



DOI:10.1145/3451150

BY AHMED ALI, SHAMMUR CHOWDHURY,  
MOHAMED AFIFY, WASSIM EL-HAJJ, HAZEM HAJJ,  
MOURAD ABBAS, AMIR HUSSEIN, NADA GHNEIM,  
MOHAMMAD ABUSHARIAH, AND ASSAL ALQUDAH

# Connecting Arabs: Bridging the Gap in Dialectal Speech Recognition

AUTOMATIC SPEECH RECOGNITION refers to the process through which speech is converted into text. Over the decades, automatic speech recognition has achieved many milestones, thanks to advances in machine learning and low-cost computer hardware. As a result, the best systems for English have achieved a single-digit word error rate (WER) and, in some conversational tasks, performance is comparable to human transcribers. This led researchers to debate

whether the machine has reached human parity in speech recognition.<sup>9,16</sup>

Unlike English, speech recognition in Arabic faces many challenges, even with such advanced techniques. Arabic poses a set of unique challenges due to its rich dialectal variety, with modern standard Arabic (MSA) being the only standardized dialect.<sup>4</sup>

MSA is syntactically, morphologically, and phonologically grounded on classical Arabic, the language of the Qur'an (Islam's Holy Book). Lexically, however, it is much more modern.<sup>8</sup> MSA is taught in schools across the Arab region and is the main language in news broadcasts, parliaments, and formal speech. This is one of the main reasons why MSA has been the main choice for speech and language technology for the last two decades. The current WER for MSA automatic speech recognition (ASR) is about 13%,<sup>2,9</sup> and is worse for dialectal ASR, where the WER averages 30%.<sup>1,3</sup>

Remarkably, the 400 million Arabic native speakers (estimated in 2020) use Dialectal Arabic (DA) as their means of communication in day-to-day speech. MSA is not the native language of any Arab. Dialects used to be primarily spoken, not written. However, this has changed with the rise of Web 2.0, when DA became a written, as well as a spoken, language.

An objective comparison of the varieties of Arabic dialects could potentially lead to the conclusion that Arabic dialects are historically related, and that they are not mutually intelligible languages like English and Dutch. Normal vernacular can be difficult to understand across different Arabic dialects. The tomato example in Figure 1 shows lexical variation across multiple Arabic countries.<sup>5</sup> How different is tomato in Arabic compared to English? In English, there is one lexical form for tomato and two phonological variations, with a 16.7% difference between *to-MAY-to* and *to-MAH-to*, while DA has 10 lexical variations with 67% average-character difference and 15 phonological varia-



tions with 87% average-phoneme difference. Despite the fact that there has been a great deal of speech recognition research in MSA, there is limited effort toward building a platform with standard lexicon and training data to benchmark results and to advance the state of the art in Dialectal ASR.

Dialectal spoken Arabic poses three main challenges: *lack of resources*, *lack of standard orthographic rules*, and *lack of definition* (that is, Arabic is a language with many dialects or sets of languages).

### Spoken Arabic Dialect Identification

From the speech perspective, dialectal

variation increases the level of ambiguity, due to the addition of different linguistic and acoustic representations.

**Arabic Dialectal Corpora:** There have been numerous efforts to produce *spoken Arabic data set resources*. The CallHome task within the 1996–1997 NIST benchmark evaluations framework<sup>15</sup> reported the first transcribed Arabic dialect dataset. In 2003 and 2004, NIST evaluations provided more DA data, mainly in the Egyptian and Levantine dialects, as part of language recognition evaluation. Data from the Iraqi dialect was first obtained as resources for two main research programs: global

autonomous language exploitation (GALE),<sup>14</sup> a U.S. Defense Department Defense Advanced Research Projects Agency (DARPA) program carried out between 2006 and 2009, and the spoken-language communication and translation system for tactical use (TRANSTAC) program, aimed to help the U.S. soldier communicate with non-English speakers using a portable bidirectional translator. These datasets exposed the research community to the challenges in spoken DA.

One of the main challenges of processing dialectal speech is to first identify the dialect of the spoken content. Arabic has more than 30 dialects that can be categorized geographically, socially, or phonologically. However, obtaining complete coverage of major dialects is still challenging.

For the Arabic dialect identification (ADI) task, the multi-genre broadcast (MGB-3) challenge<sup>1</sup> in ASRU-2017 provided a dataset, ADI-5, containing four regional Arabic dialects and MSA, with 53 hours of speech as training data. As a continuation of enriching the DA datasets, with fine-grained analysis of dialectal Arabic speech in the MGB-5 challenge,<sup>3</sup> the ADI-17 dataset was released. This includes about 3,000 hours of dialectal Arabic crawled from YouTube covering 17 Arabic-speaking



Unlike English, speech recognition in Arabic faces many challenges due to its rich dialectal variety, with modern standard Arabic (MSA) being the only standardized dialect.

countries. Furthermore, another 58 hours of speech, manually annotated, was provided as development and test sets of ADI-17. This is currently the largest available spoken dialectal Arabic dataset for ADI. Figure 2 shows mapping between ADI-17 and ADI-5. Further data on available speech for spoken dialectal Arabic is provided in the accompanying table.

**Dialectal systems:** To design the **dialect identification system**, the current state of the art adopts many approaches developed for speaker and language recognition.

The task is explored using different supervised and semi-supervised architectures covering diverse topics ranging from domain adaptation to end-to-end learning.

Attention also was given to linguistic feature extraction. For acoustic representation, techniques ranged from simple frame-level spectral features to i-/x-vectors latent representations. Many researchers explored different data augmentation techniques, such as speed and volume perturbation, to increase the diversity and amount of training data. Others exploited approaches such as time-scale modification for balancing the low-resource dialect. The most popular architecture, based on the latest MGB-5 ADI challenge, is convolutional neural network. A high performance of around 95% accuracy is seen in the broadcast domain (the best result from the MGB-5 challenge) using this CNN architecture.

The accuracy of such a model shows closeness between some dialects with shared features while also highlighting the discriminating

characteristics between different dialectal forms of Arabic. Figure 3 shows the confusion matrix for the 17 Arabic dialects. However, this performance can differ across different channels; for example, the performance of modeling narrowband telephony data is less accurate than the broadband broadcast domain. With respect to deploying ADI in ASR, it is argued both ways: building a different ASR system for each dialect versus combining all dialects in a single unified system.<sup>7</sup>

### Arabic Speech Recognition by Human and Machine

In English, **enough** is the correct spelling, and **enuf** is a wrong one, although its pronunciation is close enough. In Arabic, MSA transcription has twice as many errors as in English. This is mainly due to the lack of diacritization, which causes problems particularly in determining the location of the vowels. The high degree of complexity in Arabic morphology causes a high degree of affixation. English reports a 5.8% Human error rate,<sup>16</sup> while broadcast news in Arabic has 10%.<sup>9</sup>

The phrase “as I told you” can have more than 20 lexical representations in DA. In the DA speech-recognition challenges, authors reported that four native speakers had 15% inter-annotation disagreement for the Egyptian dialect<sup>1</sup> and 40% inter-annotation disagreement for Moroccan dialects.<sup>3</sup> This is the upper-bound WER for a perfect ASR system, assuming WER is an appropriate metric to evaluate dialectal ASR.

The speech team at the Qatar Computing Research Institute (QCRI)



and the Massachusetts Institute of Technology Computer Science and Artificial Intelligence Lab (MIT CSAIL) have built several Arabic ASR systems. They explored grapheme and phoneme representations for acoustics units and hidden Markov models with time-delay neural network for acoustic modeling; for language units, they investigated word, character, and word-pieces; and for language modeling, N-Gram and recurrent neural networks have been studied.<sup>10,11,13</sup>

Recently, researchers from Kanari AI, QCRI and John Hopkins University developed an end-to-end multi-dialect Arabic ASR system using transformer architectures. Their research led to 12.5%, 27.5%, 33.8% WER; a new performance milestone for the broadcast news, the Egyptian, and the Moroccan ASR challenges respectively. It was noticeable that the mistakes produced by the end-to-end transformer showed high similarity with the expert linguist transcription. However, their results suggest that human performance in the Arabic language is still considerably better than the machine, with an absolute WER gap of 3.6% on average.<sup>9</sup>

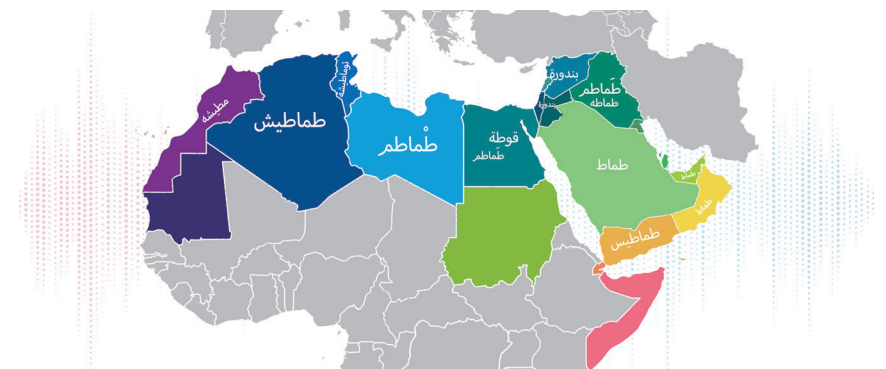
The WER for MSA achieved less than 13% in the MGB-2. This is now trusted by multiple users, such as the Al Jazeera network, BBC media-monitoring, and many government entities, thanks to the 1,200 transcribed hours and the 130 million words shared by the MGB-2 challenge, which is the case of broadcast news. Nevertheless, the wild DA is still lagging behind, with an average of 28% WER in MGB-3 and 34% in MGB-5.

**Code-switching** is one of the main challenges in spoken DA. In an eight-hour corpus collected over two days of meetings of the United Nations Economic and Social Commission for West Asia (ESCWA) in 2019, it was observed that more than 2.5 hours of the collected speech demonstrated intrasentential code-switching, where the alternation between Arabic and English is happening within the same sentence. In some cases, as with Algerian, Tunisian, and Moroccan native-speakers, the switch is between Arabic and French. In cases like this, there is an urge for a bilingual ASR, rather than just a robust multilingual ASR for the dialectal Arabic content.

**Most notable multi-dialectal Arabic speech datasets. With \* representing the datasets freely available.**

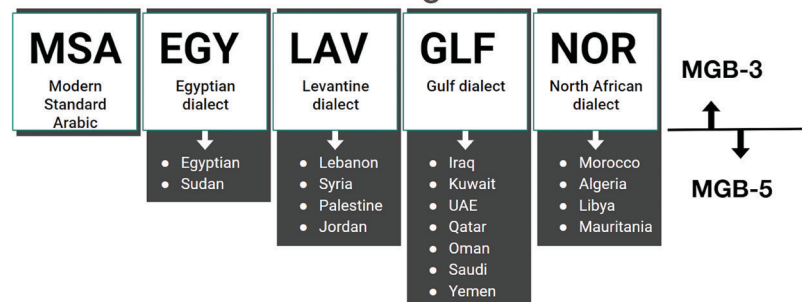
Datasets	Channels	Dialect Labels	Duration
<b>Regional/Country Level Dialect Variation</b>			
ADI-5*	Broadcast News	5 (Regional)	74h
VarDial2018* (only test set is available)	Multimedia (YouTube)	5 (Regional)	26h
GALE Phase 2 Arabic Broadcast Conversation Speech	Broadcast News	2 (MSA vs Dialect)	251h
Multi-Language Conversational (Telephone Speech 2011)	Telephone	4 (Regional)	117h
NIST LRE 2017 (most recent from the series)	Telephone	4 (Regional)	-
ADI-17*	Multimedia (YouTube)	17 (Arabic countries)	3091h
<b>Within Dialect Variation</b>			
TARIC	Recorded conversation	1 (Tunisian)	20h
KALAM'DZ	Multimedia (YouTube, Online Radio and TV)	8 Algerian dialectal variation (Hilali-Saharan, Hilali-Tellian, High-plains, Ma'qilian, Sulaymite, Algiers Blanks, Sahel-Tell, and Pre-HilaliTunisian)	104h
AMCASC	Telephone	5 Algerian cities (Constantine, Oran, Algiers, Kabyle, and Saharian)	88h

**Figure 1. Sample lexical variations for a single word in dialectal Arabic across countries in the Arab region.**



**Figure 2. Mapping between ADI-5 and ADI-17.**

• **Previous datasets has 5 regional dialect class**



**-> Not enough to cover Arab world**

Figure 3. Confusion matrix for the ADI-17 challenge, with an overall accuracy: 82%.

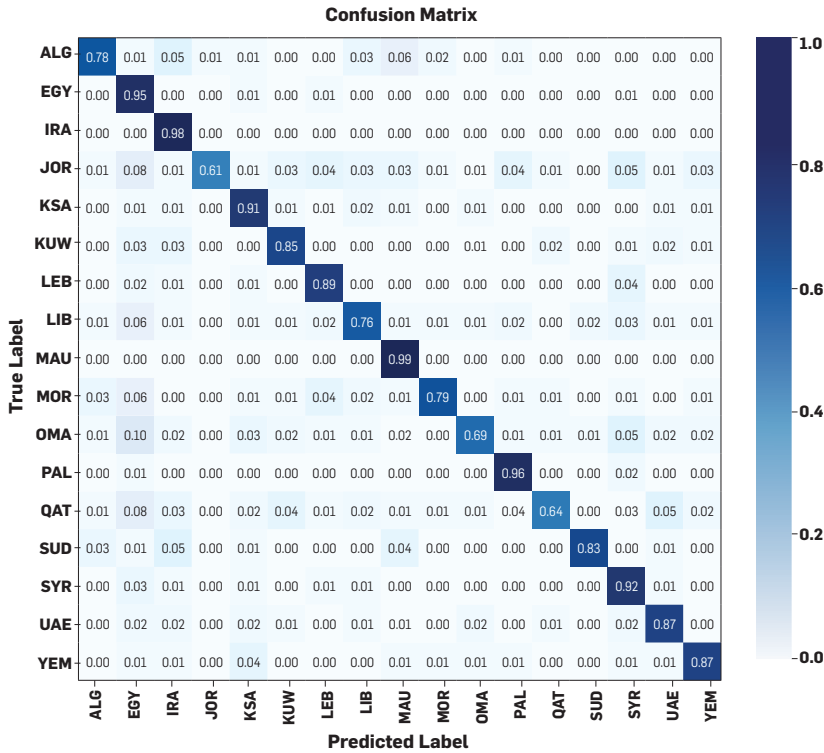


Figure 4. Example shows the integration of speech recognition, dialect identification, MT and NLP stacks happening in real time.

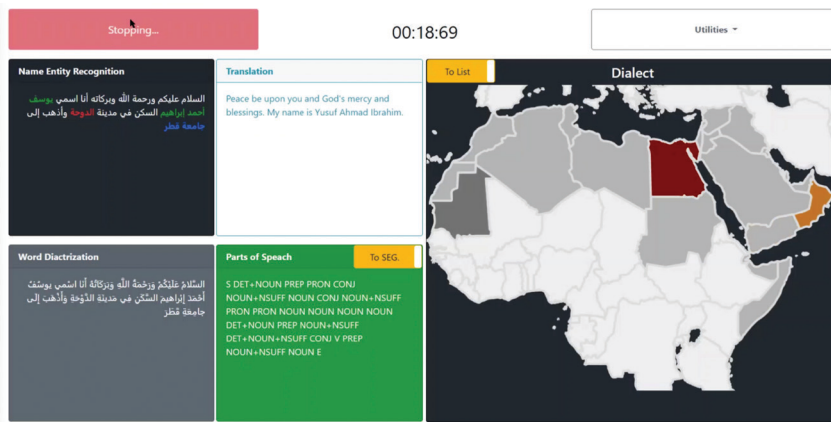


Figure 5. The first ArabicSpeech meeting in 2019 at QCRI.



**Connecting dialectal speech with NLP modules:** Figure 4 shows an application for dialectal speech processing pipeline; FarSpeech<sup>6</sup> is a system that integrates multiple language technologies starting from the voice. Initially, the application will detect which Arabic country the speaker is from using acoustic and lexical features, and based on the chosen dialect, it will choose the ASR system to convert speech into text. Secondly, this application will apply the natural language processing (NLP) stack, as part of speech tagging, named entity recognition, and vowelization restoration. Finally, the recognized text is translated into English.

**The Arabic Speech Community**  
 Recently, there has been an effort to connect the research community working on Arabic speech. *ArabicSpeech*,<sup>a</sup> founded in 2018, is an emerging research community for the benefit of Arabic speech science and speech technology. ArabicSpeech is concerned with technologies in speech and language processing focusing on both standard and dialectal Arabic language. ArabicSpeech aims to help build innovation and technology capacities for the Arabic language. It focuses on tackling large-scale computing challenges that address real priorities impacting people’s lives. The community aims to push the boundaries of Arabic speech technologies.

**Before 2018:** Earlier efforts for Arabic speech recognition were uncoordinated. In particular, there were three important projects:


- ▶ The 2002 Johns Hopkins Summer Workshop<sup>12</sup> focused on Arabic ASR.
- ▶ The GALE project was funded by DARPA to produce a system able to automatically transcribe, translate, and summarize multilingual newscasts. Arabic speech recognition was one of the core technologies of the GALE project, which was mainly concerned with Arabic broadcast news and broadcast conversation.

- ▶ The MGB challenges are evaluations of speech recognition, speaker diarization, dialect identification, and lightly supervised alignment using TV

a <https://arabicspeech.org/>



approaches to improve speech recognition systems using unlabeled data.

In summary, there is a need to build a sizeable dialectal speech corpus in the wild, and to investigate techniques for dialectal speech processing. 

#### References

1. Ali, A. et al. Speech recognition challenge in the wild: Arabic MGB-3. *IEEE Automatic Speech Recognition and Understanding Workshop*, (2017), 316–322.
2. Ali, A. et al. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. *IEEE Spoken Language Technology Workshop* (2016), 279–284.
3. Ali, A. et al. The MGB-5 challenge: Recognition and dialect identification of dialectal Arabic speech. *IEEE Automatic Speech Recognition and Understanding Workshop* (2019), 1026–1033.
4. Badawi, E.S. et al. *Modern Written Arabic: A Comprehensive Grammar*. Routledge, 2013.
5. Bouamor, H. et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the 11<sup>th</sup> Intern. Conf. Language Resources and Evaluation* (2018).
6. Eldesouki, M. et al. *FarSpeech: Arabic Natural Language Processing for Live Arabic Speech*. INTERSPEECH (2019), 2372–2373.
7. Elfeky, M.G. et al. Multi-dialectal languages effect on speech recognition: Too much choice can hurt. *Procedia Computer Science* 128, (2018), 1–8.
8. Habash, N.Y. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies* 3, 1 (2010), 1–187.
9. Hussain, A. et al. Arabic Speech Recognition by End-to-End, Modular Systems and Human; arXiv (Jan. 2020).
10. Khurana, S. et al. DARTS: Dialectal Arabic Transcription System. arXiv:1909.12163 (Sep. 2019).
11. Khurana, S. and Ali, A. QCRI advanced transcription system (QATS) for the Arabic Multi-Dialect Broadcast media recognition: MGB-2 challenge. 2016 IEEE Spoken Language Technology Workshop (SLT) (San Diego, CA, Dec. 2016), 292–298.
12. Kirchhoff, K. et al. Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop. In *Proceeding of the 2003 IEEE Intern. Conf. Acoustics, Speech, and Signal Processing*.
13. Najafian, M. et al. Automatic speech recognition of Arabic multi-genre broadcast media. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (2017), 353–359.
14. Olive, J. et al. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media, 2011.
15. Pallett, D.S. A look at NIST's benchmark ASR tests: past, present, and future. 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721) (2003), 483–488.
16. Xiong, W. et al. 2016. Achieving human parity in conversational speech recognition. arXiv preprint arXiv:1610.05256. (2016).

**Ahmed Ali**, Qatar Computing Research Institute, HBKU, Qatar.

**Shammur Chowdhury**, Qatar Computing Research Institute, HBKU, Qatar.

**Mohamed Afify**, Microsoft Advanced Technology Lab, Egypt.

**Wassim El-Hajj**, American University of Beirut, Lebanon.

**Hazem Hajj**, American University of Beirut, Lebanon.

**Mourad Abbas**, Centre de Recherche Scientifique et Technique pour Le Développement de la Langue Arabe, Algeria.

**Amir Hussein**, Kanari AI, UAE.

**Nada Ghneim**, Arab International University, Syria.

**Mohammad A.M. Abushariah**, King Abdullah II School of Information Technology, The University of Jordan, Jordan.

**Assal Alqudah**, Taibah University, Saudi Arabia.

Copyright held by authors/owners.

recordings and YouTube data. MGB-2 focused on broadcast news MSA data, while MGB-3 and MGB-5 focused on dialectal Arabic speech challenges in the wild using YouTube multi-genre speech content.

These have been great initiatives, with good research impact, but not sustainable as they rely on timely funded projects, such as DARPA, SUMMA (an EU-funded project),<sup>b</sup> or a co-located challenge, as with one of the speech conferences.

**ArabicSpeech:** In 2018, ArabicSpeech was founded by an advisory board with a mixture of researchers from academia and industry: QCRI, Johns Hopkins University, University of Edinburgh, Google, and Microsoft. This mix of academia and industry reflects the challenges and activities organized by ArabicSpeech. The community organizes an annual workshop to discuss ongoing projects related to Arabic speech. We created a special interest group in the International Speech Communication Association (ISCA). The platform now gives a one-stop location for young researchers and practitioners interested in Arabic speech. Since founding it, most of the speech teams in major companies have become interested in helping the community, with more than 300 researchers worldwide joining the ArabicSpeech community. Today, more than 4,000 hours have been shared on the ArabicSpeech resource platform. Figure 5 shows the first meeting of ArabicSpeech, organized in 2019.

At the time of writing this article,

<sup>b</sup> <http://summa-project.eu/>

we can recognize growing labs working on Arabic speech recognition-related challenges in the Arab world, including QCRI in Qatar; Kanari AI in Qatar and UAE; Microsoft Advanced Technology Lab, RDI, and Cairo University in Egypt; American University of Beirut in Lebanon; Ecole Normale de Bouzareah and CRSTDLA (The Scientific and Technical Research Center for the Development of the Arabic Language) in Algeria; HIAST and Damascus University in Syria, and King Abdullah II School of Information Technology, The University of Jordan.

#### What's Next?

There are many gaps remaining in making Arabic ASR practically useful, such as home assistance devices that can understand and speak Arabic dialects. State-of-the-art ASR systems have been trained on tens of thousands of hours transcribed verbatim, or at least semi-supervised, by choosing the most likely word sequence for each utterance. Unfortunately, Arabic dialects still need large corpora for every dialect. Most of the state-of-the-market systems are using broadcast speech data such as GALE, MGB, or the SUMMA corpus, which are mainly formal speech. In addition to the development of large-scale datasets, there is a need for a holistic approach to combine different perspectives of multi-dialect Arabic speech processing systems capable of recognizing dialect, converting it to text via ASR, and finally communicating it back in dialectal Arabic—dialectal text to speech. It is arguably convincing that self-training and unsupervised pretraining have emerged as effective