ChatGPT

Open-Source Speech Models for Dialectal Arabic (Levantine & Code-Switched)

Code-Switched Arabic Speech Recognition (Arabic + English/French)

Several projects tackle **code-switching** in Arabic speech, where speakers mix Arabic with English or French. A notable example is *Arabic-Whisper CodeSwitching Edition*, a fine-tuned OpenAI Whisper model that transcribes Arabic audio containing English words [1] [2]. This model, built on Whisper-Large-v2, was trained on a 12.5k-utterance Arabic-English dataset (from YouTube) to better handle multilingual conversations [3] [4]. It's open-source (GPL-3.0) and available on Hugging Face with a demo, making it immediately usable for transcribing everyday "kitchen table" Arabic that slips into English phrases [2]. Fine-tuning Whisper on such mixed data dramatically improves recognition of code-switched speech, and users can further fine-tune it with their own audio if needed (the training code and dataset are provided in the model repo).

Another advanced prototype comes from research on **Tunisian Arabic code-switching**. Salah Zaiem *et al.* released a **Code-Switched Tunisian ASR** model that can process Tunisian dialect mixed with French and English [5]. Their approach uses three specialized wav2vec2 models (for French, English, and Tunisian Arabic) whose outputs are fused in a "mixer" to produce the final transcription [6]. This system achieved a word error rate ~29.5% on their TunSwitch test set [7] – a strong result given the unpredictability of spontaneous code-switching. The project is open-sourced (Apache-2.0) with a Hugging Face demo and even a **Zenodo dataset** release of Tunisian audio for community use [8]. It demonstrates that with sufficient data and clever modeling, **Arabic–French–English** speech (common in North Africa) can be handled in an end-to-end pipeline. Fine-tuning on other dialects is possible: for example, one could extend this architecture to Lebanese by substituting a Levantine Arabic model and training on Arabic-French code-switched clips.

It's worth noting that **Arabic–French mixing** has been studied in Algeria and Tunisia. The newly released CAFE corpus focuses on Algerian dialect with French/English code-switches [9]. It contains ~37 hours of *in vivo* conversational speech with rich phenomena like overlapping talk, filler words, and code-switch points annotated [10]. Using off-the-shelf models on such data is challenging – Whisper-large v2 initially struggled (WER > 50%) until researchers applied advanced decoding to bring WER down to ~53.8% and Mixed-Error-Rate 0.310 [11]. This underscores that real **bilingual voice notes** (e.g. WhatsApp-style chatter) require domain-specific adaptation. The good news is that open models like Whisper or wav2vec2 can be fine-tuned on these corpora. For instance, Mohamed Rashad's code-switch model above was created by fine-tuning Whisper on a custom bilingual dataset [3]. As more code-switched corpora emerge (e.g. the **SCC22** Saudi Arabic-English test set and the CAFE/Zenodo datasets), we can expect improved models covering Franco-Arabic mixes. In summary, the community is actively building **deployable code-switched ASR** for Arabic, and these models can be further refined with new data (many authors provide training scripts and data sources for fine-tuning).

Levantine Dialect STT (Lebanese/Syrian Arabic)

Levantine Arabic (e.g. Lebanese, Syrian dialects) has historically been under-served by ASR systems focused on Modern Standard Arabic. However, recent open-source efforts are tackling this gap. For example, on Hugging Face there is a **Lebanese Arabic STT** model (by ali-issa) which fine-tuned Facebook's XLSR-53 wav2vec2 on Lebanese speech [12]. It reportedly reached a word error rate of ~41% on its evaluation set [12] – not yet low enough for production, but a promising start for a purely colloquial model. This indicates that with ~tens of hours of Lebanese audio, wav2vec2 can be adapted to the accent and vocabulary of Lebanon. Fine-tuning is absolutely possible: XLSR-53 and Whisper are pretrained on multilingual audio (including some Arabic), so they just need dialectal transcribed audio to specialize. Researchers have shown that even a few dozen hours of targeted data can significantly improve dialect ASR [13] [14]. For instance, the MGB-2 broadcast dataset (1,200 hours) included Levantine speech, and multi-dialect models trained on such data can transcribe Levantine, Gulf, Egyptian, and North African speech in one system [13]. In practice, if you have Lebanese or Syrian audio (e.g. WhatsApp voice notes with transcripts), you can fine-tune these transformer models to better handle the local accent and code-switch habits.

Beyond individual models, there are **toolkits and initiatives** supporting dialectal ASR. The Arabic ML community *ARBML* developed a toolkit called **Klaam** that wraps state-of-the-art models for easy training/inference [15] [16]. It comes with pre-trained ASR for Modern Standard Arabic and Egyptian Arabic, and even a dialect-classification model that can identify dialect labels (Egyptian, North African, Levantine, Gulf, or MSA) from speech [17]. This means one could first detect that an audio clip is, say, *Levantine Arabic*, then route it to a Levantine-optimized ASR model. Tools like NVIDIA NeMo and SpeechBrain provide similar frameworks to train speech models; indeed, open Conformer models have been trained on Arabic CommonVoice using NeMo (e.g. a Conformer-CTC model with ~40% WER on test) [18]. Academic labs are also releasing multi-dialect resources – for example, the recent **Casablanca corpus (2024)** provides transcribed audio for eight dialects including Jordanian and Palestinian Arabic [19]. Although only ~2 hours per dialect (validation/test sets) are publicly released, it's a valuable evaluation set for Levantine ASR. Likewise, the QASR dataset (by QCRI) spans ~2,000 hours across dialects [13], and the **ADI-5 corpus** (50 hours from Al-Jazeera) included Levantine speech for the MGB-3 challenge [20]. These datasets (often free for research) enable fine-tuning and benchmarking of Levantine models. In summary, **Levantine STT** is becoming more feasible: you can start with a multilingual ASR model and fine-tune on Levantine data (some community sets or self-collected transcripts), or leverage emerging pre-trained dialect models from open-source hubs. The key is that the model architecture (transformer or conformer encoders) can accommodate dialectal vocabulary and pronunciation – given the right training data.

Emotion, Prosody, and Cultural Nuance in Arabic Speech

Capturing **emotional expression** in Arabic speech goes beyond plain transcription. While most ASR models output just words, researchers have begun integrating prosody and emotion recognition into speech pipelines. There are a few open datasets targeting Arabic **speech emotion recognition (SER)**, which can be used to train classifiers or augmented ASR systems. For example, the **BAVED** corpus (Basic Arabic Vocal Emotions Dataset) contains recordings of Arabic words spoken with different levels of emotion – from neutral to high arousal (happiness, sadness, anger, etc.) [21]. Another dataset, the **Arabic Natural Audio Dataset (ANAD)**, was compiled from talk-show phone calls, labeled with three broad emotions (happy, angry, surprised) in a more natural conversational context [22]. These corpora are small (on the order of 1–2 hours of speech each), but they allow training models to detect **emotional prosody**. In fact, researchers have successfully fine-tuned wav2vec2 and HuBERT models to classify emotions on BAVED with high accuracy (≥85%) [23]. There's also the **KSU Emotions** dataset from King Saud University – ~5 hours of acted emotional speech in Modern Standard Arabic (speakers from Yemen, KSA, Syria) covering several emotions [24]. On the more spontaneous side, the recent *SERTUS* corpus contains ~23 hours of Tunisian dialect speech from YouTube, annotated for emotions like anger, satisfaction, frustration, etc., highlighting the interest in **somatic and affective expressions** in dialectal contexts (though SERTUS is a research dataset, not yet broadly open).

Toolchains for emotion: Projects are emerging that combine ASR with emotion metadata. For instance, one open GitHub project fine-tuned a wav2vec2 model to **transcribe and classify emotions simultaneously** (using BAVED and others) [25]. While no off-the-shelf ASR today will output "[angry voice]" in the transcript, it's technically possible to build a multi-task model or a two-stage pipeline (ASR + emotion classifier). Some speech toolkits (like *SpeechBrain*) support emotion recognition as a task, which could be applied to Arabic given training data. In practice, a system could tag each sentence with an emotion label or confidence. For example, after transcription, an Arabic **emotion model** could analyze the audio segment to tag it as distressed, calm, or excited. This is useful for interpreting **somatic idioms of distress** – culturally specific phrases that convey emotion. A phrase like "قلبي محروق" ("my heart is burning") might be transcribed verbatim, but an emotion-aware NLP layer could recognize it as an idiom for severe sadness. Some datasets explicitly mark laughter, crying, or sighs in transcripts (e.g. `[بكاء]` for crying is present in the Casablanca corpus) [26], providing cues to train on. By leveraging those, advanced models could insert tags or metadata for such events.

It's important to note that **cultural nuance in interpretation** often requires combining ASR with language understanding. The ASR will faithfully capture the idiom "ظهري مكسور" ("my back is broken"), but understanding that the speaker is expressing despair (not physical back pain) is a step for a natural-language or sentiment model. There are ongoing research efforts in Arabic NLP to handle idiomatic and metaphorical language – these could be integrated with speech systems. For now, one can fine-tune a language model on Arabic dialect text to identify idioms of distress, and apply it to the ASR output. In sum, while no single open-source model yet **fully understands cultural idioms or emotions** out-of-the-box, the building blocks are available: open Arabic emotion datasets, pre-trained Arabic acoustic models, and NLP tools. A developer can combine these to build a pipeline that **transcribes voice notes and tags emotional tone or key idiomatic phrases**, improving the interpretation of Lebanese or Syrian speakers' true feelings.

Data Resources and Ongoing Efforts for Fine-Tuning

Finally, a wealth of **audio datasets** and community efforts exist to support fine-tuning in these areas. On Hugging Face, collections like *MohamedRashad's Arabic Speech Datasets* aggregate resources such as MASC (a 1,000-hour Arabic corpus), the multilingual **MGB-2/MGB-3** challenge sets, Common Voice Arabic, and code-switch corpora [27] [28]. Many are available under open licenses (CC BY-NC or similar) for research. For example, MGB-2 (from 2016) contains Arabic TV recordings covering **MSA + dialects** (Levantine, Gulf, etc.), which was crucial for training early multi-dialect models [13]. Newer large-scale sets like **QASR** (2,000 hours of mixed sources) and **MGB-5** (Moroccan Arabic radio, ~16 hours) were introduced to evaluate generalization across dialects [13] [29]. If you need **Levantine data** specifically, you might leverage LDC's CallHome Levantine conversations (if accessible), or smaller open sets (e.g. the *Casablanca* corpus' Jordanian/Palestinian portions [19], or any Levantine clips in Common Voice). For **emotion and prosody**, the datasets mentioned (BAVED, ANAD, KSUEmotions) are available for download (some via Kaggle or university sites) [21] [30], and can be used to fine-tune models for emotion classification.

Crucially, **fine-tuning is very much possible** and is the recommended path to adapt ASR models to your target dialect or task. The open-source models highlighted here (Whisper variants, wav2vec2, Conformers, etc.) typically provide checkpoints and training scripts. For instance, the Arabic-Whisper code-switch model's author shared a YouTube data scraper and his fine-tuning process [31]. The Tunisian code-switched ASR team released part of their training data and an arXiv preprint explaining their unsupervised data augmentation strategy [32] [33]. With toolkits like Hugging Face Transformers, **you can fine-tune a model on your own audio** – many community models were built in exactly this way (e.g. fine-tuning XLSR on a specific dialect for a few epochs). Moreover, initiatives like the **Open Arabic ASR Leaderboard** [34] [35] actively benchmark these models on multiple dialects, which means as new models are fine-tuned and shared, they'll be tested and improved by the community. In conclusion, the ecosystem for **spoken Arabic dialect tech** is rapidly growing: from ready-to-deploy models on Hugging Face for Lebanese or code-switched speech, to research prototypes integrating emotion, and extensive datasets for customization. This enables developers and researchers to build speech systems that *truly understand the expressive, code-mixed nature of Levantine colloquial Arabic*.

Sources: Recent open-source models and datasets for Arabic speech and dialects, including Hugging Face model cards and academic publications [1] [5] [17] [22], as well as community repositories and benchmark studies on multi-dialect Arabic ASR [13] [26].

---

[1] [2] [3] [31] MohamedRashad/Arabic-Whisper-CodeSwitching-Edition · Hugging Face
https://huggingface.co/MohamedRashad/Arabic-Whisper-CodeSwitching-Edition

[4] MohamedRashad/arabic-english-code-switching · Datasets at Hugging Face
https://huggingface.co/datasets/MohamedRashad/arabic-english-code-switching

[5] [6] [7] [8] [32] [33] SalahZa/Code_Switched_Tunisian_Speech_Recognition · Hugging Face
https://huggingface.co/SalahZa/Code_Switched_Tunisian_Speech_Recognition

[9] [10] [11] [2411.13424] CAFE A Novel Code switching Dataset for Algerian Dialect French and English
https://arxiv.org/abs/2411.13424

[12] ali-issa/lebanese-stt · Hugging Face
https://huggingface.co/ali-issa/lebanese-stt

[13] [14] [29] [34] [35] Open Universal Arabic ASR Leaderboard
https://arxiv.org/html/2412.13788v1

[15] [16] [17] [20] GitHub - ARBML/klaam: Arabic speech recognition, classification and text-to-speech.
https://github.com/ARBML/klaam

[18] MostafaAhmed98/Conformer-CTC-Arabic-ASR · Hugging Face
https://huggingface.co/MostafaAhmed98/Conformer-CTC-Arabic-ASR

[19] [26] UBC-NLP/Casablanca · Datasets at Hugging Face
https://huggingface.co/datasets/UBC-NLP/Casablanca

[21] [22] [23] [24] [30] GitHub - OmarMohammed88/AR-Emotion-Recognition: An implementation of the paper titled "Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset" https://journals.scholarpublishing.org/index.php/TMLAI/article/view/11039
https://github.com/OmarMohammed88/AR-Emotion-Recognition

[25] [2110.04425] Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset
https://arxiv.org/abs/2110.04425

[27] [28] Arabic Speech Datasets - a MohamedRashad Collection
https://huggingface.co/collections/MohamedRashad/arabic-speech-datasets