EgypDisSpeech-9:
A Pilot Dataset and Feasibility Study of Egyptian Arabic Non-Standard Speech with Motor Speech Disorders

Islam Alsohby, Ahmed Hani

Abstract

Speech technologies rarely include recordings from Arabic speakers with motor speech disorders, which limits both clinical research and accessibility tools for Arabic speakers. We present **EgypDisSpeech-9**, a pilot feasibility study describing the collection protocol, metadata, and initial descriptive statistics for a small Egyptian Arabic dataset recorded from **nine** volunteer participants with non-standard speech (e.g., dysarthria related to ALS and other etiologies). The goal is not to release raw audio (IRB not currently obtained) but to report our methods, ethical safeguards, and lessons learned to guide future scaled collections and to encourage the community to develop Arabic clinical speech resources. We hypothesize that even a small, well-documented pilot will reveal practical barriers (recruitment, annotation alignment, recording variability) that must be addressed before large-scale data release. We conclude with recommendations and a plan to obtain formal ethics approval for an expanded, shareable dataset.

Keywords: Egyptian Arabic, Disordered Speech, Dysarthria, ALS (Amyotrophic Lateral Sclerosis), Speech Dataset, Non-standard Speech, Accessibility, Speech Recognition.

Introduction

Automatic speech recognition (ASR) and assistive communication systems typically perform poorly on atypical speech produced by motor-speech disorders such as dysarthria and ALS. Established dysarthric corpora (e.g., TORGO, UA-Speech) have enabled important progress in English ASR for disordered speech, but equivalent resources for Arabic — and specifically Egyptian Arabic — are virtually absent in the public research record. The Multi-Genre Broadcast (MGB-2) and related Arabic corpora provide broad dialectal coverage for broadcast speech, but they do not capture clinical or motor-speech disorders. The lack of Arabic disordered-speech datasets constrains evaluation and adaptation of models for Arabic speakers with communication impairments.

Purpose. This paper documents a small pilot dataset collection (EgypDisSpeech-9) and provides an explicit protocol, descriptive metadata, and ethical reflection intended to: (1) supply the research community with a reproducible pilot description, (2) surface the main operational and ethical issues in collecting Arabic clinical speech, and (3) propose concrete next steps toward a larger IRB-approved, shareable dataset.

Hypothesis. We hypothesize that collecting high-quality, consistently annotated Egyptian disordered speech is feasible but will be hampered by (a) heterogeneity in recording devices and environments, (b) difficulties in aligning transcripts to short, disfluent utterances, and (c) ethical constraints limiting public release — all of which must be explicitly addressed to produce a useful public corpus. This pilot aims to measure and document these constraints rather than to claim broad generalizability.

Methods

Participants

Nine adult native Egyptian Arabic speakers with clinically diagnosed motor-speech impairments volunteered for the pilot. Recruitment was convenience-based through local clinics and community networks. To protect privacy, participants are anonymized with numeric IDs (P01–P09). Basic non-identifying metadata were recorded: age decade, self-reported gender, condition class (e.g., ALS, post-stroke dysarthria), and a clinician-provided intelligibility category when available. No personally identifying data are included in this manuscript.

Ethical safeguards

Because institutional review board (IRB) approval was not available at the time of collection, we limited what we would publish and how we used the data. Each participant provided written informed consent for local research use, audio-only capture, and de-identified reporting. Raw audio and identifying metadata are **not released**. We followed internationally accepted human-subjects principles regarding informed consent, risk minimization, and confidentiality (Belmont Report; Declaration of Helsinki) in the collection and reporting process. The study therefore reports aggregated metadata and method details only; raw recordings remain under controlled storage until formal IRB approval is secured for sharing.

Recording protocol and materials

Recordings occurred in clinic or quiet home settings using commonly available devices (smartphone or USB headset), depending on participant mobility and access. Each session included: (1) a short consent and demographics interview, (2) read sentences from a small standardized prompt set (10–20 short Arabic sentences adapted for local dialect), and (3) spontaneous speech (open prompt asking participants to describe a recent memory for ~30–60 seconds). Recorded formats were WAV or high-quality MP3 sampled at device defaults (documented per file). File names use anonymized IDs and a timestamp. No images or video were recorded.

Annotation and metadata

A native Arabic transcriber produced orthographic transcriptions in Arabic script. For each utterance we annotated: utterance ID, speaker ID, recording device, environment (clinic/home), approximate duration, and a 1–5 intelligibility rating provided by a speech-language pathologist or trained annotator. Where clinical severity scores were available in patients' medical records, we included a coarse severity label (mild/moderate/severe) without publishing identifying clinical notes.

Analysis plan (descriptive)

Because the goal was feasibility and documentation, analysis is descriptive: number of utterances per speaker, distribution of durations, device variability, transcriber agreement on intelligibility ratings, and qualitative account of annotation challenges (e.g., deletions, overlapping coughs, phonetic idiosyncrasies). We do not report or evaluate ASR model performance on the recordings in this pilot.

Results (Pilot observations)

Participation and content. Nine participants completed sessions yielding short read and spontaneous utterances. The number of utterances per participant varied due to participant stamina and severity; sessions ranged from brief (a few minutes) to longer (~20 minutes), with shorter average utterance length than typical broadcast corpora.

Annotation challenges. Annotators reported frequent disfluencies, variable pronunciations of function words, and partial words that complicated strict orthographic transcription. Aligning short audio clips to exact transcript tokens required manual review. Inter-annotator agreement on the 1–5 intelligibility scale showed moderate agreement (qualitative observation), suggesting that multi-rater scoring and clearer rating rubrics are necessary for scale.

Operational findings. Variability in recording devices produced measurable differences in loudness and background noise; smartphone recordings were convenient but introduced compression artifacts in some MP3 files. Recruitment of clinical participants was feasible through local clinics but required flexible scheduling and hosting to accommodate fatigue. These pilot results support the hypothesis: collection is feasible but will require standardization (single-device protocol when possible), clearer annotation guidelines, and formal ethical oversight before public release.

Discussion and Recommendations

This pilot confirms that a small Egyptian Arabic dataset of disordered speech can be collected ethically and with useful metadata — provided that strong privacy procedures and future IRB approval are in place. Key recommendations before scaling:

1. **Obtain IRB / Ethics Committee approval** and implement a data governance plan specifying who can access data and under which conditions (data use agreements). International ethics guidance (Declaration of Helsinki and Belmont principles) recommends ethics review for human subjects research and protections for vulnerable populations.
2. **Standardize recording hardware and environment** where possible (e.g., provide headsets or specify device models) to reduce channel variability that would confound acoustic analyses and model training.
3. **Adopt multi-rater intelligibility and severity scales** with training to improve annotation reliability; consider including acoustic experts or clinicians for clinical labels.
4. **Design a clear sharing policy**: a staged release (metadata and transcripts publicly; audio available under controlled access) can balance openness with participant privacy.

Past dysarthric corpora (e.g., TORGO) have enabled research while maintaining necessary controls; learning from their licensing and documentation practices is advisable.

Limitations

EgypDisSpeech-9 is a pilot with nine participants and heterogeneous recording conditions; results are not generalizable. The dataset is small for model training but valuable for exposing practical barriers. Because IRB approval was not available at the time of collection, raw audio is not shared; this limits external validation but preserves participant privacy pending formal ethics approval.

Conclusion

EgypDisSpeech-9 documents an initial, ethically cautious effort to collect Egyptian Arabic speech from speakers with motor-speech disorders. The pilot demonstrates feasibility and surfaces concrete technical and ethical issues that future, IRB-approved collections must resolve. We invite collaboration with clinical partners and institutional review boards to expand the corpus and enable responsible public release for Arabic ASR and assistive technology research.

Acknowledgments

Thanks to the volunteer participants and local clinicians who advised the protocol. No commercial funding supported this pilot.

References

(Selected works cited in the text)

Ali, A., et al. (2016). The MGB-2 Arabic Multi-Genre Broadcast dataset. *Proceedings of Interspeech 2016*. Retrieved from https://arabicspeech.org

PMC. (2024). Voice signals database of ALS patients with different dysarthria (VOC-ALS). *PubMed Central (PMC)*. https://pmc.ncbi.nlm.nih.gov/

PubMed. (n.d.). Improving acoustic models for dysarthric speech: Studies demonstrating ASR challenges and adaptation strategies. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation, 46*(4), 523–541. https://doi.org/10.1007/s10579-012-9190-9

World Medical Association. (2013). *World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects*. JAMA, 310(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of*

human subjects of research. U.S. Department of Health & Human Services. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/