applied sciences

MDPI

Article

A Supervised Learning Method for Improving the Generalization of Speaker Verification Systems by Learning Metrics from a Mean Teacher

Ju-Ho Kim [1], Hye-Jin Shim [1], Jee-Weon Jung [2] and Ha-Jin Yu [1,*]

[1] School of Computer Science, University of Seoul, Seoul 02504, Korea; wngh1187@naver.com (J.-H.K.); shimhz6.6@gmail.com (H.-J.S.)
[2] Naver Corporation, Naver Green Factory, Seongnam 13561, Korea; jeewon.leo.jung@gmail.com
* Correspondence: hjyu@uos.ac.kr

Abstract: The majority of recent speaker verification tasks are studied under open-set evaluation scenarios considering real-world conditions. The characteristics of these tasks imply that the generalization towards unseen speakers is a critical capability. Thus, this study aims to improve the generalization of the system for the performance enhancement of speaker verification. To achieve this goal, we propose a novel supervised-learning-method-based speaker verification system using the mean teacher framework. The mean teacher network refers to the temporal averaging of deep neural network parameters, which can produce a more accurate, stable representations than fixed weights at the end of training and is conventionally used for semi-supervised learning. Leveraging the success of the mean teacher framework in many studies, the proposed supervised learning method exploits the mean teacher network as an auxiliary model for better training of the main model, the student network. By learning the reliable intermediate representations derived from the mean teacher network as well as one-hot speaker labels, the student network is encouraged to explore more discriminative embedding spaces. The experimental results demonstrate that the proposed method relatively reduces the equal error rate by 11.61%, compared to the baseline system.

Keywords: speaker verification; mean teacher; supervised learning; metric learning

Citation: Kim, J.-H.; Shim, H.-J.; Jung, J.-W.; Yu, H.-J. A Supervised Learning Method for Improving the Generalization of Speaker Verification Systems by Learning Metrics from a Mean Teacher. *Appl. Sci.* **2022**, *12*, 76. https://doi.org/10.3390/app12010076

Academic Editor: Arcangelo Castiglione

Received: 26 November 2021
Accepted: 20 December 2021
Published: 22 December 2021

1. Introduction

Speaker verification (SV) is the task of authenticating whether a speaker of an unknown input utterance matches the target speaker, and it is widely used in applications, such as voice assistant systems [1,2]. Recent SV systems are primarily studied as an open-set scenario that tests using the utterances of speakers not seen in the training phase, requiring strong generalization [2,3]. Considering these characteristics of SV, many researchers have aimed to extract discriminative speaker embeddings from utterances by exploiting deep neural networks (DNNs). A speaker embedding is a high-dimensional feature vector that represents the characteristics of a speaker. This paper focuses on studying an improved training method of a DNN-based SV system to explore a better speaker separable representation space.

We noted from the results of a study that solely averaging DNN parameters after each step in the training phase can converge to better local minima [4]. This technique is called "temporal averaging"; the temporal averaging of weights can lead to more stable and accurate results than the final weights when the training has been completed. Leveraging this knowledge, in the field of semi-supervised learning, Tarvainen et al. [5] proposed a novel framework that uses a temporal averaging model, called mean teacher (MT), to use unlabeled data for training. The MT framework comprises a teacher–student [6] setting, where the teacher network (i.e., MT) is updated with an exponential moving average (EMA) of a set of student network parameters at each training step. In other words, the MT is the temporal averaging model of the student network and can generate a relatively

accurate pseudo label for unlabeled data. Thus, the student network learns unlabeled data by reducing the Euclidean distance from the MT network predictions.

The MT network can be regarded as a temporal ensemble model in terms of aggregating information about the student network at each training step. In addition, the MT model, as a teacher network, provides high-quality predictions for the student network to learn unlabeled data stably [5]. From these perspectives, we consider the MT network to be an ensemble teacher [7] and hypothesize that it can be a sufficient auxiliary model to assist student network training even in the supervised learning domain. Therefore, this study proposes a method that can enhance the generalization performance of the SV task by adapting the MT framework in a supervised learning condition.

We modified the following two factors in the existing MT framework. First, SV is essentially a task of comparing the similarity between embeddings. Therefore, we let the student network learn directly from the speaker embedding output from the MT network, not the pseudo speaker labels. Second, the consistency loss function between the student and the MT network is changed to a cosine similarity-based metric learning that uses the negative pairs together, rather than the mean square error (MSE), which considers only positive pairs. Thus, the student network can explore the embedding space suitable for SV tasks with a small intraclass variance and large interclass variance by mimicking the evaluation scenario in the training process [8].

The contributions of this study are:

- We introduced a novel supervised training method utilizing the MT framework for the SV task of the open-set evaluation condition.
- We analyzed the effectiveness of the proposed framework's components through ablation and comparison experiments.

To train and evaluate the models, we used the entire VoxCeleb2 [9] and the VoxCeleb1 [10] datasets, respectively. The experimental results revealed that the proposed method demonstrated a relative error reduction of 11.61% compared to the baseline system.

2. Related Work

DNN-based open-set speaker verification: The majority of the early work on DNN-based SV usually trained a model for the speaker classification task on a training dataset, using classification-based objective functions, such as softmax [11,12]. The last output layer of the trained model was omitted, and the output of the final hidden layer was exploited as a speaker embedding. However, open-set SV requires a discriminative embedding space for unseen speakers rather than an accurate classification on training datasets. Therefore, several studies have employed variants of the softmax loss function applying an additional angular margin to reduce the variance within a class [13,14]. In contrast, open-set SV can essentially be treated as a metric learning problem [2,8]. Thus, many recent studies have improved the generalization performance by optimizing metric-learning-based cost functions rather than classification-based cost functions [15,16]. They trained the model to decrease the distances between speaker embeddings extracted from the same speaker utterances and increase the distances between speaker embeddings of different speakers. In addition to these loss functions, the prior studies have investigated diverse methods, such as data augmentation [17,18], network architectures [19,20], and system frameworks [21,22].

Temporal averaging and mean teacher: Loss values oscillating or bouncing without convergence during DNN training are a common training failure indicator [23]. To alleviate this issue, Polyak et al. [4] proposed a temporal averaging method that combines the weights collected until the end of the training. In the loss landscape, temporal averaging brings it closer to the bottom of the valley by averaging the weights of the points that oscillate back and forth [24]. Therefore, this method can be used to improve the convergence of an optimization algorithm [25] or further enhance the generalization performance of the model during the evaluation phase [26].

Furthermore, the MT network, a temporal averaging model, was also used as a teacher model in semi-supervised image classification tasks [5]. Concretely, the MT network pro-

vides relatively accurate pseudo labels of unlabeled data for training the student network. The architectures of the student and MT networks are identical. The weight set of the MT network, ξ, is updated through the EMA of student weights, θ, formulated as follows:

$$\xi_t = \alpha\xi_{t-1} + (1-\alpha)\theta_t, \tag{1}$$

where t indicates the training step, and α is a smoothing coefficient hyperparameter. However, the parameter set of the student network is trained via the backpropagation algorithm. The loss function of the student network is a linear combination of classification and consistency costs. The classification cost is only applied to labeled data and adopts a categorical cross-entropy (CCE) function. For all data, including unlabeled data, the consistency loss function (e.g., MSE) is constrained to reduce the Euclidean distance between the prediction of the student and MT networks.

The critical point of this method is to form an MT network that progressively aggregates information from the student network in an EMA fashion. Consequently, the MT network is expected to produce more reliable probability predictions that serve as a high-quality representation to guide the student network training.

3. Baseline

Recently, based on the assumption that handcrafted features may not be optimal, data-driven systems fed by less processed features, such as spectrograms or raw waveforms, have been explored in audio domains [10,27–29]. In the SV field, it is hypothesized that a model directly fed the raw waveform can appropriately aggregate various frequency bands of utterances and potentially extract discriminative features by the first one-dimensional convolution layer [29,30]. Therefore, we used raw waveforms as input in all experiments and exploited RawNet2 [31,32], a representative neural speaker embedding extractor based on raw waveform input, as the baseline system.

Table 1 describes the structure of the baseline with several modifications to improve the performance of the original RawNet2. First, we increased the number of residual blocks [20] in the model from six to eight. Second, instead of the gated-recurrent unit, we employed attentive statistics pooling [33] to aggregate frame-level features into an utterance level feature. Third, we reduced the output dimension (embedding size) of the fully connected layer from 1024 to 512. More specific details related to the system architecture are described in the literature [31] and the author's GitHub (https://github.com/Jungjee/RawNet, accessed on 13 October 2020).

Table 1. DNN architecture of the baseline system. L is the length of the input sample. Conv refers to the convolutional layer, and Res means the residual block. For convolutional layers, numbers inside parentheses refer to the filter length, stride size, and number of filters. ASP and FC are the attentive statistics pooling and fully connected layers, respectively. AFMS represents the alpha feature map scaling module, proposed by [31].

Block	Block Structure	# Blocks	Output Shape
1D-Conv	Conv(3, 3, 128)	1	$L/3 \times 128$
Res1	BN and LeakyReLU Conv(3, 1, 128) BN and LeakyReLU Conv(3, 1, 128) Maxpool(3) AFMS	2	$L/3^3 \times 128$

Table 1. *Cont.*

Block	Block Structure	# Blocks	Output Shape
Res2	BN and LeakyReLU Conv(3, 1, 256) BN and LeakyReLU Conv(3, 1, 256) Maxpool(3) AFMS	3	$L/3^6 \times 256$
Res1	BN and LeakyReLU Conv(3, 1, 512) BN and LeakyReLU Conv(3, 1, 512) Maxpool(3) AFMS	3	$L/3^9 \times 512$
Pooling	ASP	1	1024
Embedding	FC(512)	1	512

4. Proposed Method

We aim to achieve a better generalization performance for open-set SV scenarios. By leveraging the knowledge that a network with temporal averaging weights can yield more accurate and stable results [4], we expect that the temporal averaging model can be a sufficient auxiliary model that provides pseudo labels in addition to ground truth hard labels. Therefore, this study proposes to adapt the MT [5] framework for supervised learning SV. The following subsections describe the structure and process of the proposed method.

4.1. Architecture

In semi-supervised learning, the conventional frameworks learn a manifold of unlabeled data by reducing the distance between the predictions of the teacher and student networks after applying different augmentation or noise to the same input [5,34]. However, in supervised learning in which all labels exist, it is more effective to fully use the labels rather than the indirect learning scheme described above. Thus, we intend to design a model to directly decrease the within-class variance by explicitly reducing the distance between embeddings from different data of the same class (i.e., different utterances of the same speaker).

Figure 1 depicts the overall structure of the proposed method. Let M be a minibatch, $M \in \mathbb{R}^{N \times U \times T}$ where N, U, and T refer to the number of speakers, number of utterances per speaker, and length of a sample, respectively. In addition, M is divided into m and m' for the input of the two networks. Each of these minibatches has half of the utterances for each speaker, where $m, m' \in \mathbb{R}^{N \times \frac{U}{2} \times T}$. Thus, m and m' comprise the pairs of different utterance sets for each speaker.

The student and MT networks consist of an encoder f and converter g, where the projector q is added only to the student network. θ and ξ indicate the set of parameters for the student and MT networks, respectively. Table 2 describes the student network architecture of the proposed method. The encoder extracts the representations from the input utterances. We used RawNet2, the baseline model (see Table 1) of this study, as the encoder. Then, the converter transforms the intermediate embedding extracted from the encoder into the speaker embeddings in a new space for metric learning. Additionally, the student network exploits the projector to project the embedding into a different space the from MT network, avoiding direct predictions in the same representation space to prevent the collapse of the embedding [35]. Hence, the student network can explicitly explore expansive embedding spaces different from the MT network, increasing the discriminability [35]. The converter and projector have identical structures. Finally, the student and MT networks extract the

embeddings Z and Y' from the separated minibatches m and m', respectively, which are formulated as follows:

$$Z = q_\theta(g_\theta(f_\theta(m))), \quad Y' = g_\xi(f_\xi(m')), \tag{2}$$

where $Z, Y' \in \mathbb{R}^{N \times \frac{U}{2} \times D}$, and D is the embedding dimension (512 in this study). After training is completed, the SV performance is evaluated using the embeddings Z extracted from the student network, as indicated by the red arrow in Figure 1.

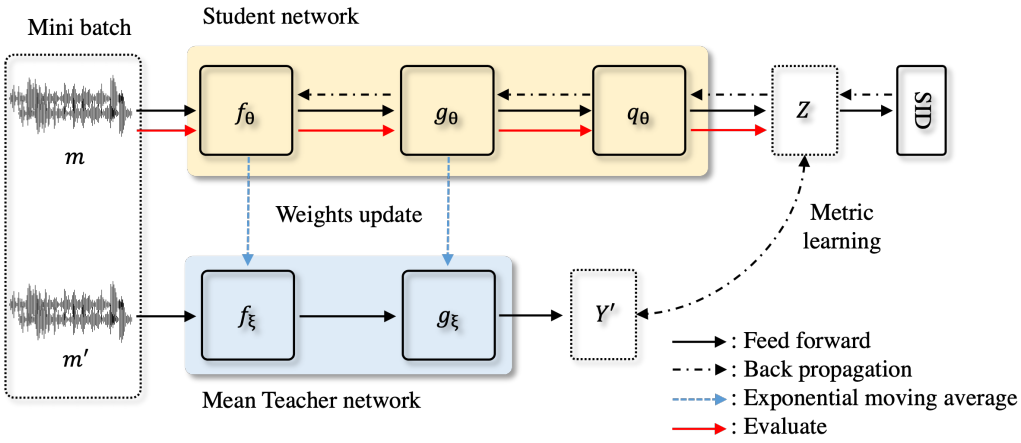

Figure 1. Overall structure of the proposed method. The student and mean teacher (MT) networks comprise an encoder f and converter g, and the projector q is added only to the student network. The two networks are fed by separate minibatches m and m', and extract the embeddings Z and Y', respectively. The student network is trained by cosine similarity-based metric learning using Y', along with speaker identification. The parameter set of the MT network ξ is updated with the exponential moving average of the student network parameter set θ.

Table 2. The student network architecture of the proposed method. RawNet2 used as the baseline in this study is exploited as the encoder. The mean teacher network has an identical structure as the student network except for the projector.

Block	Block Structure	Output
Encoder	RawNet2	512
Converter	FC(512) BN and LeakyReLU FC(512)	512
Projector	FC(512) BN and LeakyReLU FC(512)	512

4.2. Model Update

The MT network parameter set ξ is updated via the EMA of the student network parameter set θ as in Equation (1). We set the smoothing coefficient hyperparameter to 0.99. Therefore, the MT network linearly reduces the weight of the past student network, which is far from optimal, and aggregates the latest information. Consequently, we expect the MT network to provide stable and accurate speaker embeddings to train the student network to be a temporal ensemble teacher.

In contrast to the MT network, the weights of the student network are updated using the backpropagation algorithm. There are two loss functions for the student network: consistency and classification costs. First, to discriminatively project the utterances of

unseen speakers to the representation space, it is necessary to learn to reduce intraclass variance and increase interclass variance. Thus, we designated the consistency cost as a cosine similarity-based metric learning loss function using embeddings from the two networks, rather than the MSE between predictions, which considers only positive pairs, as in a previous study [5]. It is expected that the embedding space can yield greater interclass variance using a negative pair as well as a positive pair between the two network embeddings. In addition, using the cosine similarity metric, which is a measure of similarity between the embeddings in SV evaluation, it is possible to directly learn the embeddings that have a smaller intraclass variance for SV.

We used a centroid-based metric learning loss function, such as the generalized end-to-end (GE2E) [16] or angular prototypical (AP) network loss function [8]. In the case of GE2E, we modified it to fit the proposed method. The original GE2E loss operates using the cosine similarity between each query embedding and the centroids in the minibatch. Herein, the query refers to each utterance for all speakers, and the centroid c is derived by averaging embeddings that belong to the same speaker. If the speaker of a query and a centroid are equal, the centroid is calculated while excluding that query. In the proposed method, only the student network is backpropagated. Hence, only the embedding from the student network Z is used as a query during training and is referred to as half GE2E (GE2E-H).

$$c_j = \frac{1}{U}\sum_{l=1}^{U/2}(Z_{jl} + Y'_{jl}), \tag{3}$$

$$c_j^{(-i)} = \frac{1}{U-1}(\sum_{\substack{l=1 \\ l\neq i}}^{U/2} Z_{jl} + \sum_{l=1}^{U/2} Y'_{jl}), \tag{4}$$

where c_j and $c_j^{(-i)}$ represent the centroid of the jth speaker and the centroid calculated excluding the ith utterance, respectively. In addition, Z_{jl} is the lth student network embedding of the jth speaker. The similarity matrix S is defined as the scaled cosine similarity between the student network embeddings and all centroids:

$$S_{ji,k} = \begin{cases} w \cdot cos(Z_{ji}, c_j^{(-i)}) + b & \text{if } k = j, \\ w \cdot cos(Z_{ji}, c_k) + b & \text{otherwise} \end{cases}, \tag{5}$$

where $S_{ji,k}$ denotes the scaled cosine similarity between the jth speaker's ith student network embedding and the kth speaker's centroid, and w and b are learnable parameters. The final GE2E-H loss is as follows:

$$L_{G_H} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{U/2} log\frac{exp(S_{ji,j})}{\sum_{k=1}^{S} exp(S_{ji,k})}. \tag{6}$$

Along with speaker identification training over the entire training dataset for additional discriminative power [36], the loss function of the student network is defined as follows:

$$L_S = L_{G_H} + L_C, \tag{7}$$

where L_C is the classification cost, specified as CCE. The loss also computes the symmetric $\tilde{L}_S$ by entering the opposite minibatch to train the student network across the entire training set. Thus, $\tilde{L}_S$ is derived by feeding m' to the student network and m to the MT network, as in [35]. The final loss function of the proposed method is as follows:

$$L = (L_S + \tilde{L}_S)/2. \tag{8}$$

5. Experiments and Result

5.1. Dataset

We used the VoxCeleb2 dataset [9] for training, and the VoxCeleb1 dataset [10] for evaluation. The VoxCeleb2 dataset comprises over one million utterances from 6112 speakers. We exploited three test sets. The original test set (Vox1-O) consists of 37,611 enrollment-test utterance pairs from 40 speakers. Second, the extended test set (Vox1-E) contains a list of 579,818 pairs. Finally, the hard test set (Vox1-H) comprises a list of 550,894 pairs with the same nationality and gender. The Vox1-E and Vox1-H test sets contain 1251 and 1190 speakers from the entire VoxCeleb1 dataset. Every utterance is recorded in mono with a 16 kHz sampling rate and 16-bit resolution. In addition, we augment the input data using room-impulse response simulation and MUSAN corpus [37].

5.2. Experimental Configurations

We employed the raw waveforms as input, and the minibatch was configured by setting the length of the input utterance to 59,049 samples (≈3.69 s) in the training phase. In the testing phase, to prevent a mismatch with the training utterance input length, we applied a test time augmentation by sampling 10 temporal segments at regular intervals to make the length of an input utterance the same as in the training [9].

The baseline and the proposed systems were trained using 200 and 1920 batch sizes, respectively, which reported the best performance empirically in each experiment. In baseline experiments, the AMSGrad optimizer [38] was used with a learning rate (LR) of 0.001, decaying exponentially with every iteration. Furthermore, we used a LARS [39] optimizer with an LR of 3 and cosine annealing LR policy [40] with warm up [41] for the first three epochs in the proposed system. These are the optimal combination of learning hyperparameters found by the evaluation results of internal experimentation for each experiment. We applied a weight decay with $\lambda = 0.0001$.

5.3. Results

We compared the performance of the experiments conducted in this paper with various recently reported studies and display the results in Table 3. The baseline system is a RawNet2-based model with several modifications, and reported improved performances based on the equal error rate (EER) compared to the original RawNet2. Moreover, the result of the proposed model exhibited further enhanced performance compared to the baseline system used as the encoder. This result indicates that the supervised MT framework proposed in this study can improve the generalization of SV system. In addition, the proposed system demonstrates superior performance compared to each model that feeds various input features, including the raw waveforms.

Table 3. Performance comparison results with other speaker verification systems with various models and input features. C: contrastive.

Model	Features	Loss	Vox1-O	Vox1-E	Vox1-H
			EER (%)	EER (%)	EER (%)
ResNet-50 [9]	Spectrogram	Softmax+C	3.95	4.42	7.33
Thin ResNet-34 [42]		Softmax	2.87	2.95	4.93
Stats-vector [43]	MFCC	Softmax	3.29	3.39	5.94
ResNet-34-SE [44]		AS-softmax	3.10	3.38	5.93
RawNet2 [32]	Raw waveform	Softmax	2.48	2.57	4.89
Y-vector [45]		AM-softmax	2.72	2.38	3.87
Our baseline		Softmax	2.24	2.18	4.08
Our proposed		GE2E-H+Softmax	1.98	1.88	3.81

To analyze the effectiveness of the proposed framework's components, we performed comparison and ablation experiments. Table 4 presents the performance according to the modification from the existing MT framework to the proposed method. All systems presented in Table 4 use a 1920 batch size and exploit the CCE loss function for classification training, along with their respective consistency loss functions. In the first row, NP refers to using negative pairs, BC indicates batch configurations, S represents using the same minibatch with added noise as the conventional MT, and D refers to the minibatch comprising different utterances from each speaker. In addition, in the LT column, which represents the learning target of the student network, P and E refer to the prediction and embedding of the MT network, respectively. System #1, which reflects the original MT framework training scheme, exhibited a noticeable performance decline compared to the baseline (2.24% vs. 4.98% on Vox1-O). Therefore, we modified the original MT framework (System #1) to be suitable for supervised SV tasks. System #2 encourages the student network to learn embeddings of the MT, not predictions, considering the characteristics of the SV task. In Systems #3 and #4, we exploited the minibatch comprising different utterance pairs for each speaker and let the student network learn to maximize the distance between negative pairs to fully employ all labels. When comparing Systems #1 and #4, it is effective to consider the negative pairs and learn the speaker embeddings of the MT with minibatches of different utterances. However, it did not exceed the baseline performance. The comparison of Systems #4 to #7 demonstrated that cosine similarity-based metric learning could improve the performance of SV. Finally, System #7, the proposed method, reported an EER of 1.98% on the Vox-O trial, with a relative error reduction of 11.61% compared with the baseline.

Table 4. Comparison and ablation experiment results of the proposed method. NP: negative pair, BC: batch configuration, S: same batch, D: different batch, LT: learning target, P: prediction, E: embedding.

System	Consistency Loss	NP	BC	LT	Vox1-O	Vox1-E	Vox1-H
					EER (%)	EER (%)	EER (%)
#1 (Org_MT)	MSE	×	S	P	4.98	4.86	8.61
#2		×	S	E	3.56	3.15	5.61
#3		×	D	E	2.37	2.18	4.35
#4		✓	D	E	2.28	2.13	4.22
#5	GE2E	✓	D	E	2.27	2.21	4.3
#6	AP	✓	D	E	2.18	2.05	4.11
#7 (Proposed)	GE2E-H	✓	D	E	1.98	1.88	3.81

Table 5 displays the comparison results according to the batch size of the baseline system and proposed method. For the proposed method, the batch size is the product of the number of speakers and number of utterances per speaker included in a single minibatch. The experimental results indicate that, in the proposed framework, unlike the baseline system, a large batch size is effective and exhibits the best performance when the batch size and number of utterances per speaker are 1920 and 4, respectively. These results can be posited as follows: the proposed method works stably when performing metric learning in large-scale batches using more accurate embedding from the temporal averaging model, MT.

Table 5. Performance comparison of the baseline and proposed framework according to the batch size and number of utterances per speaker on the Voxceleb1 test set.

Batch Size	Baseline	Proposed Method		
		# Utterances per Speaker		
		4	6	8
200	2.24	2.1	2.23	3.04
800	2.36	2.12	2.07	2.19
1920	2.51	1.98	2.09	2.28

6. Conclusions

We proposed applying the MT framework to improve the generalization performance of open-set SV tasks. The conventional MT framework is used in a semi-supervised manner, with the same data for minibatches but different augmentations for the teacher and student networks. In this study, we used the MT network in a supervised manner with minibatches of different utterances from each speaker, increasing the cosine similarities of the embeddings from the two networks, decreasing the intraclass variances. We also investigated a method to determine discriminative embedding spaces suitable for open-set SV using diverse cosine similarity-based metric learning cost functions with positive and negative embedding pairs. The proposed system exhibited an EER of 1.98% on the VoxCeleb1 test set with a relative error reduction of 11.61% compared to the baseline system, the improved RawNet2. In future work, we intend to conduct further experiments to demonstrate the superiority of the proposed method for SV tasks using encoders with different input features, other than raw waveforms.

Author Contributions: Conceptualization, J.-H.K., H.-J.S. and J.-W.J.; Investigation, J.-H.K.; Supervision, H.-J.S., J.-W.J. and H.-J.Y.; Writing—original draft, J.-H.K.; Writing—review & editing, J.-H.K., H.-J.S., J.-W.J. and H.-J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported and funded by the Korean National Police Agency. [Project Name: Real-time speaker recognition via voiceprint analysis/Project Number: PR01-02-040-17].

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hansen, J.H.; Hasan, T. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Process. Mag.* **2015**, *32*, 74–99. [CrossRef]
2. Bai, Z.; Zhang, X.L. Speaker recognition based on deep learning: An overview. *Neural Netw.* **2021**, *140*, 65–99. [CrossRef] [PubMed]
3. Bendale, A.; Boult, T.E. Towards open set deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1563–1572.
4. Polyak, B.T.; Juditsky, A.B. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **1992**, *30*, 838–855. [CrossRef]
5. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv* **2017**, arXiv:1703.01780.
6. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
7. Freitag, M.; Al-Onaizan, Y.; Sankaran, B. Ensemble distillation for neural machine translation. *arXiv* **2017**, arXiv:1702.01802.
8. Chung, J.S.; Huh, J.; Mun, S.; Lee, M.; Heo, H.S.; Choe, S.; Ham, C.; Jung, S.; Lee, B.J.; Han, I. In Defence of Metric Learning for Speaker Recognition. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020.
9. Chung, J.S.; Nagrani, A.; Zisserman, A. VoxCeleb2: Deep Speaker Recognition. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018.
10. Nagrani, A.; Chung, J.S.; Zisserman, A. Voxceleb: A large-scale speaker identification dataset. *arXiv* **2017**, arXiv:1706.08612.
11. Variani, E.; Lei, X.; McDermott, E.; Moreno, I.L.; Gonzalez-Dominguez, J. Deep neural networks for small footprint text-dependent speaker verification. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4052–4056.

12. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.
13. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5265–5274.
14. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.
15. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4080–4090.
16. Wan, L.; Wang, Q.; Papir, A.; Moreno, I.L. Generalized end-to-end loss for speaker verification. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4879–4883.
17. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May, 2018.
18. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19. September 2019; pp. 2613–2617. [CrossRef]
19. Snyder, D.; Garcia-Romero, D.; Povey, D.; Khudanpur, S. Deep Neural Network Embeddings for Text-Independent Speaker Verification. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 999–1003.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Sang, M.; Xia, W.; Hansen, J.H. Open-set Short Utterance Forensic Speaker Verification using Teacher-Student Network with Explicit Inductive Bias. *arXiv* **2020**, arXiv:2009.09556.
22. Tao, F.; Tur, G. Improving Embedding Extraction for Speaker Verification with Ladder Network. *arXiv* **2020**, arXiv:2003.09125.
23. Vogl, T.P.; Mangis, J.; Rigler, A.; Zink, W.; Alkon, D. Accelerating the convergence of the back-propagation method. *Biol. Cybern.* **1988**, *59*, 257–263. [CrossRef]
24. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, UK, 2016; Volume 1.
25. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
26. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
27. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, 13–15 February 2017; pp. 1–5.
28. Fu, S.W.; Tsao, Y.; Lu, X.; Kawai, H. Raw waveform-based speech enhancement by fully convolutional networks. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 006–012.
29. Ravanelli, M.; Bengio, Y. Speaker recognition from raw waveform with sincnet. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 1021–1028.
30. Jung, J.W.; Heo, H.S.; Kim, J.H.; Shim, H.J.; Yu, H.J. RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. *arXiv* **2019**, arXiv:1904.08104.
31. Jung, J.W.; Shim, H.J.; Kim, J.H.; Yu, H.J. α-feature map scaling for raw waveform speaker verification. *J. Acoust. Soc. Korea* **2020**, *39*, 441–446.
32. Jung, J.W.; Kim, S.B.; Shim, H.J.; Kim, J.H.; Yu, H.J. Improved RawNet with Feature Map Scaling for Text-independent Speaker Verification using Raw Waveforms. *arXiv* **2020**, arXiv:2004.00526.
33. Okabe, K.; Koshinaka, T.; Shinoda, K. Attentive Statistics Pooling for Deep Speaker Embedding. *arXiv* **2018**, arXiv:1803.10963.
34. Samuli, L.; Timo, A. Temporal ensembling for semi-supervised learning. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017; Volume 4, p. 6.
35. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Guo, Z.D.; Azar, M.G.; et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv* **2020**, arXiv:2006.07733.
36. Kye, S.M.; Jung, Y.; Lee, H.B.; Hwang, S.J.; Kim, H.R. Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs. *arXiv* **2020**, arXiv:2004.02863.
37. Snyder, D.; Chen, G.; Povey, D. Musan: A music, speech, and noise corpus. *arXiv* **2015**, arXiv:1510.08484.
38. Reddi, S.J.; Kale, S.; Kumar, S. On the Convergence of Adam and Beyond. In Proceedings of the International Conference on Learning Representations, Vancouver, Canada, 30 April–3 May 2018.
39. You, Y.; Gitman, I.; Ginsburg, B. Large batch training of convolutional networks. *arXiv* **2017**, arXiv:1708.03888.
40. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
41. Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; He, K. Accurate, large minibatch sgd: Training imagenet in 1 h. *arXiv* **2017**, arXiv:1706.02677.

42. Nagrani, A.; Chung, J.S.; Xie, W.; Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.* **2020**, *60*, 101027. [CrossRef]
43. Hong, Q.B.; Wu, C.H.; Wang, H.M.; Huang, C.L. Statistics Pooling Time Delay Neural Network Based on X-Vector for Speaker Verification. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6849–6853. [CrossRef]
44. Zhou, J.; Jiang, T.; Li, Z.; Li, L.; Hong, Q. Deep Speaker Embedding Extraction with Channel-Wise Feature Responses and Additive Supervision Softmax Loss Function. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 2883–2887.
45. Zhu, G.; Duan, Z. Y-vector: Multiscale waveform encoder for speaker embedding. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021.