ليبيا، واستُبعد قبل التقييم كلّ متحدّث حُكم بأنّه يتحدّث العربية الفصحى الحديثة أو لهجة غير ليبية. ولأنّ كلّ مقطع يحمل تصنيفًا حقيقيًّا واحدًا (ليبي)، نُبلّغ عن الاستدعاء (recall) بدلًا من الدقّة، ونُنظّمه هرميًّا: الاستدعاء الليبي الدقيق، والاستدعاء المغاربي الإقليمي، والاستدعاء ضمن أفضل ثلاثة. يُسنِد نظام ADI-20 التصنيف الليبي الصحيح إلى 14.1٪ فقط من المقاطع (فاصل ثقة 95٪: 10.2–19.2٪)، مُوجّهًا 53.1٪ من الكلام الليبي إلى التونسية و18.1٪ إلى العربية الفصحى الحديثة. ويبلغ نموذج HuBERT استدعاءً ليبيًّا قدره 48.1٪ (43.9–52.7٪) واستدعاءً ضمن أفضل ثلاثة قدره 73.3٪، متجاوزًا ADI-20 بمقدار 34.0 نقطة في الاستدعاء الليبي الدقيق ($p < 0.001$). ويُظهر ADI-20 انهيارًا إقليميًّا دالًّا، إذ ينخفض من 15.6٪ على المناطق الساحلية إلى 6.3٪ على فزّان الجنوبية؛ بينما يبقى HuBERT مستقرًّا عبر المناطق. ويرتفع الاستدعاء ارتفاعًا طفيفًا مع طول المقطع، ونُبلّغ عن الفروق بين المتحدّثين والمتحدّثات بوصفها فروقًا وصفية لتداخلها مع تركيبة المصادر والمناطق. ويتتبّع تدقيق المصدر على مستوى الحلقات نسبة 1.10٪ من مقاطع التقييم إلى مجموعة ADI-17. وعلى هذه المقاطع يبلغ النموذجان المدرَّبان على تلك المجموعة استدعاءً يقارب 0.93، مقابل 0.13 و0.48 على بقيّة البيانات، في حين لا يُظهر badrex، غير المبني على ADI-17، أيّ تحسّن. وإزالة هذه المقاطع تُغيّر كلّ الأرقام الرئيسة بأقلّ من 0.01. وقد أُثبت المصدر لنسبة 57.7٪ من المقاطع، ولذلك لا يمكن استبعاد وجود تداخل إضافي. تُظهر النتائج أنّ التغطية «المغاربية» لا تعني تغطية ليبية، وأنّ الاكتفاء بدقّة كلّية واحدة يُخفي المواضع التي يفشل فيها النموذج فعليًّا.

1. Introduction

Spoken Arabic dialect identification (DID) decides, from audio alone, which regional variety of Arabic a speaker is using. It has become a standard task in Arabic speech processing, driven by shared tasks such as the Multi-Genre Broadcast (MGB) challenges [1], [2] and the recent Nuanced Arabic Dialect Identification (NADI) evaluations [4]. Progress is usually reported as macro-averaged accuracy over a fixed label set of countries or regions. This framing has a blind spot: a dialect that is not in the label set, or that is merged into a broad regional bucket, can appear to be "covered" while never actually being recognised.

Libyan Arabic sits squarely in that blind spot. It is spoken by roughly six million people across a territory larger than any other North African country, yet it is one of the least represented varieties in public speech resources. In the label sets of widely used models it is either absent entirely or collapsed into a single "Maghrebi" or "North African" class shared with Tunisian, Algerian, Moroccan and Mauritanian Arabic. A model can therefore score well on a Maghrebi bucket while systematically mistaking Libyan speech for a neighbouring dialect. Macro-accuracy on the model's own label set will never reveal that failure mode.

Two gaps follow from this. Empirically, no published study reports how well existing spoken Arabic DID systems perform on Libyan speech: the Libyan work that exists is text-based [6], [7], and the speech benchmarks that carry a Libya label report only aggregate accuracy over their own label sets, never a Libya-specific figure on independent audio. Methodologically, the measures used to report those benchmarks, macro-accuracy and macro-F1 over a balanced label set, cannot express the failure mode that matters here, in which a model places speech in the right region and still names the wrong country. This paper addresses both. It measures off-the-shelf performance on Libyan speech directly, and it reports that performance as a hierarchy of recall levels that separates regional from country-level correctness.

We take the position of a practitioner who has Libyan audio and wants to know, before building anything, whether existing tools work on it. We evaluate three publicly available Arabic DID models, without any fine-tuning, on a large region-labelled corpus of Libyan Arabic. Our contributions are:

- A zero-shot benchmark of three off-the-shelf models, covering the coarse-regional, mid-grained country, and fine-grained regimes, on 79,395 segments of Libyan speech from 664 human-verified speakers.
- A metric design suited to a single-true-class corpus: recall reported as a hierarchy (exact Libyan, Maghrebi-region, top-3) rather than a single accuracy that conflates distinct error types.
- An error analysis showing *where* Libyan speech is sent when it is missed, a per-region breakdown revealing a significant collapse on southern Libyan speech for the strongest model, and robustness slices by duration, gender, and speaker-level aggregation.
- An episode-level provenance audit of the evaluation set, which traces 1.10% of segments to the ADI-17 catalogue and shows that the two catalogue-trained models reach about 0.93 recall on that audio against 0.13 and 0.48 on the rest, while a model not built on ADI-17 shows no such gain. Removing the traced audio shifts every headline figure by less than 0.01.

2. Related Work

Benchmarks and shared tasks. Spoken Arabic DID was standardised by the MGB-3 Arabic task, which introduced a 5-way regional split (ADI-5: Modern Standard Arabic, Egyptian, Gulf, Levantine, and North African/Maghrebi) [1]. MGB-5 extended this to 17 country-level dialects (ADI-17) with a large collection of broadcast and web audio [2]. Most recently, the ADI-20 task expanded coverage to 20 Arabic varieties and was run as a subtask of the 2025 Nuanced Arabic Dialect Identification (NADI) evaluation [3], [4]. Libya appears as a label in ADI-17 and ADI-20 but is absent from the coarse ADI-5 scheme, where it is subsumed into the North African class.

Robustness and generalisation. A recurring finding is that DID models degrade sharply out of their training domain. Sullivan et al. [5] showed that Arabic DID systems are highly sensitive to genre and recording conditions, with accuracy dropping substantially when the evaluation genre differs from training. This motivates evaluating models zero-shot on an independent corpus rather than on held-out data from the same source.

Libyan dialect identification. Work specific to Libyan Arabic has, until now, been almost entirely text-based, targeting the written dialect in social-media text [6], [7]. To our knowledge there is no prior evaluation of spoken Libyan dialect identification against off-the-shelf models, which is the gap this paper addresses.

Speech foundation models. The three systems evaluated here are all built on self-supervised or weakly supervised speech encoders, and their behaviour on Libyan Arabic is partly a property of those encoders. wav2vec 2.0 introduced masked-prediction pre-training on raw audio [10]; HuBERT refined the objective using iteratively refined hidden-unit targets [11]; and WavLM added denoising and speaker mixing to improve robustness on noisy and overlapped speech [13]. Multilingual scaling followed: XLS-R pre-trains on 128

languages [12] and Massively Multilingual Speech (MMS) on more than a thousand [14], while Whisper takes the weakly supervised route, training on large volumes of paired multilingual audio and text [15]. Dialect and language identification is normally built on top of these encoders by pooling frame-level representations, either with a lightweight classification head, as in the HuBERT and Whisper-encoder systems evaluated here, or with a dedicated speaker-embedding architecture such as ECAPA-TDNN [16], which underpins the language-identification systems trained on VoxLingua107 [17]. The relevance to this study is that multilingual pre-training coverage is not the same as dialect discrimination. MMS and Whisper both see Arabic in quantity, but the Arabic they see is dominated by broadcast Modern Standard Arabic and by the better-resourced dialects, so a fine-tuned classification head inherits whatever separation between Libyan and its neighbours the encoder already provides, and cannot create it. Recent work has begun to benchmark dialects and regional varieties against these encoders directly [18]; that is the closest analogue to the evaluation we report, though its Arabic label set is regional and has no Libyan class.

3. Corpus and Evaluation Set

The evaluation uses a region-labelled corpus of spontaneous Libyan Arabic speech that we assembled from publicly available Libyan broadcast and online media: 163 sources covering news, talk shows, interviews, and documentary programming. Audio was resampled to 16 kHz mono, and speech regions were segmented into single-speaker utterances of 3 to 10 seconds (median 6.0 s). Speaker turns were clustered by speaker embedding within each source, and every resulting speaker was assigned to one of the three historical regions of Libya, Tripolitania (west), Cyrenaica (east), and Fezzan (south), from the speaker's documented origin together with the regional coverage of the source. Speaker identities are disjoint across the corpus partitions.

Segments were first diarised and then merged within each source at a cosine similarity of at least 0.70 between speaker embeddings, producing 2,227 per-source speaker clusters. These clusters were linked across sources at a similarity of at least 0.72 so that a speaker appearing in several shows formed one identity, while uncertain pairs and any pair that would join two different regions were confirmed by hand rather than merged by rule. This linking produced 2,042 distinct speakers. Fig. 1 summarises the pipeline.

Quality and duration filters were then applied to the segments, and each speaker was measured using the audio that survived. An automatic evidence floor excluded any speaker with fewer than 5 segments or less than 45 seconds of speech. Below that point, there is not enough audio to characterise the speaker or to read a model's behaviour on them. This rule removed 1,309 of the 2,042 speakers, the largest single reduction in the pipeline, and left 733 for human review.

Each of these 733 speakers was reviewed by an annotator raised in Libya. The judgement used up to ten clips of at most ten seconds each, sampled to favour clean and longer segments, rather than a single sample. Human review rejected 69 speakers: 29 for Modern Standard Arabic, 28 for unusable audio, and 12 as non-Libyan. A further 32 speakers were kept with corrected region labels because a wrong region is a labelling error and not a reason to discard a speaker. After the 69 rejections, 664 speakers remained, and every speaker in the evaluation set had been reviewed by a human.

Of the 79,395 segments, 73,216 came from speakers confirmed in their original region, while 6,179 came from the 32 speakers whose region was corrected. Review was per speaker, not per segment, so an occasional Modern Standard Arabic or code-switched segment from an otherwise confirmed dialectal speaker may remain.

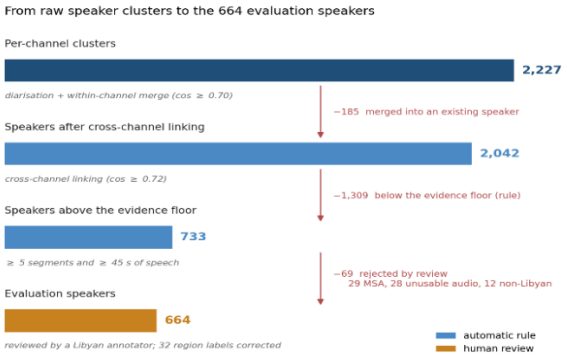


Fig. 1: From per-source speaker clusters to the 664 evaluation speakers. Blue stages are automatic rules, the final stage is human review.

Recording conditions are those of the source media rather than of a controlled collection, so we report a per-segment signal-to-noise proxy rather than a calibrated SNR. Each segment is framed into 25 ms windows; we take the RMS energy of each frame in decibels and report the difference between the 90th and the 10th percentile of those frame energies, that is, the active-speech level above the quiet noise floor within the same segment. This is a model-free percentile estimate in the spirit of NIST- and WADA-style measures. It is comparable across segments but should not be read as an absolute SNR. On this measure the evaluated segments have a median of 27.3 dB and 93.6% sit at or above the 15 dB clean-speech threshold (Cyrenaica 95.7%, Tripolitania 92.5%, Fezzan 90.7%); segments falling below a 3 dB floor were rejected during corpus construction. Speaker age is not annotated in the corpus and therefore cannot be reported.

We evaluate on all 79,395 segments whose final dialect label is one of the three Libyan regions: 135.4 hours, 664 speakers, and 163 sources. The sources are 59 YouTube channels and 104 TikTok accounts; the TikTok material is the larger share by source count but only 15.2 of the 135.4 hours. All three models are used zero-shot, so no model was trained, selected, or tuned on this corpus and no in-study fine-tuning occurred. We therefore evaluate on all segments rather than a held-out split, which maximises statistical power. The released manifest does contain speaker-disjoint train, development, and test partitions of 58,017, 6,558, and 14,820 segments respectively, but because no model was trained on this corpus the split boundary does not affect the protocol. Table 1 summarises the composition by region. The set is speaker-imbalanced by gender (539 male, 107 female, 18 unlabelled speakers), and clips per speaker are strongly right-skewed: mean 119.6, median 26, interquartile range 13–106, 90th percentile 283, maximum 6,949. Both skews are reasons we report speaker-clustered intervals and speaker-level aggregation alongside segment-level recall, and we return to them in the robustness analysis.

Table (1): Composition of the Libyan evaluation set by region, with recording-quality statistics. Clips per speaker are reported as the median because the distribution is strongly right-skewed. SNR is the per-segment percentile proxy defined in the text.

Region	Segments	Speakers	Sources	Hours	Clips/spk (med.)	Median SNR (dB)	% ≥ 15 dB	% female spk
Tripolitania (west)	32,831	275	72	56.7	**23**	**26.2**	**92.5**	**15.6**
Cyrenaica (east)	33,822	241	64	56.4	**26**	**28.6**	**95.7**	**12.4**
Fezzan (south)	12,742	148	45	22.3	**31**	**27.0**	**90.7**	**23.0**
Total	**79,395**	**664**	**163**	**135.4**	**26**	**27.3**	**93.6**	**16.1**

4. Models and Evaluation Protocol

4.1. Models

We selected the models against four explicit criteria: publicly released weights; a direct Arabic dialect identification output rather than general-purpose speech recognition; usability with no fine-tuning, as the zero-shot protocol requires; and complementary label granularity, so that the coarse-regional, mid-grained country, and fine-grained regimes are each represented. Three systems satisfy all four (Table 2). They span the three label granularities available in public Arabic DID checkpoints, which is not the same as spanning the design space of the field; the architectures that are published but not released are discussed in the Limitations.

- **badrex** (badrex/mms-300m-arabic-dialect-identifier) [20]: a Massively Multilingual Speech (MMS) 300M model [14] fine-tuned for 5-way regional identification (MSA, Egyptian, Gulf, Levantine, Maghrebi). It has no Libyan or country-level label; the finest correct answer it can give for Libyan speech is "Maghrebi".

- **hubert** (IbrahimAmin/hubert-arabic-spoken-dialect-classifier) [19]: a HuBERT-base model [11] fine-tuned for 18-way dialect classification, including an explicit Libyan (LIB) label.
- **ADI-20** (Elyadata/ADI-whisper-ADI20) [3]: the 20-class system built on Whisper encoders [15] that won the spoken-DID subtask of the NADI 2025 shared task [4] (79.83% accuracy on the official test set), including a Libya label.

4.2. Metric design

Our corpus has a single true label for every segment (Libyan), so the standard multi-class metrics behave degenerately rather than merely differently. Positive-class precision is fixed at 1.0 by construction and F1 becomes a deterministic function of recall, so neither adds information; Section 5.6 derives this for all five commonly requested measures. We therefore report recall, the fraction of Libyan segments a model correctly identifies, as the headline. To compare models whose label sets differ in granularity, we organise recall as a hierarchy:

- **Libyan recall**: top-1 prediction is exactly Libyan (not defined for badrex, which lacks the label).
- **Maghrebi-region recall**: top-1 prediction is Libyan or another Maghrebi variety (Tunisian, Algerian, Moroccan, Mauritanian). This is the "coarsely correct" answer, and the only correct answer available to badrex.
- **Top-3 Libyan recall**: Libyan appears among the model's three highest-scoring labels.

Modern Standard Arabic is treated as a full error for all models, since our segments are dialectal speech. We report 95% confidence intervals from a speaker-level bootstrap (2,000 resamples), resampling speakers rather than segments to respect the clustering of segments within speakers. Because all models predict on identical segments, we also report paired comparisons: a paired speaker-level bootstrap of the difference between two models, and McNemar's test on the discordant segment pairs as a confirmatory check. We aggregate to the speaker level in two ways: by averaging per-speaker recall (speaker-macro), and by taking each speaker's modal predicted label, which we call a plurality vote. We use the term plurality rather than majority deliberately. With eighteen or twenty candidate labels, a model's correct label can win a speaker's vote on well under half of that speaker's segments, so the modal-label rule is not a majority rule and the two give different numbers; Section 5.5 reports both.

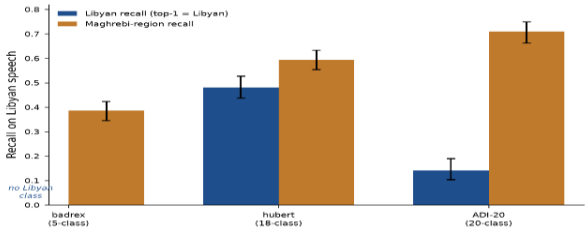


Fig. 2: Zero-shot Libyan and Maghrebi-region recall for the three models, with 95% speaker-bootstrap confidence intervals. Exact Libyan recall is undefined for badrex, which has no Libyan label.

5. Results

5.1. Main benchmark

No model identifies Libyan speech reliably. Table 2 gives the recall hierarchy, and Fig. 2 puts Libyan recall next to Maghrebi-region recall.

ADI-20 gives the correct Libyan label to 14.1% of segments (95% CI 10.2–19.2%). The same model scores nearly 80% on the NADI 2025 test set [4]. Every interval here comes from the speaker bootstrap of Section 4.2, so it answers one question: how far would the number move if the corpus held a different set of 664 Libyan speakers. HuBERT reaches 48.1% at country level (CI 43.9–52.7%) and finds Libyan inside its top three guesses for 73.3% of segments (CI 69.3–77.3%). badrex has no Libyan label, so it is scored only at the regional level, where its Maghrebi class takes 38.6% of segments (CI 34.5–42.6%).

The two country-level models are inverted relative to each other. ADI-20 has the highest Maghrebi-region recall, 71.0%, together with the lowest Libyan recall, 14.1%. It hears the speech as Maghrebi and then names the wrong country. HuBERT is the other way round, at 59.3% region and 48.1% country. Finer label granularity does not by itself produce better Libyan identification.

The gap between the two is wide relative to its uncertainty. HuBERT leads ADI-20 by +0.340 on exact Libyan recall (CI [+0.312, +0.372]) and by +0.375 on top-3 recall (CI [+0.330, +0.417]). ADI-20 leads on Maghrebi-region recall by +0.117 (CI [+0.057, +0.178]). McNemar's test agrees in all three cases ($p < 0.001$), with discordant segments splitting 29,232 to 2,212 in HuBERT's favour on exact Libyan recall, 32,156 to 2,345 on top-3, and 11,287 to 20,562 against it on Maghrebi recall. At $n = 79{,}395$ the χ^2 values are large by construction, and the test counts every segment as independent when segments from one speaker are not. We therefore treat the speaker-clustered intervals as the uncertainty statement and read the discordant counts as the effect size. badrex is left out of the exact-Libyan comparisons, since a Maghrebi hit is a different outcome.

Table (2): Zero-shot recall on Libyan speech. Values are top-1 recall with 95% speaker-bootstrap confidence intervals (2,000 resamples). A dash denotes a metric not defined for that model's label set.

Model	# labels	Libyan recall	Maghrebi recall	Top-3 Libyan
badrex	5	—	0.386 [0.345, 0.426]	—
hubert	18	0.481 [0.439, 0.527]	0.593 [0.554, 0.635]	0.733 [0.693, 0.773]
ADI-20	20	0.141 [0.102, 0.192]	0.710 [0.663, 0.751]	0.357 [0.303, 0.427]

5.2. Provenance of the evaluation set

The models under test are built on public Arabic dialect resources, among them ADI-17 [8], ADI-20 [3], and the QCRI Aljazeera Speech Resource (QASR) [9]. We could not verify the exact training composition of each released checkpoint. Any evaluation audio drawn from those resources is no longer out-of-domain, so we audited our own corpus at the episode level rather than comparing source names against a published landing page. Each episode carries a metadata record naming where the audio came from. Fig. 3A gives the result.

Source identity is known for 98.7% of evaluation segments. Provenance is established for 57.7%: 45.9% self-mined from a named source, 10.7% self-mined from TikTok accounts, and 1.10% traced to the ADI-17 catalogue. The remaining 41.7% come from named Libyan programmes that we mined ourselves, but whose provenance field was never filled, so the record alone cannot show that they are absent from any training set.

The catalogue-traced portion is 876 segments, 1.5 hours, 20 speakers and 8 episodes, from three sources. Those three sources contribute no other episodes to the evaluation set, so the traced audio is bounded by sources we can name.

Recall on that audio separates the models sharply (Fig. 3B). ADI-20 reaches 0.941 (CI [0.854, 0.975]) on the traced segments against 0.132 (CI [0.093, 0.186]) on the rest of the corpus. HuBERT reaches 0.928 (CI [0.844, 0.965]) against 0.476 (CI [0.431, 0.523]). badrex, which is not built on ADI-17, shows no gain at all: 0.333 against 0.387. The badrex result makes a simple "this audio is easier" reading hard to sustain, and the contrast is consistent with the two other models having been exposed to this material in training. We claim no more than consistency. A provenance tag records where we obtained a file, not that the file entered a particular checkpoint's training data, and 876 segments from 20 speakers is a narrow base for a population claim.

Removing the traced audio changes little. Libyan recall moves from 0.4815 to 0.4765 for HuBERT and from 0.1411 to 0.1322 for ADI-20, while badrex moves from 0.3862 to 0.3867. Excluding all 20 affected speakers instead of only their traced segments gives 0.4780, 0.1315 and 0.3857. Both shifts are smaller than the confidence intervals of Section 5.1, so we report the full corpus as the primary evaluation and treat the removal as a sensitivity check. Restricting the traced subset to the 17 speakers who appear only in traced episodes

leaves the contrast intact, at 0.950 for ADI-20 and 0.933 for HuBERT against 0.260 for badrex.

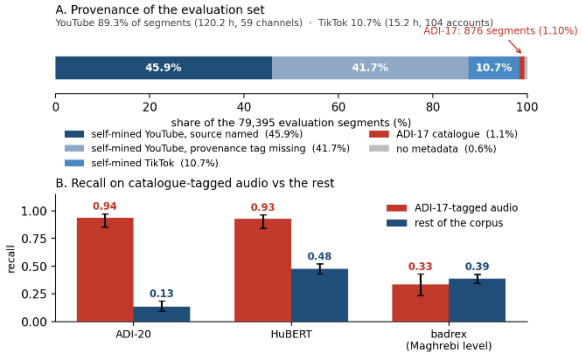


Fig. 3: Provenance of the evaluation set and its effect on recall. (A) share of segments by provenance. (B) recall on catalogue-traced audio against the rest of the corpus, with 95% speaker-bootstrap confidence intervals; badrex is scored at the Maghrebi level.

5.3. Where does Libyan speech go?

Fig. 4 shows the full top-1 label distribution for each model. Every segment is Libyan, so all mass outside the Libyan bar is error, and the shape of that mass separates the models.

ADI-20 shows the clearest pattern. It sends 53.1% of Libyan speech to Tunisian and 18.1% to Modern Standard Arabic. The confusion does not spread evenly across neighbours. It concentrates on Tunisian, the Maghrebi variety next door. One reading that fits is a model with a broad western-Maghreb acoustic profile that does not separate Libyan from Tunisian inside it. Our data cannot separate that from a label prior or a training-proportion effect.

HuBERT's errors are thinner and wider, across Egyptian (8.8%), MSA (7.2%) and Algerian (6.0%). badrex divides Libyan speech almost evenly between Maghrebi, Gulf and Levantine, which suggests that at 5-way granularity the Libyan cues are not distinctive enough to hold even the correct region.

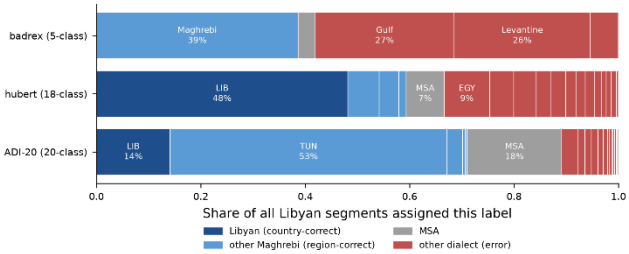


Fig. 4: Top-1 label distribution of Libyan speech under each model. Every segment is Libyan, so all mass outside the Libyan bar is error.

5.4. Regional robustness

Table 3 and Fig. 5 break Libyan recall down by region.

HuBERT is flat: 0.515 for Tripolitania, 0.454 for Cyrenaica, 0.467 for Fezzan. Its coastal-minus-Fezzan gap of 0.017 is not significant (95% CI [-0.057, 0.091]).

ADI-20 is not flat. Recall falls from 0.156 on the coastal regions to 0.063 in southern Fezzan, a gap of 0.093 whose interval excludes zero (95% CI [0.039, 0.155]). Fezzan is the region with the least available audio, and it is where ADI-20 loses most. Whether that comes from a real acoustic distance from the coastal Maghreb continuum, or from thinner training coverage, is not something this evaluation can separate. Either way, Fezzan is the population that a benchmark reporting only "Libyan" coverage would hide.

Table (3): Libyan recall by region, with the coastal (Tripolitania + Cyrenaica) minus Fezzan gap and its bootstrap significance.

Model	Tripolitania	Cyrenaica	Fezzan	Coastal–Fezzan gap
hubert	0.515	0.454	0.467	0.017 (n.s.)
ADI-20	0.194	0.120	0.063	0.093 (sig.)

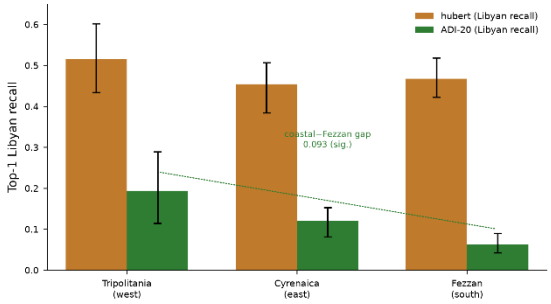


Fig. 5: Libyan recall by region for the two country-level models, with 95% speaker-bootstrap confidence intervals.

5.5. Duration, gender, and speaker-level aggregation

Recall rises with segment duration for both country-level models and stays low throughout (Fig. 6). HuBERT goes from 0.399 on 3–5 s segments to 0.554 on 7–10 s segments, ADI-20 from 0.119 to 0.162. Longer context helps. It does not close the gap.

Both models score lower on segments from female speakers: HuBERT 0.491 male against 0.427 female, ADI-20 0.151 against 0.091. We report this as descriptive. The corpus is strongly speaker-imbalanced, one speaker contributes 6,949 segments, and gender is confounded with source, region and clips-per-speaker composition, so a difference of this size cannot be assigned to gender alone.

Aggregating to the speaker level raises the apparent numbers, and the plurality vote raises them most. HuBERT's per-speaker plurality-vote recall reaches 0.750 against 0.481 per clip, and its speaker-macro recall is 0.529, while ADI-20 stays at 0.170. Table 4 collects these slices.

The size of that jump is itself a result, and its cause is the shape of the errors. HuBERT's correct predictions all land on one label, while its errors scatter across seventeen competing labels. A speaker can therefore be labelled Libyan on well under half of their segments and still have Libyan as their most frequent label, because no single wrong label collects more. Of the 498 speakers HuBERT wins this way, 142 have Libyan on half or fewer of their segments, and the lowest winning share is 18.2%. Under a strict majority rule, which requires Libyan on more than half of a speaker's segments, HuBERT scores 0.536. We report both, and read 0.536 as the stricter statement about identifying a speaker. ADI-20 gains almost nothing from pooling, because its errors sit on Tunisian, and voting cannot remove an error that keeps winning.

Table (4): Robustness slices of Libyan recall: segment-duration buckets, speaker gender, and speaker-level aggregation

	Duration			Gender		Speaker-level		
Model	**3–5 s**	**5–7 s**	**7–10 s**	**male**	**female**	**clip**	**macro**	**plur. vote**
hubert	0.399	0.490	0.554	0.491	0.427	0.481	0.529	0.750
ADI-20	0.119	0.143	0.162	0.151	0.091	0.141	0.157	0.170

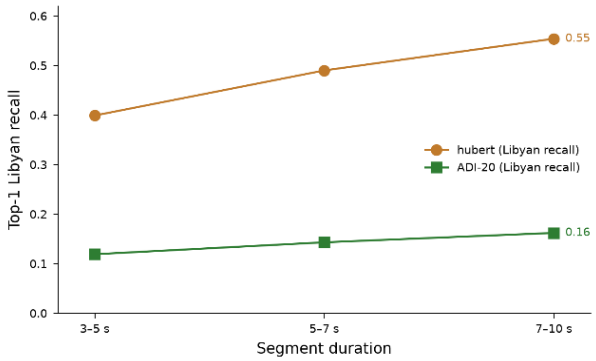


Fig. 6: Libyan recall by segment-duration bucket for the two country-level models.

5.6. Precision, F1, balanced accuracy, MCC, and Cohen's κ

Every evaluation segment carries the same true label, which makes the standard summary measures degenerate here. We give their values rather than omit them.

Precision is 1.0 by construction for every model, because a false positive would need a segment whose true label is not Libyan, and the corpus contains none. A model that predicted Libyan once, correctly, would score the same. F1 reduces to 2R/(1+R) and carries no information beyond recall: 0.650 for HuBERT and 0.247 for ADI-20 at the country level, and 0.557 for badrex at the Maghrebi level. Balanced accuracy and Matthews' correlation coefficient are undefined, since both need counts from a negative class that a single-label corpus cannot contain; implementations that average recall over the observed classes return the recall value itself, which is a software convention and not a result. Cohen's κ is exactly zero for all three models, because with one true class the observed and chance-expected agreement are both equal to the predicted-positive rate R, so $\kappa = (R - R)/(1 - R) = 0$.

These values describe the evaluation design, not the models. The recall hierarchy stays as the headline, since it is the only family of measures that separates the error types this corpus can exhibit.

6. Discussion

Three findings emerge. First, region coverage is not country coverage: ADI-20 recognises Libyan speech as broadly Maghrebi 71% of the time but names the correct country only 14% of the time, calling most of it Tunisian. A benchmark that reports only accuracy on a model's own label set, or that merges Libyan into a Maghrebi bucket, would score this as a success. Second, a high aggregate score does not transfer: ADI-20's near-80% NADI accuracy coexists with 14% Libyan recall on independent audio, which echoes the genre- and domain-sensitivity reported for Arabic DID [5]. Third, the failures are structured rather than random: they concentrate on the geographically adjacent Tunisian label, on the under-represented southern Fezzan variety, and on female speakers, all directions that a single macro number obscures.

For a practitioner with Libyan audio, the practical reading is that the mid-grained HuBERT model is the most usable off-the-shelf option (48% per clip, 75% per speaker under a plurality vote, though only 54% under a strict majority), that the coarse badrex model is unreliable even at the regional level, and that the ADI-20 system is not usable for Libyan identification without adaptation despite its benchmark ranking. The pattern also connects to the wider multilingual speech literature: the encoders underlying these systems are pre-trained on many languages [12], [14], [15], but breadth of language coverage does not by itself produce discrimination between closely related varieties within one language, which is the same gap multilingual speech recognition reports for low-resource dialectal Arabic [5].

7. Conclusion

We benchmarked three publicly available Arabic dialect identification models on 135.4 hours of region-labelled Libyan speech, with no fine-tuning, and found that none recognises Libyan speech reliably: the best per-clip Libyan recall is 48.1% (HuBERT) and ADI-20 reaches only 14.1%, sending most Libyan speech to Tunisian or MSA. The study contributes the first spoken-DID benchmark for Libyan Arabic, a recall-hierarchy metric suited to a single-true-class corpus, and an error analysis that localises failures by destination label, region, duration, and gender. Its limitations are that the corpus is drawn from broadcast and online sources and is male-skewed, so results may not transfer to conversational or telephone speech; that human validation was carried out per speaker rather than per segment; that speaker age is not annotated and so could not be reported; that the provenance audit is metadata-level rather than audio-level and establishes provenance for 57.7% of segments, so further overlap with the models' training data cannot be excluded; and that we evaluate the publicly released models rather than the full set of published systems, because as of August 2026 no ECAPA-TDNN, x-vector, or CTC-based Arabic dialect identification checkpoint is publicly available, and training one ourselves would abandon the zero-shot protocol that defines this study. For practitioners, the benchmark gives a realistic expectation of off-the-shelf performance on Libyan audio, identifies the mid-grained model as the most usable starting point, and motivates targeted fine-tuning, especially for southern Fezzan speech, as the next step. All evaluation code, per-segment predictions, and figures will be released on publication to support reproduction and extension to further models.

Abbreviations and Acronyms

ADI Arabic Dialect Identification

ADI-5/17/20 the 5-, 17-, and 20-class Arabic Dialect Identification label sets

CI Confidence Interval

DID Dialect Identification

HuBERT Hidden-Unit Bidirectional Encoder Representations from Transformers

MGB Multi-Genre Broadcast (challenge series)

MMS Massively Multilingual Speech

MSA Modern Standard Arabic

NADI Nuanced Arabic Dialect Identification (shared task)

QASR QCRI Aljazeera Speech Resource

n.s. not statistically significant

sig. statistically significant

Appendix A. Appendix A. Confusion matrices

Full confusion matrices for the three models over all 79,395 Libyan segments. Each row is the single true class (Libyan); the columns are the model's own label set, so every off-diagonal cell is an error.

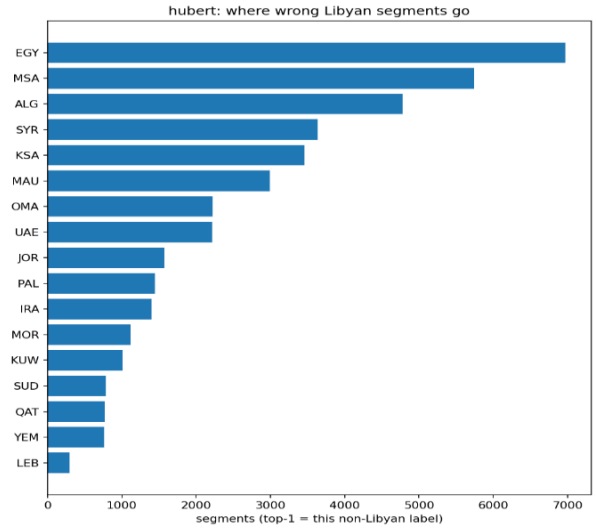


Fig. A1: Predicted-label distribution for HuBERT (18 labels).

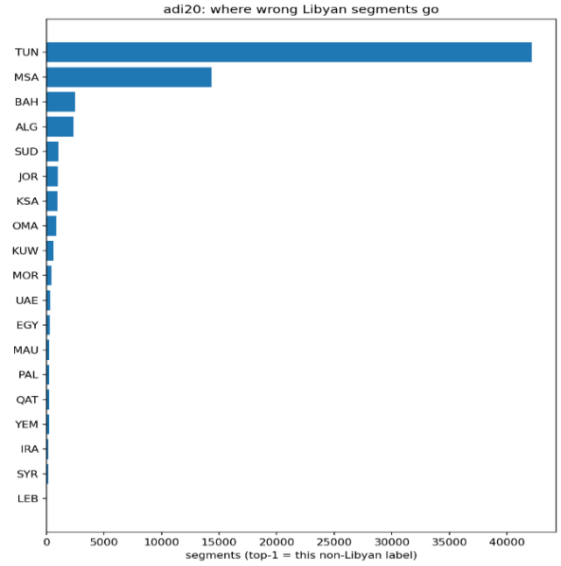


Fig. A2: Predicted-label distribution for ADI-20 (20 labels).

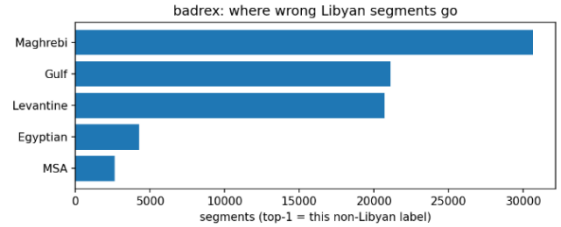


Fig. A3: Predicted-label distribution for badrex (5 labels).

REFERENCES

[1] Ali, A., Vogel, S., Renals, S., et al., (2017), Speech Recognition Challenge in the Wild: Arabic MGB-3, in Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, pp. 316–322.

[2] Ali, A., Shon, S., Samih, Y., et al., (2019), The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech, in Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, pp. 1026–1033.

[3] Elleuch, H., Mdhaffar, S., Estève, Y., Bougares, F., (2025), ADI-20: Arabic Dialect Identification Dataset and Models, in Proc. Interspeech 2025, Rotterdam, Netherlands, pp. 2775–2779. DOI: 10.21437/Interspeech.2025-884.

[4] Talafha, B., Toyin, H. O., Sullivan, P., Elmadany, A., Juma, A., Djanibekov, A., Zhang, C., Alshehhi, H., Aldarmaki, H., Jarrar, M., Habash, N., Abdul-Mageed, M., (2025), NADI 2025: The First Multidialectal Arabic Speech Processing Shared Task, in Proc. Third Arabic Natural Language Processing Conference: Shared Tasks, pp. 720–733.

[5] Sullivan, P., Elmadany, A., Abdul-Mageed, M., (2023), On the Robustness of Arabic Speech Dialect Identification, in Proc. Interspeech 2023, Dublin, Ireland, pp. 5326–5330. DOI: 10.21437/Interspeech.2023-1005.

[6] Ghumeid, N. A., Essgaer, M., (2024), Addressing the Libyan Arabic Dialect Identification: A Comparative Study of Ensemble Classification Methods, in Proc. IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA), Libya. DOI: 10.1109/MI-STA61267.2024.10599739.

[7] Essgaer, M., Massud, K., Al Mamlook, R., Ghmaid, N., (2025), Computational Linguistics Meets Libyan Dialect: A Study on Dialect Identification, International Journal of Intelligent Systems and Applications (IJISA), 17(6), pp. 118–130. DOI: 10.5815/ijisa.2025.06.09.

[8] Shon, S., Ali, A., Samih, Y., Mubarak, H., Glass, J., (2020), ADI17: A Fine-Grained Arabic Dialect Identification Dataset, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 8244–8248.

[9] Mubarak, H., Hussein, A., Chowdhury, S. A., Ali, A., (2021), QASR: QCRI Aljazeera Speech Resource — A Large Scale Annotated Arabic Speech Corpus, in Proc. 59th Annual Meeting of the Association for Computational Linguistics, pp. 2274–2285. DOI: 10.18653/v1/2021.acl-long.177.

[10] Baevski, A., Zhou, Y., Mohamed, A., Auli, M., (2020), wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, in Advances in Neural Information Processing Systems, 33, pp. 12449–12460.

[11] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., (2021), HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp. 3451–3460.

[12] Babu, A., Wang, C., Tjandra, A., et al., (2022), XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale, in Proc. Interspeech 2022, pp. 2278–2282. DOI: 10.21437/Interspeech.2022-143.

[13] Chen, S., Wang, C., Chen, Z., et al., (2022), WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing, IEEE Journal of Selected Topics in Signal Processing, 16(6), pp. 1505–1518.

[14] Pratap, V., Tjandra, A., Shi, B., et al., (2024), Scaling Speech Technology to 1,000+ Languages, Journal of Machine Learning Research, 25(97), pp. 1–52.

[15] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., (2023), Robust Speech Recognition via Large-Scale Weak Supervision, in Proc. 40th International Conference on Machine Learning (ICML), PMLR 202, pp. 28492–28518.

[16] Desplanques, B., Thienpondt, J., Demuynck, K., (2020), ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, in Proc. Interspeech 2020, pp. 3830–3834.

[17] Valk, J., Alumäe, T., (2021), VoxLingua107: A Dataset for Spoken Language Recognition, in Proc. IEEE Spoken Language Technology Workshop (SLT), pp. 652–658.

[18] Feng, T., Huang, K., Xu, A., Shi, X., Lertpetchpun, T., Lee, J., Lee, Y., Byrd, D., Narayanan, S., (2026), Voxlect: A Speech Foundation Model Benchmark for Modeling Dialects and Regional Languages Around the Globe, in Proc. 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). DOI: 10.1145/3770854.3785687.

[19] Amin, I., (2025), HuBERT Arabic Spoken Dialect Classifier [model], Hugging Face. https://huggingface.co/IbrahimAmin/hubert-arabic-spoken-dialect-classifier

[20] badrex, (2025), MMS-300M Arabic Dialect Identifier [model], Hugging Face. https://huggingface.co/badrex/mms-300m-arabic-dialect-identifier