SPEECH DATA FOR AI · MAGHREB FIRST

Arabic dialect speech data, recorded where each dialect is spoken.

Consented, native-transcribed speech for training and evaluating speech recognition, translation and voice assistants — including the natural French–Arabic code-switching people use in everyday conversation.

Original data Recorded first-hand, not scraped.	**Adult contributors** Written consent before recording.	**Native review** Dialect-specific transcription and QA.

Current dialect status

Dialect	Status	Scope
Tunisian Arabic · ar-TN	Pilot in preparation	Core dialect; French–Arabic code-switching included where specified.
Libyan Arabic · ar-LY	Next · on request	Available on request, subject to scope, local coordination and client need.
Algerian Arabic · ar-DZ	Phase 2 · on request	Available on request, subject to scope, local coordination and client need.

Other dialects and languages may be scoped on request. The roadmap includes Moroccan and Egyptian Arabic, then Gulf and Levantine Arabic, and selected francophone West African languages, subject to local coordination, legal review and client need.

Services

Speech collection	Scripted, prompted and spontaneous conversations; speaker and environment profiles agreed to the buyer specification.
Transcription & annotation	Native-dialect verbatim transcription, speaker turns, time alignment, code-switching, non-speech events and agreed QA tags.
LLM data & evaluation	Arabic-dialect tasks, evaluation, translation verification and other language-data work scoped to an agreed brief.
TTS voice data	Clean speech recordings for voice synthesis only where contributors give separate, explicit permission.
Image & Arabic OCR	Arabic text-in-image collection and annotation, including documents and handwriting, on request.

TECHNICAL & DELIVERY OPTIONS

Built around your specification.

The following is a standard starting point. Capture settings, transcript schema, naming and delivery can be agreed for each project.

Recording and delivery

Recording sessions	Paired conversation (usually 2 speakers, sometimes 3) plus short read prompts; about 50–60 minutes recorded within a 90-minute slot.
Capture master	WAV, 48 kHz, 32-bit float (or 24-bit), uncompressed, one mono track per speaker, subject to the agreed protocol.
Client delivery	WAV 16-bit PCM mono at 16 kHz by default for speech recognition; 44.1/48 kHz for clean or TTS data and 8 kHz for telephony can be agreed.
Transcripts	Dialect-native verbatim Arabic script; French/English in Latin script as spoken; time-aligned JSON, TSV or TextGrid as agreed.
Metadata & transfer	JSON or CSV manifest; secure SFTP or the client's own cloud bucket; batch reports and folder structure agreed in advance.
Quality control	Audio checks for clipping, noise, silence, overlap and duplicates; native-speaker review and a second-pass sample check against agreed targets.

Contributor rights and handling

Projects are designed for adult participants (18+), with written consent before recording, pseudonymous speaker IDs, and identity/contact details kept separate from audio and transcripts. Voice synthesis or cloning requires a separate opt-in. Project-specific consent scope, data protection arrangements, acceptance criteria and licence terms must be confirmed in writing before collection.

Engagement process

1. Share the language, use case, target duration, schema and timeline.	**2.** Request a free 30-minute Tunisian sample under NDA, subject to confirmation of scope and availability.
3. Agree a pilot scope, technical specification, rights and acceptance criteria.	**4.** Move to production after the pilot is reviewed and approved.

Leadership

Mourad Ben Chedli
Founder · Project lead
PMP-certified project manager. 20 years in sales, projects and operations across Africa, the Middle East, Europe and Canada, including IT projects for the Government of Canada.

Ichrak Gharbi
Linguistic lead
Arabic language teacher, University of Manouba graduate. Leads transcription guidelines, reviewer training and dialect quality control.

Barhoumi Mustapha
Chief Technical Officer (CTO)
Web development across PrestaShop, PHP, JavaScript, HTML/CSS and APIs; e-commerce, data automation, scraping, performance, SEO and databases.

CONTACT

contact@arabicaivoice.com
arabicaivoice.com/contact.html
Brussels · Tunis · Dubai

Arabic AI Voice is a business line of Altis CHR Solutions SRL, a Belgian company based in Brussels. Company no. 1036.762.922 · VAT BE 1036.762.922. Scope, availability, sample terms, local authorisations and project-specific licences are subject to written confirmation. No third-party certifications or unverified performance claims are stated in this sheet.