Duolin HUANG, Qirong MAO, Zhongchen MA, Zhishen ZHENG, Sidheswar ROUTRAY, Elias-Nii-Noi OCQUAYE, 2021. Latent discriminative representation learning for speaker recognition. *Frontiers of Information Technology & Electronic Engineering,* 22(5):697-708. https://doi.org/10.1631/FITEE.1900690

Latent discriminative representation learning for speaker recognition

Key words: Speaker recognition; Latent discriminative representation learning; Speaker embedding lookup table; Linear mapping matrix

Corresponding author: Qirong MAO
E-mail: mao_qr@ujs.edu.cn
ORCID: https://orcid.org/0000-0002-0616-4431

Motivation

1. Speaker recognition (SR) is one of the most widely used biometric recognition technologies, and provides unique advantages in remote authentication.

2. The representations of different speakers are discriminative, and should preserve the correlation between different utterances from the same speaker.

3. Most methods now focus mainly on learning discriminative features from speech. However, it is still a challenging task to identify an appropriate mode to learn discriminative and relevant features in SR.

Main idea

1. Mel-frequency cepstral coefficients (MFCCs) and identity vector (i-vector) are critical features, and they are often used for speaker recognition tasks.

2. A latent discriminative representation learning (LDRL) method is designed which integrates latent relevance learning and latent discriminative learning.

3. Dictionary learning is used to find a latent representation space that can be used to specify relationships between utterances.

4. The proposed model is assessed by TIMIT and Apollo datasets.

Method

1. Using a novel objective function, we can learn latent representations that are relevant and discriminative for SR. Specifically, LDRL from supervised learning is divided into three blocks: basic latent representation learning, latent relevance learning, and latent discriminative learning.

2. The proposed embedding lookup table can be used to learn correlations between different utterances of the same speaker. At the end, latent discriminative representations with relevance are used for classification.

Major results

Architecture of the latent discriminative representation learning method:

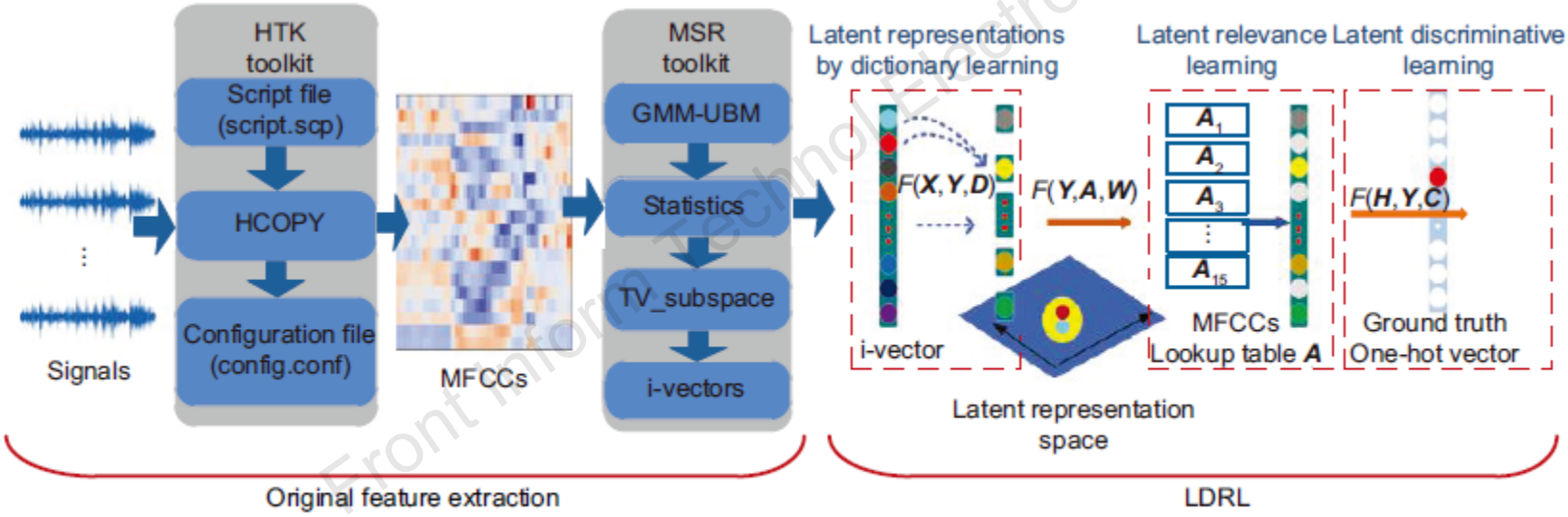

Fig. 1 Architecture of the latent discriminative representation learning method for speaker recognition based on dictionary learning

Major results (Cont'd)

Test results of our method and related methods on the TIMIT dataset with 120 speakers:

Table 7 Speaker recognition accuracy (SRA) comparison based on the TIMIT dataset for 120 speakers

Method	SRA (%)
GMM-UBM (Al-Kaltakchi et al., 2016)	95.00
CLSTM^{+} (Kumar et al., 2018)	96.31
ELM (Al-Kaltakchi et al., 2017)	96.67
UAI^{+} (Peri et al., 2019)	97.42
LDRL	**99.79**

+ denotes that the TIMIT dataset was not used as a test set in the literature. We reimplemented the code of this method and tested it on the TIMIT dataset. Best result is in bold

Major results (Cont'd)

Test results of our method and related methods on the TIMIT dataset with 630 speakers:

Table 9 Speaker recognition accuracy (SRA) comparison based on the TIMIT dataset for 630 speakers

Method	SRA (%)
GMM-UBM (Yoshimura et al., 2018)	90.79
i-vector/PLDA (Yoshimura et al., 2018)	90.95
CLSTM+ (Kumar et al., 2018)	95.13
VAE model (Yoshimura et al., 2018)	96.51
UAI+ (Peri et al., 2019)	97.10
LDRL	**99.21**

+ denotes that the TIMIT dataset was not used as a test set in the literature. We reimplemented the code of this method and tested it on the TIMIT dataset. Best result is in bold

Major results (Cont'd)

Test results of our method and related methods on the Apollo dataset:

Table 10 Top five speaker recognition accuracy (SRA) comparison based on the Apollo dataset for 183 speakers

Method	SRA (%)
Baseline (Dev)	58.17
LDRL (Dev)	91.15
Baseline (Eval)	47.00
LDRL (Eval)	83.33

Conclusions

1. A novel speaker recognition approach based on latent discriminative representation learning (LDRL) has been proposed.

2. We have demonstrated the LDRL effectiveness by comparison with other methods during testing on the TIMIT and Apollo datasets.

3. By comparing the results of experiments and visualization, it is proved that the learned representations are not only discriminative, but also relevant.

Duolin HUANG received his BS and MS degrees in computer science and technology from Jiangsu University, Zhenjiang, China, in 2017 and 2020, respectively. His research interests include multimedia analysis and pattern recognition.

Qirong MAO is currently a professor of the School of Computer Science and Communication Engineering, Jiangsu University. She received her MS and PhD degrees from Jiangsu University, Zhenjiang, China, in 2002 and 2009, respectively, both in computer application technology. Her research interests include affective computing, pattern recognition, and multimedia analysis. She has published over 50 technical articles. She is a member of IEEE.