Data Engineering and Communications Technologies
Volume 83, 2021, Pages 25–36

Automated Pipeline for Training Dataset Creation from Unlabeled Audios for Automatic Speech Recognition (Conference Paper)

Romanovskyi, O.[a], Iosifov, I.[a], Iosifova, O.[a], Sokolov, V.[b], Kipchuk, F.[b], Sukaylo, I.[b]

[a]Ukraine Ender Turing OÜ, Tallinn, Estonia
[b]Borys Grinchenko Kyiv University, Kiev, Ukraine

Abstract

In the paper, we present a software pipeline for speech recognition to automate the creation of training datasets, based on desired unlabeled audios, for low resource languages and domain-specific area. Considering the commoditizing of speech recognition, more teams build domain-specific models as well as models for local languages. At the same time, lack of training datasets for low to middle resource languages significantly decreases possibilities to exploit last achievements and frameworks in the Speech Recognition area and limits the wide range of software engineers to work on speech recognition problems. This problem is even more critical for domain-specific datasets. The pipeline was tested for building Ukrainian language recognition and confirmed that the created design is adaptable to different data source formats and expandable to integrate with existing frameworks.

Author keywords

ASR; asynchronous graphs; automatic speech recognition; dataset creation pipeline; natural language processing; NLP

Funding details

Funding sponsor	Funding number	Acronym
Ministry of Education - Singapore	CCNU19TS022	MOE

Funding text

This scientific work was partially supported by RAMECS and self-determined research funds of CCNU from the colleges' primary research and operation of MOE (CCNU19TS022). The research team is grateful to Ender Turing OÜ for defining the business problem, comments, corrections, inspiration, and computational.

About this paper

https://link.springer.com/chapter/10.1007/978-3-030-80472-5_3

ISSN: 2194-5365
Print ISBN: 978-3-030-80472-5
DOI: 10.1007/978-3-030-80472-5_3
EID: 2-s2.0-85111941280

Source Type: Book Series
Document Type: Conference Paper
Publisher: Springer, Cham