Interspeech 2022
18-22 September 2022, Incheon, Korea

Prototypical speaker-interference loss for target voice separation using non-parallel audio samples

Seongkyu Mun, Dhananjaya Gowda, Jihwan Lee,
Changwoo Han, Dokyun Lee, Chanwoo Kim

Samsung Research, Seoul, South Korea
sk1213.mun@samsung.com

Abstract

In this paper, we propose a new prototypical loss function for training neural network models for target voice separation. Conventional methods use paired parallel audio samples of the target speaker with and without an interfering speaker or noise, and minimize the spectrographic mean squared error (MSE) between the clean and enhanced target speaker audio. Motivated by the use of contrastive loss in speaker recognition task, we had earlier proposed a speaker representation loss that uses representative samples from the target speaker in addition to the conventional MSE loss. In this work, we propose a prototypical speaker-interference (PSI) loss, that makes use of representative samples from the target speaker, interfering speaker as well as the interfering noise to better utilize any non-parallel data that may be available. The performance of the proposed loss function is evaluated using VoiceFilter, a popular framework for target voice separation. Experimental results show that the proposed PSI loss significantly improves the PESQ scores of the enhanced target speaker audio.

Index Terms: Non-parallel data, source separation, speaker recognition, prototypical loss, speaker representation

1. Introduction

Deep learning based audio signal processing is increasingly becoming popular for solving the cocktail party problem [1, 2, 3, 4, 5]. A noisy signal, which is mixed by a clean and noise signal, is used as a pair with the clean signal for training a speech enhancement or source separation model [6, 7]. In the training phase, the enhancement model is trained to minimize a loss function that captures the dissimilarity between the clean and enhanced audio pair, such as mean squared error (MSE) between the spectrographic representations of the audio pairs. During the inference phase, the enhancement model receives a noisy signal as input and generates an estimated clean signal as output.

In certain scenarios, such as processing speech on one's personal device, one may have additional non-parallel audio samples of the target speaker whose speech needs to be separated from an interfering speaker or noise. This task is referred to as target voice separation, and several solutions have been proposed for this problem [1, 4, 5, 8]. Pre-enrolled speaker representations from pretrained models are popularly used to derive target speaker embeddings in recent works such as deep extractor [5], SpeakerBeam [4], and VoiceFilter [3]. In [8], a speech encoder-decoder approach with an intermediary speaker mask extraction module that utilizes embeddings from a reference audio of the target speaker to compute the mask. In [9], the speaker separation and speaker encoder modules are jointly trained to improve the utility of both parallel as well as non-parallel data from the target speaker.

A speaker representation loss (SRL) that computes the distance between speaker embeddings derived from the enhanced and a clean target speaker sample is proposed in [1]. This was motivated by the use of a VoiceID loss [10] used to train a speech enhancement network to improve the speaker verification performance. VoiceID loss uses final output of speaker identification network and ground-truth of speaker-ID label to measure whether speaker information is distorted by the enhancement module. In terms of the representation loss, adversarial training method [11] was used for training the speech enhancement module. The discriminator generates loss flows for the enhancement module based on decision classifying the enhanced output as real or fake (enhanced) clean. These loss functions are used primarily to utilize target audio information, in addition to the MSE loss between the clean-enhanced audio pairs.

To extend the representation loss to utilize non-parallel and multiple samples, we propose a new protypical speaker-inteference (PSI) loss function for training the enhancement system. The proposed PSI loss function utilizes non-parallel audio samples from the target speaker, interfering speaker as well as from interfering noise during the training. The non-parallel audio samples are not directly matched to clean signal in spectrogram, but has some information related to clean target, interfering speaker or noise signal. This can provide additional information to prevent over-fitting caused by bias introduced by clean-noisy audio pairs. We apply the proposed method to the voice separation system for target speaker [1, 2, 3], which has been widely researched recently for commercial products. Although it can be generally applied to various audio signal processing tasks, we apply the proposed method to the target voice separation system in this work, since the proposed audio representation for non-parallel samples has some shared points with speaker-conditioning in target voice separation [1, 2]. More details will be covered in Section 2.1.

2. Target Voice Separation

A simple block schematic of a target voice separation system is shown in the upper part of Fig. 1. Unlike blind source separation, reference information of the target speaker is given for separating the target voice among multiple speakers. Various types of speaker information, such as speaker audio embeddings [1, 3], spatial features [12], video information [13], or facial image [2] can be used as reference features of the target speaker. However, considering a compatibility with other audio applications [14], we used only the audio based speaker embedding as reference features in this work. The speaker embedding is extracted from a pretrained speaker encoder, which is originally trained for the task of speaker recognition.

10.21437/Interspeech.2022-11236

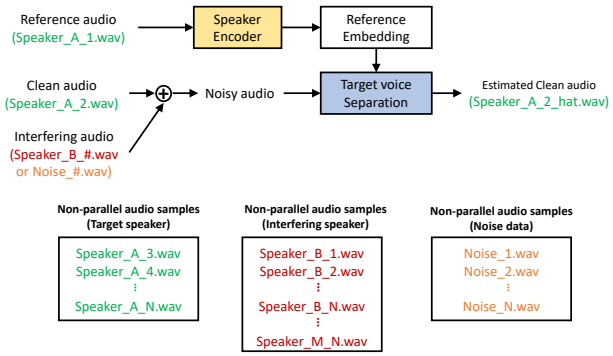

Figure 1: *The concept of the target voice separation and unpaired non-parallel audio data in training*

2.1. Speaker Representation Loss [1]

The SRL was proposed for measuring similarity of neural network perception between clean-noisy audio pairs, instead of measuring distance between raw spectrograms. This loss uses high-level speaker perceptual information to reconstruct clean spectrogram by considering the target speaker characteristic. With various loss combinations, the SRL showed better performance compared to the spectrogram distance loss. We extended this approach to utilize unpaired non-parallel audio data for providing more general information for each pair during training.

As shown in the bottom part of Fig. 1, different utterances of the same target speaker (green) can be defined as non-parallel data. Since these data are not directly matched with the input spectrogram, the training loss cannot be derived in the same way as the MSE. In order to utilize non-parallel audio information without spectrogram mixing, we extended our previous work on SRL to incorporate information from multiple samples of target, interfering speaker and noise.

2.2. Proposed approach

The SRL was used only for the clean-noisy audio pairs. However, in this work we extend SRL to be more general and applicable to unpaired non-parallel sets as well. We propose a new prototypical speaker-interference (PSI) loss, based on prototypical loss (PL) [15] used in the context of speaker verification, to utilize multiple non-parallel data groups more efficiently. It can be expected that general speaker information from multiple embeddings is more effective to derive precise SRL compared to the single sample in [1]. A similar training approach to utilizing the non-parallel set has been used for image-to-image translation [16] and voice conversion [17]. However, to the best of our knowledge, this is the first approach proposed in the field of neural network based spectrographic audio enhancement or target speaker separation.

In the case of Fig. 1, a group of target speaker A utterances (green), which are not selected for the clean sample, can be used for finding more general target embedding in the training phase. Additionally, interference leakage in the estimated target output can be prevented by using general information from the non-parallel interference set.

In the proposed PSI loss, the speaker corresponds to the target speaker and interference refers to both interfering speaker and as well as noise. As shown in Fig. 2, it can also utilize other noise information different from the clean-noisy audio pair set by using embeddings from additional noise data. It has the advantage of being able to consider various noise information for each training pair without direct spectrogram mixing.

Table 1: *Comparison of conventional VoiceFilter [3], with SRL [1] and proposed PSI loss.*

	VF	+SRL	+PSI loss
Use speaker perceptual information for loss function ?	No	Yes	Yes
Use an unpaired non-parallel sample for training? (except reference)	No	No	Yes
Use multiple audio samples for training each audio-pair ?	No	No	Yes

2.3. VoiceFilter [3]

The main feature of the proposed method is to extract speaker embedding from the audio signal and use the distance information in the embedding space. For embedding extraction, a speaker encoder is required, but it takes additional cost to train the encoder with various speaker data. For this reason, although the proposed training method can be applied to various audio enhancement applications, we adopted the VoiceFilter (VF) [3] as it already has a speaker encoder for guiding target speaker as shown in Figure 1. Therefore, speaker embeddings for non-parallel data can be obtained by the encoder already in use without additional training cost. The VF system aims to enhance a noisy magnitude spectrogram by predicting a soft mask that resembles only the target speaker voice. Having both target speaker embedding from speaker encoder and noisy spectrogram as inputs, the original VF trains a speech enhancement network to minimize the distance between the enhanced and clean magnitude spectrogram of the target speaker. In order to focus more on the novel training loss for the non-parallel set, we followed the same model configuration as that of the original VF [3]. For speaker encoder, we used angular prototypical loss based model [14] with adversarial data augmentation [18], which shows one of the best performance (1.16 % equal error rate) on VoxCeleb1 [19] test set. Table 1 shows a brief comparison between the original VF, VF with SRL and VF with the proposed PSI loss.

3. Prototypical speaker-interference loss

Let $f(\mathbf{x};\mathbf{w}) \in \mathbf{R}^D$ be the speaker encoder that maps the input $\mathbf{x}$ to an embedding space of D dimensions. Here, $\mathbf{w}$ typically denotes the parameters of a neural network model trained for speaker characterization or classification. With the sample index i, the spectrograms computed from clean and estimated audio pairs of the target speaker are denoted as $\mathbf{x}_i^T$ and $\mathbf{x}_i^E$.

The speaker representation loss between pairs of clean and estimated target speaker audio $P_i^T = (\mathbf{x}_i^T, \mathbf{x}_i^E)$, where the index i denotes one such pair, is given by

$$L_{SR} = \sum_i \{\, \beta * D_{SR}(P_i^T;\mathbf{w}) + L_{mse}(\mathbf{x}_i^T, \mathbf{x}_i^E) \,\}, \quad (1)$$

where L_{mse} denotes the MSE loss, and D_{SR} denotes the speaker distance. β denotes a constant weight (*e.g.* $\beta = 0.2$), and if it is set to 0, the loss function would be the same as the original VF. Compared to [3], we used MSE loss for spectrogram reconstruction instead of power-law compressed reconstruction error for simplicity. The speaker distance is computed as

$$D_{SR}(P_i^T;\mathbf{w}) = d(\, N(f(\mathbf{x}_i^T)),\ N(f(\mathbf{x}_i^E))\,), \quad (2)$$

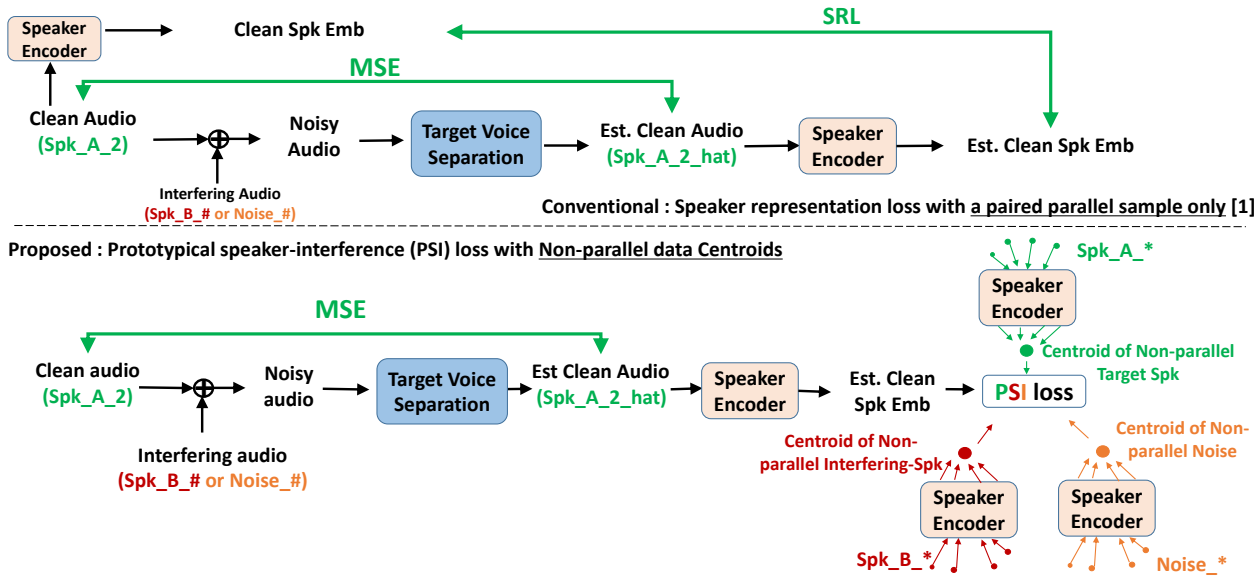

Figure 2: *Comparison of the conventional SRL approach and the proposed system framework with PSI loss. All speaker encoders in the figure indicate the same single pre-trianed model.*

where $d(\mathbf{x}, \mathbf{y})$ can denote any speaker dissimilarity metric between two audio samples. And $N(\mathbf{x})$ denotes L2 normalization function for embedding vector. In this work, we use a simple Euclidean distance between the normalized speaker embeddings derived from the clean and estimated audio samples.

To utilize non-parallel target speaker utterances, we propose to replace $\mathbf{x}_i^T$ in the representation distance (D_{SR}) with a centroid of the non-parallel target speaker embeddings. The centroid of M number of non-parallel utterances $\{\mathbf{x}_{j,m}\}$ for a target speaker with index j is computed as

$$\mathbf{c}_j^T = \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{x}_{j,m}). \tag{3}$$

We chose M as 10 heuristically, and randomly select non-parallel samples for every training clean-noisy audio pair. There was no significant difference in performance when including the parallel target sample ($\mathbf{x}_i^T$) as compared to omitting it from computing the centroids. Then the modified speaker representation distance with the target speaker centroid ($\mathbf{c^T}$) is denoted as

$$D_{\mathbf{c}}^T(P_i^T; \mathbf{w}) = d(\ N(\mathbf{c}_j^T),\ N(f(\mathbf{x}_i^E))\ ) \tag{4}$$

The loss function utilizing non-parallel target utterances can be defined by replacing D_{SR} in Eq. 1 with D_c^T. This scenario would match the depiction in the Fig. 2 without the interfering speaker (orange) and noise (red) parts.

In addition to the target speaker, information on the interfering speaker can be also used in the training phase. A centroid of non-parallel interfering speaker utterances ($\mathbf{c^I}$) can be obtained by Eq.2 for any chosen interfering speaker index. To measure distance with the interfering speaker centroid, it is possible to use the inverse masked (residual) spectrogram and triplet loss as in [1]. However, in this work we use the prototypical loss based method. A centroid of non-parallel interfering noise utterances ($\mathbf{c^N}$) can be obtained in a similar manner.

The prototypical loss [15] is for computing distances to prototype representations (or centroid) of each class, with a training procedure that mimics the test scenario. The use of multiple negative classes, interfering speaker and other noise in this work, helps to stabilise learning since loss functions can force the target speaker embedding to be far from all negatives in a batch, rather than one particular negative as in the case of triplet loss [1]. As proposed in the original paper [15], squared Euclidean distance is used as the distance metric between the estimated target embedding and centroids. For the target speaker case it can be denoted as

$$\mathbf{S}^T = \|N(f(x_i^E)) - \mathbf{c}_j^T\|_2^2 \tag{5}$$

During training, each estimated target embedding is classified against 2-classes (*i.e.* target and interfering speaker) based on a softmax over distances to each centroid:

$$D_{\text{PSI}} = \log \frac{e^{\mathbf{S}^T}}{e^{\mathbf{S}^T} + e^{\mathbf{S}^I}} \tag{6}$$

Here, $\mathbf{S}^T$ is the squared Euclidean distance between the estimated target embedding and the centroid of the non-parallel target speaker utterances. And $\mathbf{S}^I$ is the squared Euclidean distance between the estimated target embedding and the centroid of the non-parallel interfering speaker utterances. The loss function utilizing the prototypical loss is defined by replacing D_{SR} in Eq.(1) with D_{PSI}. It can be shown in the Fig. 2 right side without the noise (orange) parts.

Additionally, we also propose to use a noise centorid for non-parallel noise data. In the SRL work [1], we trained model by mixing only the interfering speaker utterance without additional noise, following the typical scenario of source separation [1, 3, 6]. However, since there is an additional noise in the real-world application, it is necessary to consider the noise in the training phase. In this paper, we propose to use the noise set as also a "speaker". It can be expected to train more noise robust model by considering various noise information from the non-parallel noise centroid. Our final proposed system is shown in Fig. 2, and for the noise including case, Eq. (6) is derived with an additional noise centroid (*i.e.* 3-classes with target, interfering speaker and noise). M in Eq. (3) is heuristically chosen as 30 for the noise centroid.

Table 2: ***PESQ and SDR improvements*** *of different methods over noisy inputs (SDR : 4.73, PESQ : 1.820) containing only interfering speaker. (TS: a Target sample, TC: a Target centroid, TIS: Target and interference samples, TIC: Target and interference centriods)*

	SDR (Δ)	PESQ (Δ)
Noisy input (Before processing)	4.73	1.820
Baseline VF [3]	+5.78	+0.641
+ SRL w/ parallel TS [1]	+5.95	+0.679
+ SRL w/ non-parallel TC	+5.98	+0.684
+ PSI loss w/ parallel TIS	+6.05	+0.703
+ PSI loss w/ non-parallel TIC	**+6.21**	**+0.705**

Table 3: ***PESQ and SDR improvements*** *of different methods over noisy inputs (SDR: 5.01, PESQ: 1.957) with both interfering speaker and noise samples. (TS: a Target sample, TC: a Target centroid, TIS: Target and interference samples, TIC: Target and interference centriods, TINS: Target, interference and noise samples, TINC: Target, interference, and noise centriods)*

	SDR (Δ)	PESQ (Δ)
Noisy input (Before processing)	5.01	1.957
Baseline VF [3]	+6.12	+0.592
+ SRL w/ parallel TS [1]	+6.31	+0.623
+ SRL w/ non-parallel TC	+6.30	+0.623
+ PSI loss w/ parallel TIS	+6.62	+0.664
+ PSI loss w/ non-parallel TIC	+6.70	+0.675
+ PSI loss w/ parallel TINS	+6.77	+0.663
+ PSI loss w/ non-parallel TINC	**+6.82**	**+0.678**

4. Experiments

4.1. Data Generation and Settings

As indicated in [1, 3], we train and evaluate our network with VCTK dataset [20] to achieve a reliable comparison between the conventional VF and our proposed method. Out of a total 109 speakers, we randomly select 99 and 10 speakers each for training and validation, followed by the data generation workflow also referred in [3]. Hyper parameter settings are the same as the conventional VF, except for the speaker embedding dimension being set to 256 based on [14].

In order to evaluate the noise robust performance, audio events from VGG-sound dataset [21] were used for noise samples. The database was chosen due to its large scale of 310 different real-word event classes. The noise samples were randomly selected followed by the train/test split from the dataset configuration, regardless of the event class. In addition to the inference mixing method [3], noise was additionally mixed in the range of 0-10 speech-to-noise ratio (SNR) to clean target source for training phase. For evaluation, two type of test sets were generated. One contains only the interfering speaker, and the other contains noise along with the interfering speaker. The interfering speaker utterance and a noise sample were mixed in same power ratio first, and then the mixed sample was added to a target speech sample about SDR-5-level for evaluating robustness on both speech interference and sound event. In addition to the experiments described in the proposed approach, comparison experiments were conducted using a embeddings from parallel speaker/noise samples, instead of using non-parallel sample centroids in Eq. (3).

To extract more discriminative information form noise samples, we conducted experiments using an additional noise embedding extractor [22] instead of a speaker encoder and an additional loss for noise only, but there was no significant difference in performance improvement. Therefore, for the simplicity, we used a speaker encoder for noise embedding extraction.

4.2. Evaluation and Results

To evaluate the performance of different VF models, we use source to distortion ratio (SDR) [23] and perceptual evaluation of speech quality (PESQ) [24] improvements. In both cases, a higher number indicates a better resemblance of the estimated utterance to the clean utterance of the target speaker. After conducting experiments of the individual task for three-times, the mean results are shown in Table 2 for noisy data with interfering speech and Table 3 for speech inference and sound event noise.

The performance improvement was meaningful, especially in the last row of each table when the most diverse information from noise samples was utilized as well. This is mostly because of a practical issue in the VF system wherein the distance between the reference and target clean utterance is not close enough, compared to distance between reference and interference utterance in embedding space. Even when using a speaker encoder, the intra-speaker distances computed on the embeddings derived from individual samples is far less reliable as compared to using a batch of samples. As can be seen from the experimental results, this issue is mitigated by utilizing centroids from multiple non-parallel samples, which can guide the training process to find the right target point in the embedding space.

Note that this improvement was achieved by only extending a loss function under the same training data configuration and VF structure, without any additional changes to network parameter. Compared to previous researches on loss functions, such as different loss function experiments in Table 2 of [25], Table 2 of [26] and Table 1-2 of [27], our steady SDR improvements can be meaningful considering the fact that the PESQ scores are also improved along with SDR.

5. Conclusions

In this paper, we extended the conventional target voice separation system with additional training criteria to utilize non-parallel data for training. Experimental results show that the most improved performance was achieved by utilizing various non-parallel samples from target, interference speaker and noise source. This improvement was achieved without any additional changes to the network parameters, structures, or database used in the popular VoiceFilter approach as our baseline setting. The improvements were obtained only with the application of our proposed PSI loss based distance metric utilizing non-parallel data centroids for target, interference speaker and noise.

Our proposed loss function is not limited to target voice separation framework, but it can also be applied to other speech-related tasks or frameworks. Note that reason for choosing target voice separation as baseline for the proposed loss is the readily available pre-trained speaker encoder for extracting target or interference embeddings. Our proposed loss function can be used directly without any additional modifications to network structure in other speech enhancement tasks or even voice conversion tasks [28] that use a speaker encoder in their training phase.

6. References

[1] Seongkyu Mun, Soyeon Choe, Jaesung Huh, and Joon Son Chung, "The sound of my voice: speaker representation loss for target voice separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7289–7293.

[2] Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang, "Facefilter: Audio-visual speech separation using still images," *arXiv preprint arXiv:2005.07074*, 2020.

[3] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif Saurous, Ron Weiss, Ye Jia, and Ignacio Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *INTERSPEECH*, 2019, pp. 2728–2732.

[4] Delcroix Marc et al., "Single channel target speaker extraction and recognition with speaker beam," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 5554–5558.

[5] Jun Wang et al., "Deep extractor network for target speaker recovery from single channel speech mixtures," *Proc. Interspeech 2018*, pp. 307–311, 2018.

[6] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *arXiv preprint arXiv:2008.09586*, 2020.

[7] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, *Audio source separation and speech enhancement*, John Wiley & Sons, 2018.

[8] Yunzhe Hao, Jiaming Xu, Jing Shi, P. Zhang, Lei Qin, and B. Xu, "A unified framework for low-latency speaker extraction in cocktail party environments," in *INTERSPEECH*, 2020.

[9] X. Ji, M. Yu, Chunlei Zhang, Dan Su, Tao Yu, Xiaoyu Liu, and Dong Yu, "Speaker-aware target speaker enhancement by jointly learning with speaker embedding extraction," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7294–7298, 2020.

[10] Suwon Shon, Hao Tang, and James Glass, "VoiceID Loss: Speech enhancement for speaker verification," in *INTERSPEECH*, 2019.

[11] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[12] K. Žmolíková et al., "Learning speaker representation for neural network based multichannel speaker extraction," in *2017 IEEE ASRU*. IEEE, 2017, pp. 8–15.

[13] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[14] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.

[15] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *NIPS*, 2017, pp. 4077–4087.

[16] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[17] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.

[18] Jaesung Huh, Hee Soo Heo, Jingu Kang, Shinji Watanabe, and Joon Son Chung, "Augmentation adversarial training for unsupervised speaker recognition," *arXiv preprint arXiv:2007.12085*, 2020.

[19] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[20] Christophe Veaux, Junichi Yamagishi, and Kirsten Macdonald, "Superseded - cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[21] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[22] Qiuqiang Kong, Changsong Yu, Yong Xu, Turab Iqbal, Wenwu Wang, and Mark D Plumbley, "Weakly labelled audioset tagging with attention neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1791–1802, 2019.

[23] Emmanuel Vincent et al., "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] R W. Antony et al., "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*. IEEE, 2001, vol. 2, pp. 749–752.

[25] Francois G Germain, Qifeng Chen, and Vladlen Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[26] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[27] Yan Zhao, Buye Xu, Ritwik Giri, and Tao Zhang, "Perceptually guided speech enhancement using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5074–5078.

[28] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.