Arabic dialects speech corpora: A systematic review

Speech Communication • Review • 2025 • DOI: 10.1016/j.specom.2025.103322

Alqadasi, Ammar Mohammed Ali [a, b] ; Zeki, Akram M. [b] ; Sunar, Mohd Shahrizal [c, d] ; Hashim, Siti Zaiton Mohd [c] ; Sah hj Salam, Md [c] ; +1 author

[a] Department of Computer Engineering, Faculty of Computer and Information Sciences, Karabuk University, Karabuk, 78050, Turkey

Abstract

Speech processing applications are crucial in various domains, necessitating reliable speech recognition systems built upon suitable speech databases. However, the availability of comprehensive resources for the Arabic language remains limited compared to other languages like English. A systematic review was conducted to identify, analyze, and classify existing Arabic dialect speech databases. Initially, online digital databases and search engines were identified to collect a diverse range of manuscripts for thorough examination. The review encompassed 30 publicly accessible databases and an additional 39 self-databases, which were thoroughly studied, classified based on their characteristics, and subjected to a detailed analysis of research trends. This paper offers a comprehensive discussion on the diverse speech databases developed for various speech processing applications, highlighting the purposes and unique characteristics of Arabic speech databases. By providing valuable insights into their availability, characteristics, challenges, and research directions, this review aims to facilitate researchers' access to suitable resources for their

specific applications, encourage the creation of new datasets in underrepresented areas, and promote open and easily accessible databases. Furthermore, the findings contribute to bridging the gap in available Arabic speech databases and serve as a valuable resource for researchers in the field.

Author keywords

Arabic dialects; Arabic dialects corpora; Modern Standard Arabic; Speech corpus; Speech database; Speech recognition

Indexed keywords

Engineering controlled terms

Audio signal processing; Database systems; Reviews; Search engines; Speech communication; Speech processing

Engineering uncontrolled terms

Arabic dialect corpus; Arabic dialects; Modern standard arabic; Modern standards; Processing applications; Speech corpora; Speech database; Standard arabics; Systematic Review

Engineering main heading

Speech recognition

Funding details

Details about financial support for research, including funding sources and grant numbers as provided in academic publications.

Funding sponsor	Funding number	Acronym
International Islamic University Malaysia See opportunities by IIUM		IIUM
Universiti Teknologi Malaysia See opportunities by UTM		UTM

Funding text

The authors would like to thank the Malaysia International Islamic University and the Universiti Teknologi Malaysia for supporting and providing the opportunity to conduct this research work.

Corresponding authors

Corresponding author	A.M.A. Alqadasi
Affiliation	Department of Computer Engineering, Faculty of Computer and Information Sciences, Karabuk University, Karabuk, 78050, Turkey
Email address	ammarmohammed@karabuk.edu.tr

Customer Service

Help

Tutorials

Contact us

ELSEVIER

Terms and conditions Privacy policy

RELX™

View PDF