ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095066

DASA: DIFFICULTY-AWARE SEMANTIC AUGMENTATION FOR SPEAKER VERIFICATION

Yuanyuan Wang[1,*], *Yang Zhang*[1], *Zhiyong Wu*[1,3,†], *Zhihan Yang*[1], *Tao Wei*[2], *Kun Zou*[2], *Helen Meng*[3]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Ping An Technology, Shenzhen, China
[3] The Chinese University of Hong Kong, Hong Kong SAR, China
{wangyuan21, zhangy20}@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn

ABSTRACT

Data augmentation is vital to the generalization ability and robustness of deep neural networks (DNNs) models. Existing augmentation methods for speaker verification manipulate the raw signal, which are time-consuming and the augmented samples lack diversity. In this paper, we present a novel difficulty-aware semantic augmentation (DASA) approach for speaker verification, which can generate diversified training samples in speaker embedding space with negligible extra computing cost. Firstly, we augment training samples by perturbing speaker embeddings along semantic directions, which are obtained from speaker-wise covariance matrices. Secondly, accurate covariance matrices are estimated from robust speaker embeddings during training, so we introduce difficulty-aware additive margin softmax (DAAM-Softmax) to obtain optimal speaker embeddings. Finally, we assume the number of augmented samples goes to infinity and derive a closed-form upper bound of the expected loss with DASA, which achieves compatibility and efficiency. Extensive experiments demonstrate the proposed approach can achieve a remarkable performance improvement. The best result achieves a 14.6% relative reduction in EER metric on CN-Celeb evaluation set.

Index Terms— speaker verification, data augmentation, semantic augmentation, difficulty-aware

1. INTRODUCTION

Speaker verification is to compare two speech utterances and verify whether they are spoken by the same speaker [1]. In recent years, deep neural networks (DNNs) have significantly boosted advancements in speaker verification [2, 3, 4, 5, 6]. Now speaker verification has become an important technology in our daily life, such as biometric identification or smartphone control. However, current speaker verification systems are still unsatisfactory in real industrial scenarios, such as video analysis with non-standard ambient sound, channel number, and speech emotion. Data insufficiency is one of the most critical challenges to the robustness of performance in complex scenarios.

To address the problem of data insufficiency, data augmentation is an essential technique to increase the diversity and quantity of training samples [7]. Conventional data augmentation techniques for speaker verification, e.g., additive noises, reverberation, and speed perturbation [4, 8], are raw signal-level data augmentation. Wang *et al.* [9] investigate SpecAugment [10] and augment data by masking the spectrogram during training. The diversity of augmented samples generated by these methods is inherently limited by the direct manipulation of the raw audio [8]. Besides, these methods will also introduce huge computing cost and I/O time for augmentation.

Recently, Deep generative models [11, 12], e.g., Generative Adversarial Networks (GANs), Variational Autoencoder (VAE), have been introduced to learn the distribution of noisy speaker embeddings and generate new embeddings from the generative models learned distribution. However, these methods utilize complex deep generative models to explicitly augment samples, which significantly slow down the training process of recognition models. Xun *et al.*[13] sample noise from the pure noise distribution and directly add it to clean embeddings to generate augmented embeddings. But the distribution is derived and specialized on extra noise datasets.

These problems can potentially be solved by the implicit semantic data augmentation (ISDA) approach [14, 15]. Noise data, auxiliary model, and model modification are no longer required, and therefore, it eliminates extra computation and time cost. ISDA augments training data by translating speaker embeddings towards meaningful semantic directions, which are sampled from a zero-mean normal distribution with the dynamically estimated covariance [16]. Importantly, ISDA estimates speaker-wise covariance matrices according to speaker embeddings. We find that ISDA performs suboptimally in speaker verification tasks, because it is based on softmax training, unable to acquire the optimal embedding to estimate the appropriate covariance matrices.

In this paper, we propose a novel difficulty-aware semantic augmentation (DASA) approach to augment the training data at deep speaker embedding level rather than the raw signal level, from the point that more accurate covariance matrices require optimal embeddings. Without changing the model structure, the loss function plays a crucial role in speaker embeddings. We first use additive margin softmax (AM-Softmax) [17] loss to reduce the intra-class variation and increase the inter-class difference [17, 18]. Furthermore, AM-Softmax assumes that all the speakers have the same inter-class difference. Difficult samples are highly similar to other speakers and have insufficient discrimination, so their inter-class difference should be larger. Therefore, we are inspired by [19] and introduce difficulty-aware AM-Softmax (DAAM-Softmax) which sets inter-class difference according to sample difficulty. DAAM-Softmax can solve the over-optimization of easy samples and under-optimization of difficult samples, so it gets the optimal embeddings to estimate covariance matrices of DASA. Finally, we assume infinite sampling directions and derive an upper bound of expected loss with DASA, which can be simply adopted by most models with little computation

[*] Work performed during an internship at Ping An Technology.
[†] Corresponding author.

and time cost. In summary, our contributions are as follows:

- **Low computing cost**: Instead of augmenting the speech by processing the raw signal, DASA performs meaningful semantic perturbation in speaker embedding space with little extra computing cost.
- **Good compatibility**: The proposed DASA approach can combine with traditional data augmentation and is compatible with most models to achieve better recognition performance.
- **Outstanding improvement**: Extensive experiments conducted on VoxCeleb and CN-Celeb demonstrate the proposed method can obtain remarkable performance improvement.

2. DIFFICULTY-AWARE SEMANTIC AUGMENTATION

In this section, we present the difficulty-aware semantic augmentation (DASA) for speaker verification. The comparison between conventional data augmentation and the proposed DASA is shown in Fig.1. We assume the training set is $S = \{(\boldsymbol{u}_i, y_i)\}_{i=1}^{N}$, and $y_i \in \{1, \dots, C\}$ is the label of the i-th utterance $\boldsymbol{u}_i$ over C speaker categories. The vector $\boldsymbol{f}_i = [f_{i1}, \dots, f_{iF}]^T$ indicates the F-dimensional deep embeddings of $\boldsymbol{u}_i$ learned by deep neural network D.

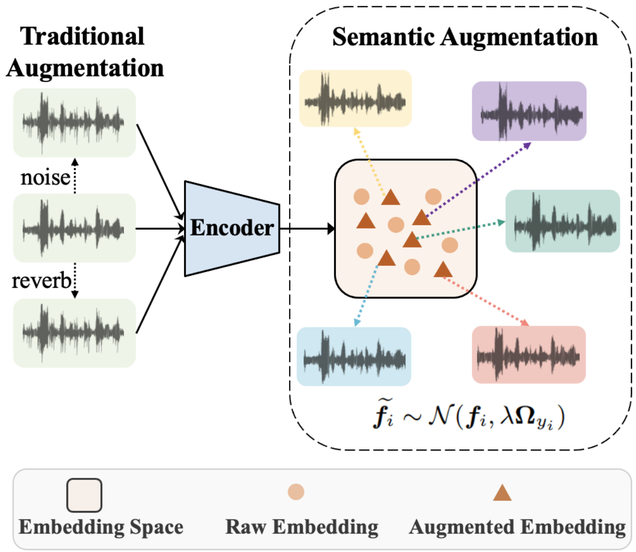

Fig. 1. **The comparison between conventional data augmentation (left) and proposed difficulty-aware semantic augmentation (right).** Conventional data augmentations directly perform on the raw signal level, such as adding noise and reverberation, whereas the DASA augments the training data at the speaker embedding level. DASA can exactly estimate speaker-wise covariance matrices in training with DAAM-Softmax, and these covariance matrices are used to establish a zero-mean normal distribution. Then DASA generates new instances by changing speaker embeddings to semantic transformation directions sampled from the normal distribution. Augmented diversified instances are represented by speech rectangles with different colors. After that, to improve efficiency, we assume an infinite number of sampling and derive a closed-form upper bound of the expected loss with DASA.

2.1. Implicit Semantic Data Augmentation

Here, we revisit the ISDA proposed by [14]. We define that speaker-wise covariance matrix $\boldsymbol{\Omega}_{y_i}$ is computed from all embeddings of training samples in class y_i, so C speakers have C covariance matrices during training. We can then randomly sample from a zero-mean normal distribution $\mathcal{N}(0, \lambda\boldsymbol{\Omega}_{y_i})$ to obtain vectors representing semantic directions of $\boldsymbol{f}_i$. The augmented embedding $\widetilde{\boldsymbol{f}}_i$ is generated by Eq.(1), where λ is a positive number that controls the strength of ISDA.

$$\widetilde{\boldsymbol{f}}_i \sim \mathcal{N}(\boldsymbol{f}_i, \lambda\boldsymbol{\Omega}_{y_i}) \tag{1}$$

Firstly, we explicitly augment each embedding $\boldsymbol{f}_i$ for M times. Then through the softmax layer, the cross entropy (CE) loss function is calculated as Eq.(2), where **w** and **b** are the weight matrices and biases of the last fully connected layer, respectively.

$$L_M = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{k=1}^{M} -log(\frac{e^{\boldsymbol{w}_{y_i}^T \boldsymbol{f}_i^k + b_{y_i}}}{\sum_{j=1}^{C} e^{\boldsymbol{w}_j^T \boldsymbol{f}_i^k + b_j}}) \tag{2}$$

When considering that $M \to \infty$, an easy-to-compute upper bound can be derived for the loss function. The expectation of the CE loss is calculated as all possible augmentation embeddings and uses Jensen's inequality $E(logX) \leq logE(X)$ to approximate:

$$\begin{aligned} L_\infty &= \frac{1}{N}\sum_{i=1}^{N} E_{\widetilde{\boldsymbol{f}}_i}(log(\sum_{j=1}^{C} e^{(\boldsymbol{w}_j^T - \boldsymbol{w}_{y_i}^T)\widetilde{\boldsymbol{f}}_i + (b_j - b_{y_i})})) \\ &\leq \frac{1}{N}\sum_{i=1}^{N} log(\sum_{j=1}^{C} E_{\widetilde{\boldsymbol{f}}_i}(e^{(\boldsymbol{w}_j^T - \boldsymbol{w}_{y_i}^T)\widetilde{\boldsymbol{f}}_i + (b_j - b_{y_i})})) \end{aligned} \tag{3}$$

According to Eq.(1), we can deduce Eq.(4), where $\Delta\boldsymbol{w} = (\boldsymbol{w}_j^T - \boldsymbol{w}_{y_i}^T)$, $\Delta b = (b_j - b_{y_i})$ and $\boldsymbol{\Phi} = (\boldsymbol{w}_j^T - \boldsymbol{w}_{y_i}^T)\boldsymbol{\Omega}_{y_i}(\boldsymbol{w}_j - \boldsymbol{w}_{y_i})$.

$$\Delta\boldsymbol{w}\widetilde{\boldsymbol{f}}_i + \Delta b \sim \mathcal{N}(\Delta\boldsymbol{w}\boldsymbol{f}_i + \Delta b, \lambda\boldsymbol{\Phi}) \tag{4}$$

Combined with the Moment Generating Function Eq.(5), we can derive an easily computable upper bound as the final loss function Eq.(6). Without explicitly augmented samples, ISDA can be performed more efficiently by calculating the upper bound of the loss.

$$E[e^{tX}] = e^{t\mu + \frac{1}{2}\sigma^2 t^2}, X \sim N(\mu, \sigma^2) \tag{5}$$

$$L_{\infty(ISDA)} \leq \frac{1}{N}\sum_{i=1}^{N} log(\sum_{j=1}^{C} e^{\Delta\boldsymbol{w}\boldsymbol{f}_i + \Delta b + \frac{1}{2}\lambda\boldsymbol{\Phi}}) \tag{6}$$

2.2. AM-Softmax with Semantic Augmentation

The above formula Eq.6 performs poorly in speaker verification, since it is based on the softmax function. AM-Softmax can better separate different speakers and make speaker embeddings from the same speaker more compact by introducing the angular margin into the softmax. So it can get better embeddings to accurately estimate covariance matrices $\boldsymbol{\Omega}_{y_i}$.

In this section, we will derive the upper bound of expected AM-Softmax loss with semantic augmentation. The formula for AM-Softmax loss commonly used in speaker verification [17] can be formulated as follows, where both $\boldsymbol{w}$ and $\boldsymbol{f}$ are normalized.

$$\begin{aligned} L_{AMS} &= -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s\cdot(cos\theta_{y_i} - m)}}{e^{s\cdot(cos\theta_{y_i} - m)} + \sum_{j=1, j\neq y_i}^{C} e^{s\cdot cos\theta_j}} \\ &= -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(\boldsymbol{w}_{y_i}^T \boldsymbol{f}_i - m)}}{e^{s(\boldsymbol{w}_{y_i}^T \boldsymbol{f}_i - m)} + \sum_{j=1, j\neq y_i}^{C} e^{s\boldsymbol{w}_j^T \boldsymbol{f}_i}} \end{aligned} \tag{7}$$

Similar to formulas Eq.(2) and Eq.(3), we also sample $M \to \infty$ times and approximate the upper bound of the expected loss:

$$\begin{aligned} L_A &= \frac{1}{N}\sum_{i=1}^{N} E_{\widetilde{\boldsymbol{f}}_i}(log(1+\sum_{j=1,j\neq y_i}^{C} e^{s\Delta \boldsymbol{w}\widetilde{\boldsymbol{f}}_i+sm})) \\ &\leq \frac{1}{N}\sum_{i=1}^{N} log(1+\sum_{j=1,j\neq y_i}^{C} E_{\widetilde{\boldsymbol{f}}_i}(e^{s\Delta \boldsymbol{w}\widetilde{\boldsymbol{f}}_i+sm})) \end{aligned} \tag{8}$$

According to equation Eq.(1), we can obtain the following Gaussian distribution:

$$s\Delta \boldsymbol{w}\widetilde{\boldsymbol{f}}_i + sm \sim \mathcal{N}(s\Delta \boldsymbol{w}\boldsymbol{f}_i + sm, \lambda\boldsymbol{\Phi}s^2) \tag{9}$$

From formula Eq.(5), the upper bound of the expected AM-Softmax loss with semantic augmentation can be obtained:

$$L_A \leq \frac{1}{N}\sum_{i=1}^{N} log(1+\sum_{j=1,j\neq y_i}^{C} e^{s\Delta \boldsymbol{w}\boldsymbol{f}_i+sm+\frac{1}{2}\lambda\boldsymbol{\Phi}s^2)}) \tag{10}$$

2.3. Difficulty-aware Semantic Augmentation

All the speakers have the same inter-class difference in AM-Softmax, so the margin in the AM-Softmax loss is set to be a fixed value. But in fact, the angle between the speaker embedding and the center of the speaker category is different for each sample [20]. When the angle is larger, it indicates that the samples are more difficult to verify. A larger margin should be set to increase inter-class difference. For simple training utterances, setting a slightly smaller margin can distinguish well. So we introduce difficulty-aware AM-Softmax (DAAM-Softmax) as shown in Eq.(12). Harder samples get larger margins, thus achieving the purpose of difficulty-aware discriminative learning.

$$DA = \frac{1-cos\theta_{y_i}}{2} \tag{11}$$

$$L_{DAAM} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(\boldsymbol{w}_{y_i}^T\boldsymbol{f}_i - m\cdot DA)}}{e^{s(\boldsymbol{w}_{y_i}^T\boldsymbol{f}_i - m\cdot DA)}+\sum_{j=1,j\neq y_i}^{C} e^{s\boldsymbol{w}_j^T\boldsymbol{f}_i}} \tag{12}$$

The DA makes the coefficient in the range of $[0,1]$, and ensures difficulty is negatively correlated with logits. It will focus more on difficult utterances that usually lead to worse results [19]. This enables further accuracy improvements across all training samples which can generate optimal deep embeddings. Therefore, the covariance matrix $\boldsymbol{\Omega}_{y_i}$ estimated with optimal embeddings is more accurate. Based on our DAAM-Softmax, the upper bound of the expected loss with the proposed difficulty-aware semantic augmentation ($\boldsymbol{DASA}$) is Eq.(13). Therefore, DASA can be easily applied to most deep models as a novel robust loss function.

$$L_{\infty(DASA)} \leq \frac{1}{N}\sum_{i=1}^{N} log(1+\sum_{j=1,j\neq y_i}^{C} e^{s\Delta \boldsymbol{w}\boldsymbol{f}_i+sm\cdot DA+\frac{1}{2}\lambda\boldsymbol{\Phi}s^2)}) \tag{13}$$

3. EXPERIMENTAL SETUP

3.1. Dataset

VoxCeleb1&2 [21, 22] and CN-Celeb [23, 24] are used in our experiments. VoxCeleb is an audio-visual dataset consisting of 2,000+ hours of short clips of human speech. The development set of VoxCeleb2 containing 5994 speakers is used to train models. We use VoxCeleb1 test set to evaluate system performance. CN-Celeb is a more challenging large-scale text-independent speaker recognition dataset containing 2800 speakers for network training and 200 speakers for system evaluation.

3.2. System Implementation

Network. The experimental systems are based on WeSpeaker [1]. To sufficiently verify the effectiveness of our method, we use ResNet34 [3] and ECAPA-TDNN [5] as speaker models, which accept 80-dimensional Fbanks as input and extract 256-dimensional speaker embeddings.

Implementation details. The initial and final learning rates are 0.1 and 5e-5, which are decreased with an exponential schedule. We train all models for 150 epochs using an SGD optimization algorithm with a Nesterov momentum. We set the weight decay as 1e-4 and the momentum as 0.9. The batch size is 128. In most experiments, we perform traditional data augmentation online, including adding noise and reverberation but no speed perturbation. Each training utterance has a probability of 0.6 for conventional data augmentation [25].

In addition, we adopt a simple but effective deferred optimization schedule [26]. In the first 60 epochs, we train models with DAAM-Softmax loss for learning good embeddings. In later epochs, we estimate the covariance matrix for DASA with $\lambda = (t/T)\cdot\lambda_0$, where t is the current iteration.

4. RESULTS AND ANALYSES

4.1. Results on VoxCeleb

Table 1. Results on VoxCeleb

Loss	**Hyperparams**	**VoxCeleb1-test**	
		EER(%)	**minDCF**
Softmax	/	1.760	0.210
ISDA [14]	$\lambda_0 = 7$	1.191	0.141
AM-Softmax [17]	$s = 32, m = 0.2$	0.936	0.093
DAM-Softmax [20]	$s = 32, m = 0.2$	1.032	0.107
DASA(Ours)	$\lambda_0 = 4$	0.927	0.094
	$\lambda_0 = 3$	**0.904**	0.098

* λ_0 indicates the strength of DASA. The scaling factor (s) is 32, and the margin (m) is 0.2 for all experiments except softmax and ISDA.

In the first experiment, we train ECAPA-TDNN model on VoxCeleb2, and test performance on VoxCeleb1 test set and results are shown in Table 1. ISDA performs worse than AM-Softmax on speaker verification, because it is based on softmax loss. So we will not conduct the ISDA experiment later. DAM-Softmax in the fourth line comes from the dynamic-margin softmax proposed in [20]. It proposes to multiply the fixed margin by a dynamic variable as follows:

$$DY = \frac{e^{(1-cos\theta_{y_i})}}{\gamma} \tag{14}$$

where γ is set to 2. Since DY is similar to Eq.(11), our DASA is compared with DAM-Softmax in all subsequent experiments. The result of DAM here is slightly worse than AM-Softmax, which may be due to the large coefficient caused by the exponential form of DY.

The last two rows are the proposed DASA in Table 1, where λ_0 represents the strength of DASA. DASA outperforms the AM-Softmax (baseline) and DAM-Softmax methods, and the performance is optimal when λ_0 is 3.

4.2. Results on CN-Celeb

Compared to VoxCeleb, CN-Celeb is a more complex and challenging dataset [23]. This is because speakers have 11 different genres of utterances, which leads to significant variation in speaking styles. In addition, most short utterances recorded at different times with different devices involve real-world noise and perceptible voices in

[1] https://github.com/wenet-e2e/wespeaker

Table 2. Results on CN-Celeb

Model	Loss	Hyperparams	CN-Celeb-Eval	
			EER(%)	minDCF
ECAPA-TDNN	Softmax	/	10.641	0.485
	AM-Softmax [17]	$s = 32, m = 0.2$	8.877	0.435
	DAM-Softmax [20]	$s = 32, m = 0.2$	8.348	0.438
	DASA(Ours)	$\lambda_0 = 0.15$	8.278	0.438
		$\lambda_0 = DA$	8.255	0.443
		$\lambda_0 = 0.2$	8.235	**0.432**
		$\lambda_0 = 0.1$	$\underline{8.161}$	0.437
		$\lambda_0 = DY$	**8.021**	0.443
ResNet34	Softmax	/	9.367	0.429
	AM-Softmax [17]	$s = 32, m = 0.2$	8.404	0.416
	DAM-Softmax [20]	$s = 32, m = 0.2$	7.977	0.407
	DASA(Ours)	$\lambda_0 = DY$	7.784	**0.403**
		$\lambda_0 = DA$	7.577	0.406
		$\lambda_0 = 0.1$	7.568	0.406
		$\lambda_0 = 0.2$	7.413	0.412
		$\lambda_0 = 0.15$	**7.379**	$\underline{0.404}$

the background. To meet the demand for more real application scenarios, we train the ECAPA-TDNN and ResNet34 models on the CN-Celelb training set and test the performance on the CN-Celeb test set in this experiment.

The hyperparameters of AM-Softmax have been carefully tuned to achieve outstanding performance and we experiment on this basis. Results in table 2 show that DASA performs significantly better than AM-Softmax in both models, where $\lambda_0 = DA$ and $\lambda_0 = DY$ mean to change the strength of semantic augmentation with Eq. (11), (14) dynamically. Bold and underline indicate the optimal and suboptimal results, respectively. DASA under both ECAPA-TDNN and ResNet34 models obtain 9.6% and 12.2% relative reduction in EER compared with AM-Softmax when evaluated on the CN-Celeb test set, respectively. In the case of ResNet34 and $\lambda_0 = 0.15$, both EER and minimum Detection Cost Function (minDCF) are almost optimal.

Remarkable improvements in the experiments further demonstrate the effectiveness of DASA. Furthermore, compared with the experiments on VoxCeleb in Table 1, the improvement on CN-Celeb is more apparent, which sufficiently verifies that DASA can pay more attention to difficult samples and performs better on complex trials of real industrial scenarios.

4.3. Ablation Study

In this section, we conduct ablation studies to explore the effect of each component in DASA. Table 3 shows the results of ECAPA-TDNN on CN-Celeb, with similar trends on other models and datasets. Compared with the proposed DASA, we only remove SA and DA in the 'w/o SA' and 'w/o DA' lines, respectively. The 'w/o SA' uses only DAAM-Softmax as shown in Eq.(12), and so does not include any hyperparameters for strength of DASA. The 'w/o DA' uses the Eq.(10) with $\lambda_0 = 0.1$, which allows for a fair comparison with our DASA. The results denote that both DA and SA can improve performance, with improvement of DA being slightly more apparent. The reduction of EER in the 'w/o SA' row verifies that DA helps the model learn better embeddings. The comparison of DASA and the 'w/o DA' line shows that performance of SA is significantly improved with DA, indicating better embeddings are beneficial for SA to estimate covariance matrices.

Since all the experiments in Table 1 and Table 2 are performed with traditional data augmentation, we conduct experiments to eliminate the impact. Comparing the fifth row with the sixth row shows that the performance of DASA is relatively improved by 14.6% without traditional augmentation, which exceeds the performance improvement with traditional augmentation. Furthermore, this indicates that DASA and traditional augmentation are complementary and can be combined to achieve higher performance.

Table 3. Ablation study of ECAPA-TDNN on CN-Celeb

Loss	Hyperparams	CN-Celeb-Eval	
		EER(%)	minDCF
DASA(Ours)	$\lambda_0 = 0.1$	$\underline{8.161}$	0.437
	$\lambda_0 = DY$	**8.021**	0.443
w/o SA	/	8.433	0.44
w/o DA	$\lambda_0 = 0.1$	8.632	0.439
w/o aug	$\lambda_0 = 0.15$	9.175	0.479
w/o aug, w/o DASA	/	10.739	0.485

* w/o aug means no traditional data augmentation.

More importantly, we calculate the average running time of an epoch as extra cost on Intel(R) Xeon(R) Silver 4210R CPU (2.40GHz) and GeForce RTX 3080. Compared with AM-Softmax, the cost introduced by traditional augmentation is about 213.3% due to the limitation of I/O. However, the additional cost of DASA on ECAPA-TDNN and ResNet34 is only 7.4% and 8.5%, respectively.

5. CONCLUSIONS

In this study, we present difficulty-aware semantic augmentation (DASA), a novel data augmentation approach for speaker verification. Different from conventional augmentation methods, which perform on raw speech signal level, the proposed DASA augments the training data on deep speaker embedding level without noticeable extra computing cost. Besides, DASA could be an ideal complement to existing data augmentation and be applied to various networks. Extensive experiments on VoxCeleb and CN-Celeb show that DASA is effective and performs better on more realistic and challenging trials. In the future, our method can hopefully be extended to other margin losses similar to AM-Softmax.

Acknowledgement: This work was supported by National Natural Science Foundation of China (62076144), CCF-Tencent Open Research Fund (RAGR20210122) and Shenzhen Science and Technology Program (WDZC20200818121348001).

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] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4052–4056.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[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)*. IEEE, 2018, pp. 5329–5333.

[5] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *arXiv preprint arXiv:2005.07143*, 2020.

[6] Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-yi Lee, and Helen Meng, “Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification,” *arXiv preprint arXiv:2203.15249*, 2022.

[7] Hitoshi Yamamoto, Kong Aik Lee, Koji Okabe, and Takafumi Koshinaka, “Speaker augmentation and bandwidth extension for deep speaker embedding.,” in *Interspeech*, 2019, pp. 406–410.

[8] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, “Audio augmentation for speech recognition,” in *Sixteenth annual conference of the international speech communication association*, 2015.

[9] Shuai Wang, Johan Rohdin, Oldřich Plchot, Lukáš Burget, Kai Yu, and Jan Černockỳ, “Investigation of specaugment for deep speaker embedding learning,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7139–7143.

[10] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[11] Yexin Yang, Shuai Wang, Man Sun, Yanmin Qian, and Kai Yu, “Generative adversarial networks based x-vector augmentation for robust probabilistic linear discriminant analysis in speaker verification,” in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 205–209.

[12] Zhanghao Wu, Shuai Wang, Yanmin Qian, and Kai Yu, “Data augmentation using variational autoencoder for embedding based speaker verification.,” in *INTERSPEECH*, 2019, pp. 1163–1167.

[13] Xun Gong, Zhengyang Chen, Yexin Yang, Shuai Wang, Lan Wang, and Yanmin Qian, “Speaker embedding augmentation with noise distribution matching,” in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[14] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu, “Implicit semantic data augmentation for deep networks,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu, “Regularizing deep networks with semantic data augmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3733–3748, 2022.

[16] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng, “Metasaug: Meta semantic augmentation for long-tailed visual recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5212–5221.

[17] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, “Additive margin softmax for face verification,” *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[18] Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu, “Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1652–1656.

[19] Yan Zhao, Weicong Chen, Xu Tan, Kai Huang, and Jihong Zhu, “Adaptive logit adjustment loss for long-tailed visual recognition,” 2021.

[20] Dao Zhou, Longbiao Wang, Kong Aik Lee, Yibo Wu, Meng Liu, Jianwu Dang, and Jianguo Wei, “Dynamic margin softmax loss for speaker verification.,” in *INTERSPEECH*, 2020, pp. 3800–3804.

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

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

[23] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang, “Cn-celeb: a challenging chinese speaker recognition dataset,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[24] Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang, “Cn-celeb: multi-genre speaker recognition,” *Speech Communication*, 2022.

[25] Zhengyang Chen, Bei Liu, Bing Han, Leying Zhang, and Yanmin Qian, “The sjtu x-lance lab system for cnsrc 2022,” 2022.

[26] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, “Learning imbalanced datasets with label-distribution-aware margin loss,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.