Team61 SDSV Challenge 2020 Task2 System Description

Team61

Anonymous

anonymous

Abstract

This is the system description for text independent speaker verification in SDSV Challenge 2020 Task2. We propose a system consisting a baseline X-vector front-end with domain-specific LDA/PLDA back-end and enrollment duration based score nomalization.

Index Terms: text-independent speaker verification, x-vector, PLDA

1. Introduction

In text-independent speaker verification x-vector embeddings are commonly used as a state-of-art method because of its success in embedding speaker information to fixed size vectors. [1] However a robust x-vector embedding system requires high amount of training data. It is also experimented that embeddings from utterances of different durations carry different amount of speaker information resulting in imbalanced likelihood scoring for different durations. In this task domain training data is limited and insufficient to train a x-vector system. Our approach is that train a out-of-domain baseline x-vector embedding extraction front-end trained with large amount of data (voxceleb1,voxceleb2) to have meaningful embeddings. Then a PLDA back-end is trained on domain training data and applied to have a robust domain-specific verification system. Further score normalization is also applied, for different enrollment recording durations.

2. Dataset

Voxceleb1 [2] and Voxceleb2 [3] are used for training the front-end x-vector extractor. For both datasets, test sets are also added to train dataset in order to maximum use of data. This set is augmented with noise samples from MUSAN [4], and RIRs (Room Impulse Response and Noise Database) datasets.

DeepMine Dataset [5] Task 2 [6] Train partition is used in tranining of LDA/PLDA backend. This dataset consists of 588 speakers and Persian text-independent utterances by those speakers. Dataset is split to tranining and development set as utterances from first 525 speakers constitutes tranining set, utterances from last 63 speakers constitute development set.

In order to have a consistent test conditions with the evaluation set, development trials are generated using development set according to the enrollment duration distributions of the evaluation set.

3. Methodology

3.1. X-vector Embeddings

X-vector embeddings are basically embeddings extracted from a neural-network that is trained to classify different speakers. [1] The basic structure of a x-vector consists of frame-level layers using acoustic features; a stats-pooling layer extracting stats for given utterance from frames; a feed-forward classificaton structure producing classification output for a given utterance.

3.1.1. Features

30 MFCC's of 36 Mel-Filterbanks for 16kHz sampling rate are used as features. Features are extracted from speech portion of given utterance. Kaldi energy VAD is applied for detecting speech.

3.1.2. Neural Network Topology

E-TDNN is used as neural network topology in this task. [7] The structure is almost same with [7], except in Task2 baseline configurations, last dense layer before stats pooling is omitted. Overall network consists of a time delay layer with kernel 5 and dilation factor 3; three time-delay layers with kernel 3 and dilation factor 2,3,4 respectively; dense layers in between those time layers and following 2 dense layers before stats-pooling. This part of the network works on frame-level; after stats-pooling two dense layers followed by a softmax output layer works on chunk of frames level and completes the structure.X-vector embeddings are generally extracted from preceding layer of stats pooling(also in this task), but it also can be extracted from next dense layer that is just before the softmax output.

3.2. LDA/PLDA Scoring

After extracting x-vector embeddings a dimensionality reduction, and representation modeling are often required and improves classification significantly. For this reason commonly used LDA is used for reducing representation dimension to 150, and a PLDA model is used for scoring different embeddings by a discrimination maximizing approach.

In this task, after extracting x-vectors from the robust baseline neural-network, LDA/PLDA models are trained on Task2 domain training data in order to have a suitable verification system for the evaluation domain.

3.3. Score Normalization

X-vectors from different durations of utterances, may carry different amount of information. Resulting in different likelihood score ranges for verification with different utterance durations. In this task, evaluation authentication data is the evaluation set of Task1[] which consists of single phrase utterances that are approximately 2-3 seconds long . For this reason scores are analyzed and normalized according to different enrollment durations only . Fig.1 shows likelihood score distributions for different enrollment durations in the development set.

To normalize the effect to enrollment durations for a given model enrollment trial, mean-std normalization is applied by using mean and standard deviation of corresponding model enrollment duration distribution in the dev set.

Figure 1: *Score Distributions in Dev Set*

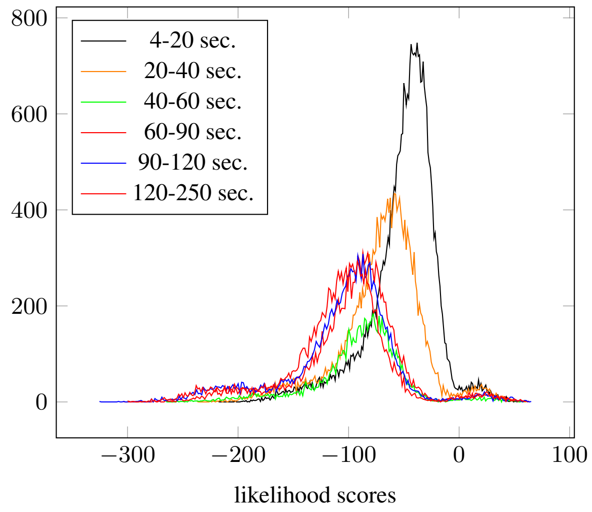


Figure 2: *Score Distributions in Eval Set*

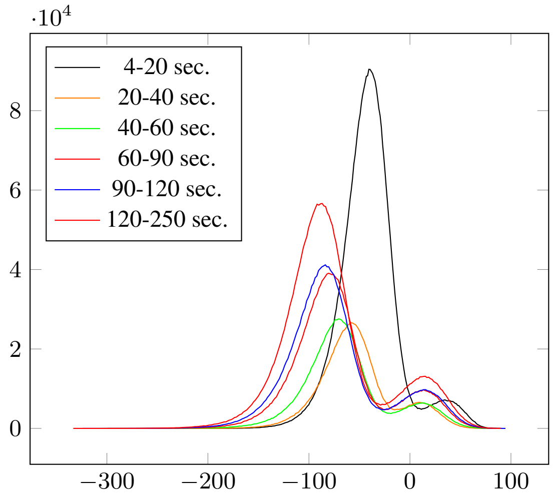
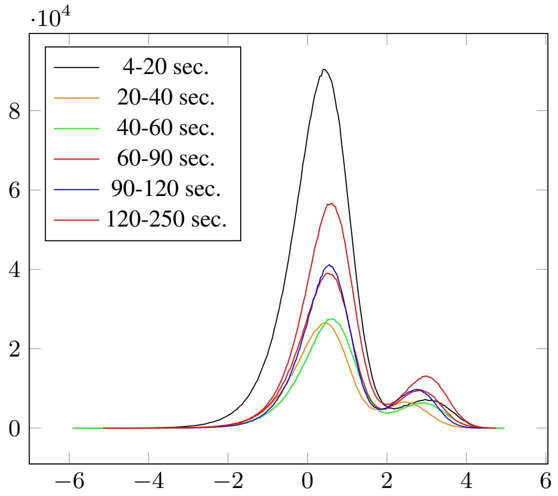


Figure 3: *Normalized Score Distributions in Eval Set*



4. Results

Table 1: RESULTS ON PROGRESS AND EVAL SET

System	Set	EER	minDCF
Baseline x-vector system	Progress	10.67	0.4324
System without normalization	Progress	8.012	0.316732
System with normalization	Progress	3.946	0.164331
Baseline x-vector system	Eval	10.67	0.4319
System with normalization	Eval	3.934	0.164624

5. Conclusions

It is seen that for text-independent speaker verification, a robust baseline model that it trained with out-of-domain data can be used with a LDA/PLDA back-end module that is trained on target domain to have a successful speaker verification system. It is also seen that x-vector embeddings are susceptable to enrollment durations and it is a more dominant factor than the domain fit of the LDA/PLDA module. In this task we used a basic normalization approach to overcome this problem, but further adaptation for utterance durations may be used in a way that taking utterance durations into account in the LDA/PLDA stage.

6. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[2] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[3] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A large-scale speaker identification dataset,” in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[4] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” *ArXiv*, vol. abs/1510.08484, 2015.

[5] H. Zeinali, L. Burget, and J. Cernocky, “A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database,” in *Proc. ASRU 2019 The 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.

[6] K. A. Zeinali, Hossein nad Lee, J. Alam, and L. Burget, “Short-duration speaker verification (SdSV) challenge 2020: the challenge evaluation plan.” arXiv preprint arXiv:1912.06311, Tech. Rep., 2020.

[7] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, “Speaker recognition for multi-speaker conversations using x-vectors,” *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5796–5800, 2019.