ROBUST SELF-SUPERVISED SPEAKER REPRESENTATION LEARNING VIA INSTANCE MIX REGULARIZATION

Woo Hyun Kang, Jahangir Alam, Abderrahim Fathan

Computer Research Institute of Montreal (CRIM)

woohyun.kang,jahangir.alam,abderrahim.fathan@crim.ca

ABSTRACT

Over the recent years, various self-supervised contrastive embedding learning methods for deep speaker verification were proposed. The performance of the self-supervised contrastive learning framework highly depends on the data augmentation technique, but due to the sensitive nature of speaker information within the speech signal, most speaker embedding training relies on simple augmentations such as additive noise or simulated reverberation. Thus while the conventional self-supervised speaker embedding systems can yield minimum within-utterance variability, the capability to generalize to out-of-set utterance is limited. In order to alleviate this problem, we propose a novel self-supervised learning framework for speaker verification which combines the angular prototypical loss and the instance mix (i-mix) regularization. The proposed method was evaluated on the VoxCeleb1 dataset and showed noticeable improvement over the standard self-supervised embedding method.

Index Terms— speaker verification, speaker embedding, representation learning, data augmentation, i-mix regularization

1. INTRODUCTION

Speaker verification is the task of verifying the claimed speaker identity based on the given speech samples and has become a key technology for personal authentication in many commercial, forensics, and law enforcement applications [1]. Commonly, utterance-level fixed-dimensional vectors (i.e. embedding vectors) are extracted from the enrollment and test speech samples and then fed into a scoring algorithm (e.g., cosine distance, probabilistic linear discriminant analysis) to measure their similarity or likelihood of being spoken by the same speaker. Classically, the i-vector framework has been one of the most dominant approaches for speech embedding [2], [3]. The widespread popularity of the i-vector framework in the speaker verification community can be attributed to its ability to summarize the distributive pattern of the speech with a relatively small amount of training data in an unsupervised manner.

In recent years, various methods have been proposed utilizing deep learning architectures for extracting embedding vectors and have shown better performance than the i-vector framework when a large amount of training data with enough diversity is available [4]. In [4, 5], a speaker recognition model consisting of a time-delay neural network (TDNN)-based frame-level network and a segment-level network was trained and the hidden layer activation of the segment-level network, denoted as x-vector, was extracted as the embedding vector. In [6], an ECAPA-TDNN architecture was proposed, which has shown state-of-the-art performance by introducing residual and squeeze-and-excitation (SE) components to the widely used TDNN-based embedding system. Although the deep embedding methods have outperformed the i-vector framework in various speaker verification benchmarks, since most of these models are trained in a fully supervised fashion, they require a large amount of speaker labeled dataset for optimization.

To overcome this limitation, a number of self-supervised embedding learning methods for deep speaker verification were proposed over the past couple of years [7], [8], [9], [10]. Many of these researches employ the contrastive learning scheme for optimization, where the embeddings from the same utterance (positive pairs) are trained to be close to each other while pushing away embeddings from different utterances (negative pairs). In order to effectively capture the utterance-dependent variability into the embedding, different types of augmentations are usually applied to the positive pair utterances. However, since speaker-dependent information can be easily distorted under severe augmentation, most speaker embedding training relies on simple augmentations such as noise/reverberation mixing [7] or frequency-/time-masking [9]. Due to this constraint, while the augmentation can help minimizing the within-utterance variability, its capability to generalize to out-of-set utterances is limited.

In this paper, we propose a novel framework for self-supervised embedding learning for speaker verification based on the instance mix scheme (i-mix) [11]. The i-mix method is a variant of the mixup augmentation [12], which has shown to improve the performance in various contrastive learning-based self-supervised embedding training for image and voice command classification. Unlike the conventional augmenta-

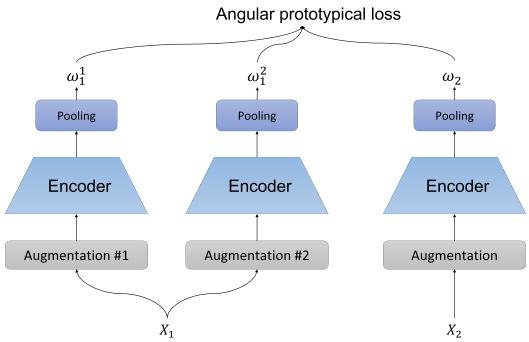

Fig. 1. The general framework for the self-supervised contrastive speaker embedding learning.

tions, which simply augments the waveform or spectrogram of the speech, the i-mix scheme aims to create a training sample with a new target identity by interpolating different samples along with their pseudo-labels. Therefore the i-mix strategy can efficiently improve the generalization of the learned representation. In light of this, we employ the i-mix regularization scheme to the self-supervised embedding learning framework, to produce a robust embedding which can perform well on verifying out-of-set speakers.

2. RELATED WORK AND BASELINE MODEL

2.1. Baseline self-supervised representation learning

Most recent self-supervised embedding learning methods use contrastive loss to produce embedding vectors with maximum utterance discriminability. As shown in Figure 1, in the self-supervised contrastive embedding learning framework, two samples are generated per utterance by applying different augmentations. The augmented samples are then passed through an encoder network to generate embedding vectors. The network is optimized via contrastive learning (e.g., prototypical loss), which minimizes the distance between the embeddings from the same utterance, while maximizing the distance between different utterance embeddings.

2.1.1. Encoder network

In our research, we have experimented with two types of encoder architectures which have shown good performance in speaker verification:

- ResNetSE34 [13]: The first architecture is the Fast ResNet, which follows the same general structure as the original ResNet with 34 layers (ResNet-34) [14] with squeeze-and-excitation [15], but only uses one-quarter of the channels in each residual block to reduce computational cost.
- ECAPA-TDNN [6]: an architecture which achieved state-of-the-art performance in text-independent speaker recognition. The ECAPA-TDNN uses squeeze-and-excitation as in the SE-ResNet, but also employs channel- and context-dependent statistics pooling and multi-layer aggregation.

The encoder network takes the acoustic feature as input and outputs the frame-level representations. The encoder outputs are aggregated via self-attention pooling, which computes the weighted average of the frame-level representations to obtain an utterance-level fixed dimensional embedding vector.

2.1.2. Angular prototypical objective

In order to train the encoder network with no speaker labels, we have used an utterance-discriminative contrastive loss, more specifically the angular prototypical loss function [10], [7]. Given a batch of prototype embedding vectors ω_i^1 and query embeddings ω_i^2, where ω_i^k indicates the embedding extracted from the i^{th} utterance X_i applied with augmentation $\#k$, the angular prototypical function is defined as follows:

$$L_{AP} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{exp(cos(\omega_i^1, \omega_i^2))}{\sum_{j=1}^{N} exp(cos(\omega_i^1, \omega_j^2))}, \qquad (1)$$

where cos represents the cosine similarity operation. Equation 1 can be interpreted as the cross-entropy loss which aims to maximize the similarity between the embeddings extracted from the same utterance, while minimizing the similarity between different utterance embeddings.

2.1.3. Speech augmentation

For training the encoder network via angular prototypical objective, we have used waveform-level data augmentations including additive noise and room impulse response (RIR) simulation [4]. In addition to the waveform-level augmentations, we have also applied augmentation over the extracted Mel-frequency cepstral coefficient (MFCC) feature, denoted here as cepsaugment, which is similar to specaugment scheme often used for automatic speech recognition (ASR) [16]. Analogous to the specaugment, in cepsaugment, a randomly selected time-cepstral bin is selected and masked before being fed into the encoder network.

2.2. Instance mix (i-mix) regularization strategy

The i-mix is a data-driven augmentation strategy for improving the generalization of the learned representation [11]. For arbitrary objective function $L_{pair}(x, y)$, where x is the input sample and y is the corresponding pseudo-label, given two data instances (x_i, y_i) and (x_j, y_j), the i-mix loss is defined

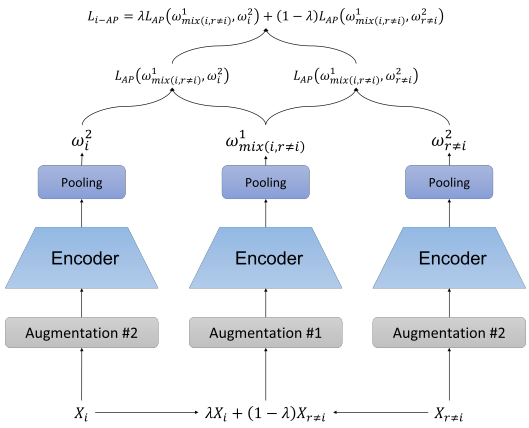

Fig. 2. The general framework for the proposed i-mix angular prototypical learning.

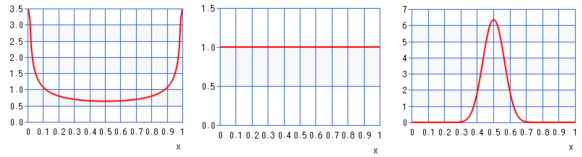

Fig. 3. Beta distributions with different α value.

as follows:

$$L_{pair}^{i-mix}((x_i, y_i), (x_j, y_j)) = L_{pair}(\lambda x_i + (1-\lambda)x_j, \lambda y_i + (1-\lambda)y_j), \quad (2)$$

where $\lambda \sim Beta(\alpha, \alpha)$ is a mixing coefficient. For cross-entropy-based L_{pair}, such as prototypical loss, equation 2 can be rewritten as,

$$L_{pair}^{i-mix}((x_i, y_i), (x_j, y_j)) = \lambda L_{pair}(x_i, y_i) + (1-\lambda)L_{pair}(x_j, y_j). \quad (3)$$

3. I-MIX ANGULAR PROTOTYPICAL OBJECTIVE

In this paper, we propose a new objective for robust self-supervised embedding learning, which combines the angular prototypical loss and the i-mix strategy. More specifically, as depicted in Figure 2, we propose to perform i-mix on the prototype embedding vectors:

$$L_{i-AP} = -\lambda \frac{1}{N}\sum_{i=1}^{N} \log \frac{exp(cos(\omega^1_{mix(i,r\neq i)}, \omega_i^2))}{\sum_{j=1}^{N} exp(cos(\omega^1_{mix(i,r\neq i)}, \omega_j^2))} - (1-\lambda)\frac{1}{N}\sum_{i=1}^{N} \log \frac{exp(cos(\omega^1_{mix(i,r\neq i)}, \omega^2_{r\neq i}))}{\sum_{j=1}^{N} exp(\omega^1_{mix(i,r\neq i)}, \omega_j^2))}, \quad (4)$$

where $\omega_{r\neq i}$ is an embedding randomly sampled from the batch $[\omega_1, \omega_2, ..., \omega_N]$ excluding ω_i, and $\omega_{mix(i,r\neq i)}$ is an embedding extracted from mixed utterance $\lambda X_i + (1-\lambda)X_{r\neq i}$.

Training the encoder network with the i-mix angular prototypical objective L_{i-AP} can be thought of as optimizing the network on a out-of-set utterance $X_{mix} = \lambda X_i + (1-\lambda)X_{r\neq i}$, which retains utterance-dependent attributes from both X_i and $X_{r\neq i}$. Hence the resulting embedding vector can generalize well on samples that are not included in the training dataset.

Analogous to the standard i-mix described in equation 3, we also use λ randomly sampled from $Beta(\alpha, \alpha)$. As depicted in Figure 3, the shape of the $Beta(\alpha, \alpha)$ distribution varies heavily depending on the α, and resulting lambda decides the expected behavior of the utterance interpolation $\lambda X_i + (1-\lambda)X_{r\neq i}$. For example, for $\alpha < 1.0$, the beta distribution is U-shaped, thus the sampled λ is likely to have value close to 1.0 or 0. In this case, the resulting interpolation can be interpreted as a babble noise augmentation, in which a relatively lower-powered speech is mixed to the original audio. Considering that the babble noise augmentation can improve the speaker recognition performance on mismatched condition [17], using λ sampled from beta distribution with $\alpha < 1.0$ may benefit the speaker embedding system.

On the other hand, using $\alpha > 1.0$ creates a bell-shaped beta distribution, which is similar to a Gaussian distribution with mean 0.5. The λ sampled from this distribution is likely to have value near 0.5, hence in the interpolation process, the two utterances will be added with similar power-level. Such overlapping speech samples are known to be challenging for the speaker recognition system, even for speakers observed during training [18]. Therefore, using $\alpha > 1.0$ may hinder the embedding encoder's learning capacity, thus resulting in an embedding vector with insufficient speaker-dependent information.

4. EXPERIMENTS

4.1. Experimental setup

In order to evaluate the performance of the proposed technique for self-supervised speaker verification, a set of experiments were conducted based on the VoxCeleb2 dataset [19]. For training the embedding networks, we used the *development* subset of the VoxCeleb2 dataset, consisting of 1,092,009 utterances collected from 5,994 speakers. The evaluation was performed according to the original VoxCeleb1 trial list [20], which consists of 4,874 utterances spoken by 40 speakers.

The acoustic features used in the experiments were 40-dimensional MFCCs extracted at every 10 ms, using a 25 ms Hamming window. The embedding networks are trained with segments consisting of 180 frames, using the ADAM optimization technique [21].

All the experimented networks were trained with initial

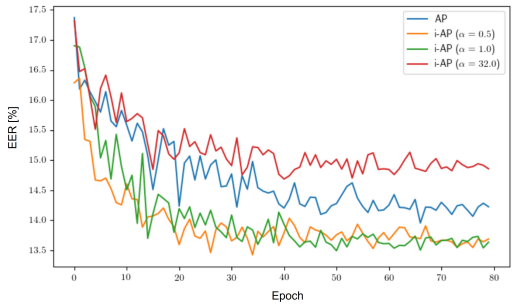

Fig. 4. The VoxCeleb1 original trial EER results of ResNetSE34 systems on different training epochs.

learning rate 0.001 decayed with ratio 0.95 for 150 epochs, and the models from the best performing checkpoint was selected. The batch size for training was set to be 200. Cosine similarity was used for computing the verification scores in the experiments.

4.2. Experimental results

4.2.1. Training analysis

In this section, we evaluate the encoder networks on each training epoch to analyze the generalization behavior of i-mix angular prototypical learning. Figure 4 depicts the epoch-wise equal error rate (EER) of systems trained with and without i-mix on the VoxCeleb1 trial set. As shown in the figure, using i-mix angular prototypical (i-AP) with $\alpha = 0.5, 1.0$ improved the performance on out-of-domain utterances (i.e., VoxCeleb1) much faster than the standard prototypical objective system (AP). This clearly indicates that the i-mix regularization can help to improve the generalization of the embedding vectors for speaker verification task.

However, it could also be observed that using $\alpha = 32.0$ failed to generalize on out-of-domain utterances. This may be attributed to the fact that, as discussed in section 3, the λ sampled from the beta distribution with $\alpha = 32.0$ is likely to have value near 0.5, thus generating augmented samples too difficult to learn.

4.2.2. Comparison between embeddings extracted from systems trained with different augmentations

In this section, we compare self-supervised speaker verification systems with different encoder architectures, augmentations, and objective functions. As depicted in Table 1, it could be noticed that the effect of augmentation varies depending on the encoder network. For example, while the ECAPA-TDNN-based system benefited from cepsaugment, the performance of ResNetSE34 systems were generally degraded when cepsaugment was used. This reassures that the selection of data augmentation method is crucial for obtaining optimal self-supervised embedding vectors.

Table 1. EER (%) comparison between the embedding networks trained with different augmentations and objectives.

Encoder	Augmentation	Objective	EER [%]
Human Benchmark [7]			15.7700
i-vector + Cosine [7]			15.2800
ResNetSE34	waveaug	AP	13.9555
		i-AP ($\alpha = 0.5$)	**13.4252**
		i-AP ($\alpha = 1.0$)	13.4942
		i-AP ($\alpha = 32.0$)	14.6872
	waveaug +cepsaug	AP	14.5228
		i-AP ($\alpha = 0.5$)	**13.8971**
		i-AP ($\alpha = 1.0$)	13.9767
		i-AP ($\alpha = 32.0$)	15.6628
ECAPA-TDNN	waveaug	AP	11.6384
		i-AP ($\alpha = 0.5$)	11.9618
		i-AP ($\alpha = 1.0$)	**11.2407**
		i-AP ($\alpha = 32.0$)	11.8240
	waveaug +cepsaug	AP	11.6013
		i-AP ($\alpha = 0.5$)	**10.6257**
		i-AP ($\alpha = 1.0$)	10.9279
		i-AP ($\alpha = 32.0$)	12.1633

On the other hand, the i-mix angular prototypical objective was able to improve the performance in all settings (i.e., architecture, augmentation). As observed in Section 4.2.1, in most cases, i-AP with $\alpha = 0.5, 1.0$ outperformed AP, while $\alpha = 32.0$ hindered the performance. Especially in ECAPA-TDNN with waveaugment and cepsaugment, i-AP ($\alpha = 0.5$) outperformed the AP with a relative improvement of 8.41% in terms of EER. This shows that with the right choice of α, the self-supervised embedding network can be improved significantly via incorporating i-mix regularization to the objective.

5. CONCLUSION

In this paper, we proposed a novel objective function for self-supervised speaker embedding learning, which combines the angular prototypical loss and the i-mix regularization scheme (i-AP). By training on augmented samples with a new target identity created via i-mix, the generalization of the embedding vectors on out-of-domain utterances can be increased.

In order to evaluate the proposed i-AP objective, we have conducted several experiments on the VoxCeleb dataset. Our results showed that the proposed i-AP can significantly improve the generalization of the self-supervised embeddings, hence outperforming the systems trained with standard angular prototypical objective.

In our future study, we will be expanding the i-AP objective function by applying different types of interpolation schemes. Moreover, we will be analyzing the effect of incorporating i-mix regularization on other self-supervised embedding learning methods, such as clustering-driven pseudo-label-based softmax objectives.

6. REFERENCES

[1] John H.L. Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[3] P. Kenny, "A small footprint i-vector extractor," in *Odyssey*, 2012.

[4] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[5] David Snyder, D. Garcia-Romero, Daniel Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *INTERSPEECH*, 2017.

[6] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, Helen Meng, Bo Xu, and Thomas Fang Zheng, Eds. 2020, pp. 3830–3834, ISCA.

[7] Jaesung Huh, Hee Soo Heo, Jingu Kang, Shinji Watanabe, and Joon Son Chung, "Augmentation adversarial training for unsupervised speaker recognition," in *Workshop on Self-Supervised Learning for Speech and Audio Processing, NeurIPS*, 2020.

[8] Sung Hwan Mun, Woo Hyun Kang, Min Hyun Han, and Nam Soo Kim, "Unsupervised representation learning for speaker recognition via contrastive equilibrium learning," 2020.

[9] Ke Ding, Xuanji He, and Guanglu Wan, "Learning speaker embedding with momentum contrast," 2020.

[10] Haoran Zhang, Yuexian Zou, and Helin Wang, "Contrastive self-supervised learning for text-independent speaker verification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6713–6717.

[11] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee, "i-mix: A domain-agnostic strategy for contrastive representation learning," in *ICLR*, 2021.

[12] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2017.

[13] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[16] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition.," in *Interspeech 2019*, 2019, pp. 2613–2617.

[17] Nitish Krishnamurthy and John H. L. Hansen, "Babble noise: Modeling, analysis, and applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1394–1407, 2009.

[18] Van-Thuan Tran and Wei-Ho Tsai, "Speaker identification in multi-talker overlapping speech using neural networks," *IEEE Access*, vol. 8, pp. 134868–134879, 2020.

[19] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.