Text-Independent Speaker Recognition Based on Adaptive Course Learning Loss and Deep Residual Network

Qinghua Zhong (✉ zhongqinghua@m.scnu.edu.cn)

South China Normal University Guangzhou Higher Education Mega Center <https://orcid.org/0000-0002-7354-6608>

Ruining Dai

South China Normal University

Han Zhang

South China Normal University

YongSheng Zhu

South China Normal University Guangzhou Higher Education Mega Center

Guofu Zhou

South China Normal University

Research

Keywords: Text-independent, Speaker recognition, Adaptive curriculum learning loss, Deep residual network, Convolutional attention statistics pooling

Posted Date: February 18th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-206450/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at EURASIP Journal on Advances in Signal Processing on July 23rd, 2021. See the published version at <https://doi.org/10.1186/s13634-021-00762-2>.

Abstract

Text-independent speaker recognition is widely used in identity recognition. In order to improve the features recognition ability, a method of text-independent speaker recognition based on a deep residual network model was proposed in this paper. Firstly, the original audio was extracted with a 64-dimensional log filter bank signal features. Secondly, a deep residual network was used to extract log filter bank signal features. The deep residual network was composed of a residual network and a Convolutional Attention Statistics Pooling (CASP) layer. The CASP layer could aggregate the frame-level features from the residual network into utterance-level features. Lastly, Adaptive Curriculum Learning Loss (ACLL) classifiers was used to optimize the output of abstract features by the deep residual network, and the text-independent speaker recognition was completed by ACLL classifiers. The proposed method was applied to a large VoxCeleb2 dataset for extensive text-independent speaker recognition experiments, and average equal error rate (EER) could achieve 1.76% on VoxCeleb1 test dataset, 1.91% on VoxCeleb1-E test dataset, and 3.24% on VoxCeleb1-H test dataset. Compared with related speaker recognition methods, EER was improved by 1.11% on VoxCeleb1 test dataset, 1.04% on VoxCeleb1-E test dataset, and 1.69% on VoxCeleb1-H test dataset.

Full Text

This preprint is available for [download as a PDF](#).

Figures

Figure 1

The proposed text-independent speaker recognition framework

Figure 2

Flow chart of residual block construction

Figure 3

The calculation diagram of the CASP layer

Figure 4

The blue line, yellow line, and green line are Softmax, AAM-Softmax and our decision boundary, respectively. m is the feature margin of AAM-Softmax. d is adaptive feature margin ranges

Figure 5

The average EER of text-independent speaker recognition at $m = 0:1 \sim 0:5$