Odyssey 2020 The Speaker and Language Recognition Workshop
1-5 November 2020, Tokyo, Japan

Selective Deep Speaker Embedding Enhancement for Speaker Verification

Jee-weon Jung, Ju-ho Kim*, Hye-jin Shim, Seung-bin Kim, and Ha-Jin Yu*†

School of Computer Science, University of Seoul, Republic of Korea

jeewon.leo.jung@gmail.com, wngh1187@naver.com, shimhz6.6@gmail.com,
kimho1wq@naver.com, hjyu@uos.ac.kr

Abstract

Utterances that are input from a distance are one of the major causes of performance degradation in speaker verification systems. In this study, we propose two frameworks for deep speaker embedding enhancement and specifically focus on distant utterances. Both frameworks input speaker embeddings extracted from front-end systems, including deep neural network-based systems, which widen the range of applications. We use speaker embeddings that are extracted by inputting raw waveforms directly into a deep neural network. The first proposed system, skip connection-based selective enhancement, adopts a skip connection that directly connects the input embedding to the output. This skip connection is multiplied by a value between 0 and 1, which is similar to the gate mechanism where the value is concurrently determined by another small deep neural network. This approach allows the selective application of enhancements, thus, when the input embedding is from a close-talk, the skip connection would be more activated. On the other hand, when embedding from a distance is input, the deep neural network would be more activated. The second proposed system, i.e., a selective enhancement discriminative auto-encoder, aims to find a discriminative representation with an encoder–decoder architecture. The hidden representation is divided into two subspaces with the objective to gather speaker information into one subspace by adding additional objective functions and letting the other subspace contain subsidiary information (e.g., reverberation and noise). The effectiveness of both proposed frameworks is evaluated using the VOiCES from a Distance Challenge evaluation set and demonstrates a 11.03 % and 15.97 % relative error reduction, respectively, compared to the baseline, which does not employ an explicit feature enhancement phase.

1. Introduction

Owing to the recent advances in deep learning across various applications, deep neural networks (DNNs) are exhibiting state-of-the-art performance in both text-dependent and text-independent speaker verification (SV) [1–6]. Although recent SV systems have shown reliable performance in clean and close-talk scenarios, distant scenarios, which include various environmental factors (e.g., reverberation and background noise), are known to degrade the performance [7, 8]. To accelerate studies on SV in distant, noisy scenarios, the Voices Obscured in Complex Environmental Settings (VOiCES) from a Distance Challenge 2019 was held [9, 10]. The data used in the VOiCES challenge was collected by re-recording the pre-recorded utterances in Librispeech dataset [11] under various realistic conditions (see Section 3 for details on VOiCES).

A number of studies have proposed techniques in various aspects to compensate for the performance degradation caused by distant input speech [8, 12–14]. These studies have been shown to be effective in distant utterance compensation. However, according to previous studies [15], the degradation of close-talk utterance performance may occur when it is input into the compensation system. Owing to this phenomenon, it is difficult to use compensation methods directly in an environment in which utterances from various distances are input. In addition, dependency exists that, when a new speaker embedding extractor is proposed, corresponding studies for adequate dereverberation and denoising should be performed.

To mitigate these issues, we propose two distant utterance compensation systems that are independent of the front-end speaker embedding extractor. Both systems selectively perform feature enhancement, which includes reverberation and noise, using a separate speaker embedding enhancement phase. In this study, the authors propose speaker embedding enhancement systems that fulfill the following three properties, and the third property is related to practical issues:

- Both close-talk and distant utterances can be inputted into the proposed system
- Proposed systems should exist independently of the front-end speaker embedding systems
- Proposed systems should comprise a relatively simple architecture that causes minimal overhead to the overall SV process pipeline

The first proposed system is designed to selectively conduct enhancement according to the required extent of compensation when speaker embedding is input. The intuition behind this framework is that rather than dichotomously dividing a particular input utterance into a close-talk or distant utterance, it would be better to determine whether there is relatively less or more reverberation and noise and conduct compensation accordingly. We designed the proposed system to meet this reasoning by internally having a separate DNN to determine the extent of skip connection activation, which is similar to the gate mechanism. The second proposed system is based on the auto-encoder framework. By dividing the hidden representation into two subspaces, the system aims to separately extract speaker information included in embedding during encoding and decoding. One subspace is targeted to contain clean speaker information by applying additional objective functions to this hidden layer, and the other subspace is targeted to contain subsidiary information (e.g., reverberation and noise). The input features for the two systems proposed in this study are the speaker embeddings that are directly extracted from the raw waveform by the front-end speaker embedding extractor, which is used as a baseline [4]. To evaluate the performance of the SV system, the

*Equal contribution.
† Corresponding author

10.21437/Odyssey.2020-25

equal error rate (EER) was calculated using the cosine similarity.

The rest of this paper is organized as follows. Section 2 provides a brief overview of previous studies on distant utterance compensation. Section 3 provides details regarding the VOiCES dataset, and Section 4 describes the front-end speaker embedding system that was used as the baseline in this study. In Section 5, we introduce two proposed frameworks. Experiments and results are shown throughout Section 6 and the paper is concluded.

2. Previous works

Compensation for distant utterances has been conducted in several aspects in the process pipeline of SV, e.g., signal-level, acoustic feature-level, and DNN-level. Some studies have conducted the compensation of distant utterances at a signal level such as the linear prediction inverse modulation transfer function and weighted prediction error [16, 17]. In [18], a variance-normalized delayed linear prediction is proposed to determine an optimal dereverberation filter and cancel out late reverberation without significant damage to the original speech. Wang *et. al.* [19] have estimated the clean speech spectrum using a power spectral subtraction algorithm to remove noise. In addition, Han *et. al.* [20] have constructed a spectral mapper that converts noise and reverberation speech spectrogram into an anechoic speech spectrogram. Some studies, on the other hand, have concentrated on acoustic features that are robust to noise and reverberation [21, 22].

Compensation can be also performed by extracting reverberation and noise-mitigated speaker embeddings while training DNN [23] or by adopting techniques such as multi-task learning or knowledge distillation [15, 24]. In addition, distant utterances are used as a training set to construct a system that is robust to noise and reverberation. Data augmentation, which generates new training samples by synthesizing noise and reverberation in the train set, can be also conducted using techniques such as mix-up [25], spec augment [26], noise augment [27], and label smoothing [28]. For the VOiCES 2019 challenge, most participants focused on recent approaches such as learnable dictionary encoding [29], skip connection, and variants of ResNet-based [30] deep architectures.

3. VOiCES dataset

The VOiCES dataset was collected by playing the Librispeech [11] dataset through a loudspeaker and re-recording with array mics at a various distance and acoustic conditions to facilitate research on distant SV. Acoustic condition refers to four different rooms, 12 or 20 mics, different angles, and distractors such as TV, babble, and music. At its initial deployment, only development and evaluation sets were released for the VOiCES from a distance 2019 challenge, where the training set was a list of public datasets under fixed conditions: VoxCeleb1&2 [31, 32] and Speakers in the Wild (SITW) [33]). After the competition has ended, the full VOiCES dataset with meta information was released. In the full VOiCES dataset, there are $3,904$ source (close-talk) utterances (recorded by 300 speakers) and $999,424$ distant utterances, and there are 256 distant utterances per each source utterance. The challenge development set comprises a part of distant utterances of 200 speakers, and the challenge evaluation set comprises part of source and distant utterances of the rest 100 speakers. Note that, herein, we refer to close-talk utterances as a 'source' to avoid confusion with the naming

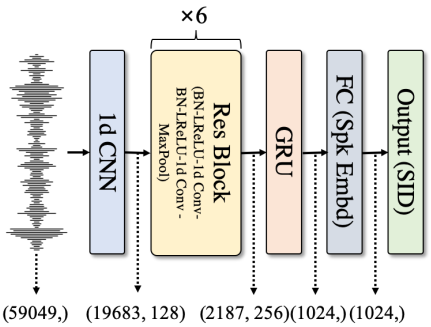

Figure 1: Illustration of the RawNet architecture. The number of input samples is fixed in the train phase to construct mini-batches. Numbers in the figure refer to the output shape with the configuration of this study. In the test phase, on the other hand, utterances with varying duration can be inputted where the different number of frame-level features is aggregated into a fixed dimensional utterance-level representation by a GRU.

protocol of the VOiCES dataset. Further details regarding the VOiCES dataset are provided in [9, 10].

In this study, we organized three subsets, i.e., train, development, and evaluation set. We composed the development and evaluation sets using the VOiCES challenge dataset and part of the full VOiCES dataset as the train set. The development and evaluation sets are identical to those from the VOiCES challenge to make the comparison possible with previous studies, which results in some overlap between the train and development sets. The train set includes 2602 source (Librispeech) utterances from 200 speakers and distant utterances simulated using these source utterances. This results in $600k+$ utterances in the train set. We excluded any utterances that belonged to one of 100 speakers among the full VOiCES dataset ($900k+$ utterances) to make sure that only unknown speakers existed in the evaluation set[1]. It would be important to note that even though the number of distant utterances in the train set is more than $600k+$, the actual diversity may not be as diverse as it seems in terms of providing variant data for training DNN. This occurs because severe duplication in reverberation and noise exists; recording two different source utterances in identical configuration will yield two distant utterances with similar reverberation and noise.

4. RawNet

Pre-processing methods before inputting an acoustic feature into DNN are becoming streamlined with recent advances in deep learning. When DNNs were first used to extract speaker embeddings, Mel-frequency cepstral coefficients (MFCCs) and log Mel-energy features were abundantly used [5]. These acoustic features utilize human knowledge to emphasize more discriminant features. Although these conventional acoustic features are still widely used, many recent studies also explore spectrograms and raw waveforms as input to DNN with the expectation that data-driven DNN can better extract discriminative information without human priors. When spectrograms are pro-

[1]It is ideal to exclude only the utterances that overlap with the evaluation set; however, this cannot be done because the meta information of utterances in the VOiCES challenge evaluation set is removed.

cessed by convolutional neural networks (CNNs), the receptive fields of CNN can consider only adjacent time and frequency regions in layers that are close to the input layer. The receptive field widens in deeper representations, close to the output layer. On the other hand, when raw waveforms are processed by CNNs, the aggregation of frequency responses across all frequency bands can be extracted. In addition, the frequency bands that aggregate are adaptive to the data and task, which is thought to be advantageous when directly using the raw waveform [4, 34]. Among frameworks that utilize raw waveform as an input, RawNet [4] is a recently proposed speaker embedding extractor for SV.

RawNet adopts a convolutional neural network-gated recurrent unit (CNN-GRU) architecture. The CNN part is a variant of ResNet [30], which includes max pooling according to [35] and extracts frame-level features from the raw waveform. Then, GRU aggregates varying frame-level features into a single utterance-level feature, which is independent of the duration of an input utterance. The output of GRU at the last timestep is connected to one fully-connected layer, which is used as the speaker embedding. The output layer indicates speaker identification results that are identical to the d-vector [5] framework; the output layer is removed after the training is complete. Figure 1 provides an overview of the RawNet system. In this study, we used speaker embeddings extracted from RawNet for both baseline and as input speaker embeddings for two proposed deep speaker embeddings enhancement systems.

5. Deep speaker embedding enhancement

In this session, we introduce two proposed systems. Both systems can input source utterances as well as distant utterances and are independent of the front-end speaker embedding extraction system. The following two subsections describe two systems for speaker embedding enhancement, respectively.

5.1. Skip connection-based selective enhancement (SCSE)

The first proposed system is referred to as skip connection-based selective enhancement (SCSE), which comprises a DNN that enhances speaker embedding (SEDNN), a skip connection, and another small skip decision DNN that decides the extent of activating the skip connection similar to the gate mechanism in GRU (SDDNN). The overall illustration of this system is provided in Figure 2. SEDNN is a denoising auto-encoder and SDDNN measures the extent of reverberation and noise residing in an input speaker embedding to modulate the degree of compensation by scaling the skip connection. A skip connection connects the input speaker embedding directly to the enhanced speaker embedding.

During the train phase, SEDNN is trained to minimize the mean squared error (MSE) objective function between the input and enhanced speaker embedding. When a source utterance x is input, the target x' is itself; when a distant utterance is input, the target is the source utterance that was used to make the distant utterance. Binary cross-entropy (BCE) objective function is used to train the SDDNN, where the binary label of 1 is given for source utterance to make the skip connection fully working. A binary label of 0 or 1 is multiplied by the input speaker embedding instead of SDDNN's output and added to the SEDNN's output.

In [36, 37], it has been reported that when compensation is conducted in the speaker embedding space, compensation may not be effective although the MSE value is very low. This phenomenon is analyzed as a result of losing the discriminative power of a speaker embedding by changing its values in a high-dimensional, abstract embedding space. Leveraging this knowledge, the last component of the proposed system is a speaker identification layer where a categorical cross-entropy (CCE) objective function is used. This component is added with the perspective to maintain the discriminative power of the enhanced speaker embedding. Hence, the final objective function used to train the SCSE system is described as:

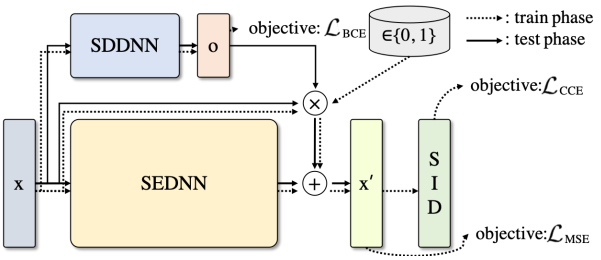

Figure 2: Illustration of the proposed skip connection-based selective enhancement system (Section 5.1). x and x' refer to the original and enhanced speaker embedding, respectively. Both SDDNN and SEDNN are simple DNN comprising few fully-connected layers.

$$\mathcal{L} = \mathcal{L}_{MSE} + \mathcal{L}_{BCE} + \mathcal{L}_{CCE}, \tag{1}$$

where $\mathcal{L}_{MSE}$ measures the reconstruction (dereverberation and denoising of distant utterances) error, $\mathcal{L}_{BCE}$ measures the distance detection error, and $\mathcal{L}_{CCE}$ measures the speaker identification error using the enhanced speaker embedding.

In the test phase, the speaker embedding x is input into SEDNN and SDDNN. Skip connection that connects x and x' is multiplied by the output of SDDNN with a sigmoid activation function, which is a real number between 0 and 1, and produces a scaled skip connection. Our intuition behind using the sigmoid activation function rather than, for example, the step function is because most input utterances cannot be decided dichotomously into source or distant. Enhanced speaker embedding is derived by adding the output of SEDNN and scaled skip connection. Using this scheme, in an ideal scenario, we expect that the skip connection will be more activated to conduct enhancement less for utterances with less noise and reverberation, and will be activated at its least to conduct more enhancement in the opposite case.

5.2. Selective enhancement discriminative auto-encoder (SEDA)

The second proposed system is referred to as selective enhancement discriminative auto-encoder (SEDA), which adopts an auto-encoder architecture that is composed of an encoder, decoder, and two parallel intermediate hidden layers (x' and n in Figure 3). Figure 3 illustrates the overall system. Speaker identification is conducted using the output to maintain discriminative power as in the previous subsection. x' is used as the enhanced speaker embedding, and n is intended to contain noise and reverberations.

The architectural design follows a discriminative auto-encoder (DCAE) structure, which has been proposed to extract x' representation that reduced intra-class variance and increased

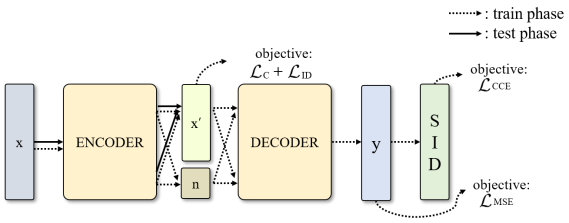

Figure 3: Illustration of the proposed selective enhancement discriminative auto-encoder system (Section 5.2). x and y refer to the original speaker embedding and the reconstructed, respectively. Both ENCODER and DECODER are a simple DNN comprising few fully-connected layers. x' and n refer to the enhanced speaker embedding and residual information hidden layer, respectively.

inter-class variance [38]. Inspired by DCAE, SEDA is configured to collect reverberation and noise in n and to contain clean speaker information in x'. To achieve this goal, four objective functions are adopted to train the SEDA system: MSE, center [39], internal dispersion [38], and CCE.

When training SEDA, the goal for y is reconstruction if the input is a source utterance and denoising and dereverberation into the source utterance of the identical condition when the input is a distant utterance. We used the MSE loss to achieve this goal which is described as:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\alpha||y_i - x_{si}||_2^2, \qquad \alpha(\cdot) = \begin{cases} 256, & if\ x_i = source, \\ 1, & if\ x_i = distance, \end{cases} \tag{2}$$

where y_i and x_{si} represent the system's output vector and source embedding of sample i, respectively. N is the size of mini-batch, α is a sample weight, and x_i is the input embedding. To match the unbalanced proportion between the number of source utterances and distance utterances in the training set, the sample weight of 256 and 1 is given according to the proportion in the train set.

SEDA uses the intermediate hidden layer x' as the enhanced speaker embedding, which is ideally noise-isolated, rather than the reconstruction y. To separate the noise from the embedding and improve the feature, additional objective functions are applied to x'. By adopting these objective functions for x' only, we aim to minimize the intra-class variance and maximize the inter-class variance; n was configured to contain the subsidiary information. We utilized center loss and internal dispersion loss. Center loss was presented to minimize intra-class variance while the embedding feature remains discriminative. To achieve this, the center loss function is used as

$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{N}||x'_i - c_{y_i}||_2^2, \tag{3}$$

where x'_i refers to the enhanced embedding of utterance i, and c_{y_i} refers to the center of class y_i.

The internal dispersion loss used in DCAE aims to maximize inter-class variance, which is given by

$$\mathcal{L}_{ID} = -\frac{1}{N}\sum_{i=1}^{N}||x'_i - H||_2^2, \tag{4}$$

where H denotes the empirical mean vector of x'_i.

The CCE objective function is also used as the objective function to conduct speaker identification using the reconstruction y. The final objective function of the proposed SEDA system can be described as:

$$\mathcal{L} = \mathcal{L}_{MSE} + \mathcal{L}_{CCE} + \gamma(\beta\mathcal{L}_C + (1-\beta)\mathcal{L}_{ID}) \tag{5}$$

where γ is a hyper-parameter that scales objective function applied to x', and β is a hyper-parameter that combines the objective function to reduce intra-class variance and increase inter-class variance at an appropriate ratio.

6. Experiments & results

6.1. Experimental configuration

All experiments in this paper were conducted using PyTorch [40], a deep learning toolkit written in Python. For the baseline speaker embedding extractor and the first proposed speaker embedding enhancement system, SCSE, the train set comprises a part of the VOiCES (excluding all utterances by speakers that coincide with those in the evaluation set), VoxCeleb1&2 datasets. The second proposed model, SEDA, was trained using the VOiCES dataset.

The baseline RawNet [4] system inputs raw waveform with pre-emphasis applied. The duration is adjusted to $59,049$ samples (≈3.69 s) for mini-batch construction in the train phase. Full duration without adjustment is used to extract speaker embeddings in the evaluation phase. The number of nodes in both GRU and speaker embedding layers is adjusted to $1,024$ by considering a larger training dataset. We also changed the architecture of residual blocks in RawNet from the initial ResNet [30] to identity mapping [41] with a max pooling according to [35].

For the SCSE system introduced in Section 5.1, SEDNN comprises three fully-connected layers with 1024 nodes. The leaky rectified linear unit (ReLU) activation [42] follows all fully-connected layers except the output layer. An output layer of $7,563$ nodes, which indicates the number of speakers, is connected to the enhanced speaker embedding, which is removed after training. SDDNN comprises two hidden layers with 256 and 64 nodes, also followed by the leaky ReLU activation, and an output layer with one node and a sigmoid activation function. The weight parameters of SDDNN are trained only for the first five epochs solely using the VOiCES dataset and are then fixed. This is done this way because, for utterances in VoxCeleb1&2 datasets, we cannot provide labels to train SDDNN. It is worth noting that in the first five epochs during the train phase, the skip connection is multiplied by an explicit label 0 or 1 instead of using SDDNN's output. The skip connection is multiplied by the output of SDDNN in the train phase after five epochs and in the evaluation phase.

In SEDA, both encoder and decoder each comprise two fully-connected layers, and the number of nodes in each layer is $1,024$. Hyperbolic tangent is used as the activation function. The number of nodes of x' and n is $1,024$ and 256, respectively.

6.2. Baseline composition

The baseline system uses the RawNet architecture with two modifications to fit data configurations of this study on the basis of the internal experiments. First, we increased the dimensionality of the speaker embedding to $1,024$. Second, we changed the number of the output layer's node to $7,563$

Table 1: Comparison of the baseline systems according to different configurations and previous studies. Systems with citation are reported performances from different papers. All performances are from a single system without an ensemble or calibration.

System	Input	Front&Back-end	EER
[44]	MFCC	x-vec/CSML	6.09
[45]	MFBANK	x-vec/PLDA	**5.65**
[46]	MFCC	ResNet/PLDA	6.18
ours-voi	waveform	RawNet/CosSim	9.98
ours-pretrn	waveform	RawNet/CosSim	10.50
ours-all	waveform	RawNet/CosSim	**7.70**

$(1, 251 + 6, 112 + 200)$. Lastly, center and between-class [43] objective functions were removed.

Table 1 describes few top performing teams' single system performances from the VOiCES challenge and our baseline system with various configurations. Top three rows [44–46] describe the performance of the top three teams in the VOiCES challenge, which are reported in single models. The evaluation EER, which is based on the best performing EER, for the development set is reported. These performances are provided as a reference because direct comparison cannot be made owing to the difference in the input feature, train set configuration, and back-end classifiers. The bottom three rows describe our three baselines with different training configurations. *ours-voi* describes the performance when using only the VOiCES dataset for training. In *ours-pretrn*, we first trained the network using VoxCeleb2 and then replaced only the output layer and conducted fine-tuning with the VOiCES dataset. *ours-all* shows the result of training all three datasets together.

The results show that among various configurations, training of all three datasets simultaneously provides the best performance. In our analysis, a relatively high EER, despite similar domain between train and evaluation in *ours-voi*, occurred owing to the severe overlap between utterances in the VOiCES train set. For example, the re-recording of ten source utterances using the same acoustic configuration produces ten additional utterances but does not provide the variations that result from recording new utterances in ten different acoustic configurations. Another phenomenon was also observed which is related to the direct use of raw waveforms as an input to DNN in our analysis. Different datasets have different amplitude where we hypothesize that DNNs which input raw waveform directly are more sensitive to amplitude differences. On the basis of comparing *ours-pretrn* and *ours-all*, we conclude that when an additional dataset with different amplitude is given, it is better to conduct training simultaneously rather than using one dataset for pre-training. We used *ours-all* as the baseline in this study because it demonstrates a reasonable performance without additional back-end classifiers.

6.3. Skip connection-based selective enhancement

Table 2 addresses the performance of the proposed SCSE system that was introduced in Section 5.1. Systems #1 through #8 show the result of comparative experiments using different optimizers [i.e., stochastic gradient descent (SGD) and Adam], learning rates, and learning rate schedulers. In *Lr_Sc*, iter refers to the decay of the learning rate every iteration, and cos refers to the cosine learning rate scheduling proposed in [47]. By comparing system #1 to #8, we conclude that the SCSE system is effective in most cases, showing a relative error reduction (RER) of 11.03 % in system #2. Systems #9 and #10 show the result of expanding the size of mini-batch to 10000 from system #2 and #7 on the basis of the successful result in Section 6.4, which did not show further improvement in the case of the SCSE system. The comparison of #2 and #11 shows the effectiveness of using speaker identification with the enhanced embedding addressed in Section 5.1. It is worth noting that for SCSE, removing VoxCeleb1&2 datasets from the train set resulted in an EER of 8.50 %, showing that using VoxCeleb1&2 datasets was essential for the successful performance. In addition, when we applied weights to the samples according to the source or distant, as described in (2), the SCSE system demonstrated an EER of 7.00 %, which did not produce additional performance improvements.

Table 2: Comparative experimental results of SCSE reported in EER and relative error reduction (RER, %) using different batch size (Bs), learning rate (Lr), learning rate scheduler (Lr_Sc), and optimizer (Optim).

Sys	Bs	Lr	Lr_Sc	Optim	EER	RER
Base	-	-	-	-	7.70	0
#1	256	0.01	Iter	SGD	7.18	6.75
#2	256	0.01	Cos	SGD	**6.85**	**11.03**
#3	256	0.001	Iter	SGD	7.28	5.46
#4	256	0.001	Cos	SGD	7.26	5.71
#5	256	0.01	Iter	Adam	7.67	0.38
#6	256	0.01	Cos	Adam	7.37	4.41
#7	256	0.001	Iter	Adam	**7.07**	**8.18**
#8	256	0.001	Cos	Adam	7.24	5.97
#9	10000	0.01	Cos	SGD	7.21	6.36
#10	10000	0.001	Iter	Adam	9.00	0
#11	256	0.01	Cos	SGD	8.30	-
#12	256	0.01	Cos	SGD	7.00	9.09

6.4. Selective enhancement discriminative auto-encoder

Comparative experiments using the proposed SEDA system were conducted to observe performance under various conditions and to determine the optimal condition. The results are described in Table 3. The train dataset of SEDA was the VOiCES dataset without the addition of the VoxCeleb1&2 datasets based on empirical experiments. The learning rate is set to 0.001, and λ for the weight decay is set to 0.0001 when the Adam optimizer is used. If the optimizer is SGD, the learning rate is set to 0.01, and we use the cosine learning rate scheduler. By comparing systems #1 through #5 to #6 through #10, we empirically determined that the use of Adam optimizer consistently demonstrated better performance than when SGD was used. By comparing systems #1 through #4, we conclude that there is no significant difference in performance when changing γ and β. Finally, systems #2 and #5 show that the use of a large mini-batch size of $10,000$ improves the performance of SEDA.

6.5. Score normalization

Score normalization techniques are frequently employed in various acoustic mismatch conditions. Most of the participants in the VOICES 2019 challenge also used score normalization techniques such as zero score normalization (z-norm) [48], test score normalization (t-norm) [49], and symmetric normalization (s-norm) [50]. z-norm applies imposter score distribution

Table 3: Comparative experimental results of SEDA reported in EER and relative error reduction (RER %) using a different optimizer (Optim), batch size (Bs), γ, and β.

Sys	Optim	Bs	γ	β	EER	RER
Base	-	-	-	-	7.70	0
#1	Adam	10000	0.001	0.2	6.49	15.71
#2	Adam	10000	0.001	0.8	**6.47**	**15.97**
#3	Adam	10000	0.0001	0.2	6.48	15.84
#4	Adam	10000	0.0001	0.8	6.48	15.84
#5	Adam	256	0.001	0.8	7.48	2.86
#6	SGD	10000	0.001	0.2	8.25	0
#7	SGD	10000	0.001	0.8	7.61	1.17
#8	SGD	10000	0.0001	0.2	7.72	0
#9	SGD	10000	0.0001	0.8	7.75	0
#10	SGD	256	0.0001	0.8	**7.41**	**3.77**

Table 4: Various score normalization techniques applied to the baseline and top performing systems from the two proposed frameworks.

System	$\times$	z-norm	t-norm	s-norm
Baseline	7.70	**7.30**	7.83	7.31
SCSE	6.85	6.58	6.84	**6.53**
SEDA	6.47	**6.31**	6.90	6.41
Ensemble	6.36	**6.19**	6.61	6.41

to the enrollment data. In contrast to the z-norm, t-norm employs imposter score distribution for the test data. s-norm uses the average of z-norm and t-norm.

We experimented the effectiveness of these techniques for our baseline and for two proposed systems, SCSE and SEDA, and report the results in Table 4. Bold font indicates the best performance for each system. The results show that score normalization consistently improved performance in most cases. z-norm demonstrated the best performance in most cases in our experiments. In addition, the score-sum ensemble of the two proposed systems has been also explored. The results show that the ensemble of two systems can lead to additional performance improvement with an EER of 6.19 % for z-norm.

7. Conclusion

In this study, we addressed the need for the selective speaker embedding enhancement independent from the front-end speaker embedding extraction and proposed two systems to conduct feature enhancement. Both proposed systems comprise simple DNNs, and these systems can process not only distant utterances but also close-talk (source) utterances. This approach mitigates the issue of performance degradation when close-talk utterances are input into the speaker embedding enhancement system that is designed for distant utterances. Compared to the baseline system, the two proposed systems (SCSE and SEDA) demonstrated an RER of 11.03 % and 15.93 % respectively, which showed the effectiveness in both close-talk and distant utterances. Various score normalization techniques were also explored and observed to be effective for mismatch conditions. The score-level ensemble of two proposed systems could further improve the performance. In our future work, we intend to integrate two proposed systems into a single speaker embedding enhancement system. Experiments using different front-end speaker embedding extraction frameworks will be also conducted.

8. Acknowledgement

This work was supported by the Technology Innovation Program (10076583, Development of free-running speech recognition technologies for embedded robot system) funded by the Ministry of Trade, Industry & Energy(MOTIE, Korea)

9. References

[1] 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.

[2] Soonshin Seo, Daniel Jun Rim, Minkyu Lim, Donghyun Lee, Hosung Park, Junseok Oh, Changmin Kim, and Ji-Hwan Kim, "Shortcut connections based deep speaker embeddings for end-to-end speaker verification system," *Proc. Interspeech 2019*, pp. 2928–2932, 2019.

[3] Youngmoon Jung, Younggwan Kim, Hyungjun Lim, Yeunju Choi, and Hoirin Kim, "Spatial pyramid encoding with convex length normalization for text-independent speaker verification," *Proc. Interspeech 2019*, pp. 4030–4034, 2019.

[4] Jee-weon Jung, Hee-soo Heo, Ju-ho Kim, Hye-jin Shim, and Ha-jin Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," in *Interspeech*, 2019.

[5] 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)*. IEEE, 2014, pp. 4052–4056.

[6] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[7] Mahesh Kumar Nandwana, Julien van Hout, Mitchell McLaren, Allen R Stauffer, Colleen Richey, Aaron Lawson, and Martin Graciarena, "Robust speaker recognition from distant speech under real reverberant environments using speaker embeddings.," in *Interspeech*, 2018, pp. 1106–1110.

[8] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee, "Bridgenets: Student-teacher transfer learning based on recursive neural networks and its application to distant speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5719–5723.

[9] Colleen Richey, Maria A Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, et al., "Voices obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

[10] Mahesh Kumar Nandwana, Julien Van Hout, Mitchell McLaren, Colleen Richey, Aaron Lawson, and Maria Alejandra Barrios, "The voices from a distance challenge

2019 evaluation plan," *arXiv preprint arXiv:1902.10828*, 2019.

[11] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[12] Yanmin Qian, Tian Tan, and Dong Yu, "An investigation into using parallel data for far-field speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5725–5729.

[13] Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio, "Batch-normalized joint training for dnn-based distant speech recognition," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 28–34.

[14] Xiaoyi Qin, Danwei Cai, and Ming Li, "Far-field end-to-end text-dependent speaker verification based on mixed training data with transfer learning and enrollment data augmentation," *Proc. Interspeech 2019*, pp. 4045–4049, 2019.

[15] Jee-weon Jung, Hee-Soo Heo, Hye-jin Shim, and Ha-Jin Yu, "Multi channel far field speaker verification using teacher student deep neural networks," *The Journal of the Acoustical Society of Korea*, vol. 37, no. 6, pp. 483–488, 2018.

[16] Bengt J Borgström and Alan McCree, "The linear prediction inverse modulation transfer function (lp-imtf) filter for spectral enhancement, with applications to speaker recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4065–4068.

[17] Ladislav Mošner, Pavel Matějka, Ondřej Novotnỳ, and Jan Honza Černockỳ, "Dereverberation and beamforming in far-field speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5254–5258.

[18] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[19] Ning Wang, PC Ching, Nengheng Zheng, and Tan Lee, "Robust speaker recognition using denoised vocal source and vocal tract features," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 196–205, 2010.

[20] Kun Han, Yuxuan Wang, DeLiang Wang, William S Woods, Ivo Merks, and Tao Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[21] Seyed Omid Sadjadi and John HL Hansen, "Blind spectral weighting for robust speaker identification under reverberation mismatch," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 5, pp. 937–945, 2014.

[22] Chanwoo Kim and Richard M Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 24, no. 7, pp. 1315–1329, 2016.

[23] Tiago H Falk and Wai-Yip Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 90–100, 2009.

[24] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[25] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[26] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[27] Warick M Brown, Tamás D Gedeon, and David I Groves, "Use of noise to augment training data: a neural network method of mineral–potential mapping in regions of limited known deposit examples," *Natural Resources Research*, vol. 12, no. 2, pp. 141–152, 2003.

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[29] Weicheng Cai, Zexin Cai, Xiang Zhang, Xiaoqi Wang, and Ming Li, "A novel learnable dictionary encoding layer for end-to-end language identification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5189–5193.

[30] 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.

[31] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[32] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[33] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson, "The speakers in the wild (sitw) speaker recognition database.," in *Interspeech*, 2016, pp. 818–822.

[34] Hannah Muckenhirn, Vinayak Abrol, Mathew Magimai-Doss, and Sébastien Marcel, "Understanding and visualizing raw waveform-based cnns," *Proc. Interspeech 2019*, pp. 2345–2349, 2019.

[35] Jee-Weon Jung, Hee-Soo Heo, IL-Ho Yang, Hye-Jin Shim, and Ha-Jin Yu, "Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification," *extraction*, vol. 8, no. 12, pp. 23–24, 2018.

[36] Jiacen Zhang, Nakamasa Inoue, and Koichi Shinoda, "I-vector transformation using conditional generative adversarial networks for short utterance speaker verification," *arXiv preprint arXiv:1804.00290*, 2018.

[37] Jee-weon Jung, Hee-soo Heo, Hye-jin Shim, and Ha-jin Yu, "Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings," *arXiv preprint arXiv:1810.10884*, 2018.

[38] Hung-Shin Lee, Yu-Ding Lu, Chin-Cheng Hsu, Yu Tsao, Hsin-Min Wang, and Shyh-Kang Jeng, "Discriminative autoencoders for speaker verification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5375–5379.

[39] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[42] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, 2013, vol. 30, p. 3.

[43] Hee-Soo Heo, Jee-weon Jung, IL-Ho Yang, Sung-Hyun Yoon, Hye-jin Shim, and Ha-Jin Yu, "End-to-end losses based on speaker basis vectors and all-speaker hard negative mining for speaker verification," *arXiv preprint arXiv:1902.02455*, 2019.

[44] Sergey Novoselov, Aleksei Gusev, Artem Ivanov, Timur Pekhovsky, Andrey Shulipa, Galina Lavrentyeva, Vladimir Volokhov, and Alexandr Kozlov, "Stc speaker recognition systems for the voices from a distance challenge," *Interspeech*, 2019.

[45] Lukáš Burget, Ondrej Novotnỳ, and Ondrej Glembek, "Analysis of but submission in far-field scenarios of voices 2019 challenge," in *Proc. Interspeech*, 2019.

[46] David Snyder, Jesús Villalba, Nanxin Chen, Daniel Povey, Gregory Sell, Najim Dehak, and Sanjeev Khudanpur, "The jhu speaker recognition system for the voices 2019 challenge," in *Proc. Interspeech*, 2019, pp. 2468–2472.

[47] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[48] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[49] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, 2000.

[50] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey*, 2010, p. 14.