Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Creating the Hu-Int dataset: A comprehensive Arabic speech dataset for gender detection and age estimation of Arab celebrities

Hussain A. Younis [a,b,*], Nur Intan Raihana Ruhaiyem [a,*], Ameer A. Badr [c], Taiseer Abdalla Elfadil Eisa [d], Maged Nasser [e], Tien-Ping Tan [a], Nur Hana Samsudin [a], Sani Salisu [a,f]

[a] *School of Computer Sciences, Universiti Sains Malaysia, Gelugor, 11800, Penang, Malaysia*
[b] *College of Education for Women, University of Basrah, 61004 Basrah, Iraq*
[c] *Department of Information Technology, Technical College of Management-Baghdad, Middle Technical University, Baghdad, Iraq*
[d] *Department of Information Systems-Girls Section, King Khalid University, Mahayil 62529, Saudi Arabia*
[e] *Computer & Information Sciences Department, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia*
[f] *Information Technology Federal University Dutse, Jigawa, Kano, Nigeria*

ARTICLE INFO

Keywords:
Dataset
Features
Speaker
Gender detection
Age estimation
Video captioning

ABSTRACT

Speech is one of the attributes that humans enjoy. Humans possess several distinguishing characteristics, including fingerprints, hands, fingers, eyes and DNA, which set each individual apart from one another. Numerous datasets specialise in speech and include many languages from around the world, such as JNAS DATASET, TCDSA DATASET, UF-VAD DATASET, NIST SRE DATASET, a Gender Dataset, TIMIT dataset, common-voice dataset, SITW, VoxCeleb1, VoxCeleb2, Arabic-Saudi Arabic Speech Dataset-2 and MGB Challenge Dataset. This research contributes to the establishment of requirements for creating a new dataset featuring Arab celebrities in the Arabic language. The Hu-Int dataset is designed with the goal of profiling new notable Arab individuals. It includes 1017 speakers after the filtering process and 93.814 videos with various formulas (e.g. mp4, wav and CSV). The algorithms Random Forest, Logistic Regression, SVM (RBF), SVM (Linear), Decision Tree and CNN-LSTM are then used. The initial experimental results for algorithmic precision yielded the following accuracy rates: 0.88%, 0.87%, 0.96%, 0.77%, 0.85% and 0.88%, respectively, for males and 0.86%, 0.84%, 0.92%, 0.79%, 0.87% and 0.89%, respectively, for females in gender classification. The features of age were divided into six classes for gender classification. Dimensionality reduction was performed using principal component analysis and linear discriminant analysis. The results showed the opposite effect, with the SVM (RBF) and hybrid CNN-LSTM algorithms outperforming many other algorithms in both gender detection and age estimation.

1. Introduction

Speech dataset is a collection of audio recordings used to train algorithms and develop speech recognition systems. It is a key component in the research and development of natural language processing (NLP) [1]. In the last few decades, there has been a significant and rapid development in the field of automated computer use due to its information processing capabilities. Perhaps the most notable of these developments is the emergence of the field of artificial intelligence (AI), which broadened the applications of computers in daily life beyond their usual roles of performing calculations and implementing traditional programmes. Computers were used in this field to simulate and model human intelligence, such as learning, remembering, generalisation, discrimination and decision-making, and other smart qualities.

1.1. Speech

Speech is defined as a series of linguistic voices produced by the pronunciation device to express the thoughts and feelings of the speaker. There are many ways for individuals to express themselves and exchange ideas with others, such as sign, writing and speech, but the language of speech is regarded as one of the most important, renowned

* Corresponding authors.
E-mail addresses: hussain.younis@uobasrah.edu.iq (H.A. Younis), intanraihana@usm.my (N.I.R. Ruhaiyem).

https://doi.org/10.1016/j.bspc.2024.106511
Received 25 July 2023; Received in revised form 7 May 2024; Accepted 24 May 2024
Available online 2 June 2024