INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Deep Feature CycleGANs: Speaker Identity Preserving Non-parallel Microphone-Telephone Domain Adaptation for Speaker Verification

Saurabh Kataria, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velázquez, Najim Dehak

Center for Language and Speech Processing, Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA

{skatari1, jvillal7, pzelasko, laureano, ndehak3}@jhu.edu

Abstract

With the increase in the availability of speech from varied domains, it is imperative to use such out-of-domain data to improve existing speech systems. Domain adaptation is a prominent pre-processing approach for this. We investigate it to adapt microphone speech to the telephone domain. Specifically, we explore CycleGAN-based unpaired translation of microphone data to improve the x-vector/speaker embedding network for Telephony Speaker Verification. We first demonstrate the efficacy of this on real challenging data and then, to improve further, we modify the CycleGAN formulation to make the adaptation *task-specific*. We modify CycleGAN's identity loss, cycle-consistency loss, and adversarial loss to operate in the *deep feature* space. *Deep features* of a signal are extracted from an auxiliary (speaker embedding) network and, hence, preserves speaker identity. Our 3D convolution-based Deep Feature Discriminators (DFD) show relative improvements of 5-10% in terms of equal error rate. To dive deeper, we study a challenging scenario of pooling (adapted) microphone and telephone data with data augmentations and telephone codecs. Finally, we highlight the sensitivity of CycleGAN hyper-parameters and introduce a parameter called *probability of adaptation*.

Index Terms: CycleGAN, Deep features, Speech domain adaptation, Telephony speech, Preserving speaker identity

1. Introduction

Telephony speaker verification draws significant interest from the speaker recognition community, thanks to the Speaker Recognition Evaluation (SRE) that is conducted regularly by National Institute of Standards and Technology (NIST). The state-of-the-art system uses x-vector speaker embedding front-end [1] and Probabilistic Linear Discriminant Analysis (PLDA) based back-end [2, 3]. In this work, we experiment with evaluation data acquired from the telephone domain. A standard way of constructing the training data for the x-vector network is to pool together telephone and microphone data. However, the microphone data is downsampled (from 16 kHz to 8 kHz) and treated with standard telephone codecs like GSM, AMR-NB to simulate telephone domain [4, 5]. Speech signal from a particular *domain* comprises of characteristic channel and transmission effects. We argue that a more principled approach like *domain adaptation* is suited better since the telephone domain consists of other channel effects in addition to different codecs [6].

Unsupervised Domain Adaptation (UDA) is suited for this scenario since access to paired data from multiple domains is usually infeasible. CycleGAN (Cycle-Consistent Generative Adversarial Network) [7] (Sec. 2) is a popular GAN-based unpaired translation/style transfer technique. Adaptation of speech features using CycleGAN works well for problems like voice conversion [8], speech enhancement [9, 10], and domain adaptation [11, 12]. In [11], authors pursued microphone speaker verification by adapting microphone evaluation data to telephone domain to overcome the *domain shift* w.r.t. their telephone domain training data. In the follow-up study [12], the authors focused on the scenario where the availability of the target domain data is restricted. Various modifications to CycleGAN are proposed in the literature, like having multiple discriminators [13] and preserving class labels in the forward and backward cycle (semantic consistency) [14]. It is known that the cycle-consistency constraint is too restrictive and limits the flexibility in predictions [15, 16, 17]. Moreover, in GANs, the loss in semantic information is compensated by matching high-level *deep features* of signals via *content loss* [18, 19].

Deep features of a signal refer to its activations obtained from a pre-trained auxiliary network [20, 21, 22]. Choosing it as a speaker embedding network in the *content loss* minimizes the distance between the speaker embeddings of signals. It also makes the formulation *task-specific* [23, 24], since our downstream task is speaker verification. This idea of preserving certain attributes is explored in non-CycleGAN solutions for linguistic [25] and speaker identity preservation [26]. The study [27] defined the cycle-consistency loss in the *deep feature space* for preserving phonetic information for voice conversion task. Our goal is to devise a *complete* re-formulation of CycleGAN where *all* loss terms and constraints are computed in the *deep feature space*. This is similar to the *task-specific* speech enhancement of [23]. To improve the generators and the discriminators of CycleGAN, we modify its identity, cycle-consistency, and adversarial loss. The focus is on improving the utility of microphone training data (via adaptation) for telephony speaker verification. Previous works have only explored adapting test sets with no speaker identity preservation [9, 10]. Note that, we refer to *microphone speech* as recordings that are originally wide-band, in consistency with NIST terminology.

Our contributions are: 1) using CycleGANs, we provide the first study for combining telephone and (adapted) microphone training data for Telephony Speaker Verification; 2) we re-formulate all components of CycleGAN to operate in deep feature space, thereby alleviating cycle-consistency strictness issue and making the formulation *task-specific* w.r.t. the downstream task; 3) introduce Deep Feature Discriminator (DFD) which uses deep features for real/fake determination in adversarial loss; 4) demonstrate the efficacy of proposal on three real test sets while quantifying the effect of telephone codecs, data augmentation, *probability of adaptation*, and hyper-parameters.

2. Description of CycleGAN

CycleGAN [7] is an unsupervised generative model which was proposed in computer vision for unpaired translation (or domain

http://dx.doi.org/10.21437/Interspeech.2021-1502

adaptation). It is based on Generative Adversarial Networks (GANs) [28] which are briefly summarized as a minimax game between generator ($\mathcal{G}$) and discriminator ($\mathcal{D}$) network where discriminator tries to distinguish between real and fake (synthesized by $\mathcal{G}$) sample, while generator tries to fool the discriminator into believing that the synthesized sample is real. The goal of GAN is to learn to generate real samples via $\mathcal{G}$. CycleGAN consists of two GANs each trying to learn *realistic* mapping (enforced by adversarial loss) from one domain to the other. The training of two GANs is linked via a cycle-consistency constraint which promotes the mappings to be reversible which, desirably, restricts the output space of generators. If the two domains microphone and telephone are denoted by $\mathcal{M}$ and $\mathcal{T}$ respectively, mathematically, we optimize:

$$\min_{\mathcal{G}_{\mathcal{M}\to\mathcal{T}},\mathcal{G}_{\mathcal{T}\to\mathcal{M}}} \max_{\mathcal{D}_{\mathcal{M}},\mathcal{D}_{\mathcal{T}}} \mathcal{L}_{\text{cyc-GAN}}, \tag{1}$$

where $\mathcal{L}_{\text{cyc-GAN}}$ is given by a weighted sum of adversarial, cycle-consistency and identity losses,

$$\begin{aligned}&\mathcal{L}_{\text{cyc-GAN}}(\mathcal{G}_{\mathcal{M}\to\mathcal{T}},\mathcal{G}_{\mathcal{T}\to\mathcal{M}},\mathcal{D}_{\mathcal{M}},\mathcal{D}_{\mathcal{T}}) = \\ &\mathcal{L}_{\text{GAN},\mathcal{M}}(\mathcal{G}_{\mathcal{T}\to\mathcal{M}},\mathcal{D}_{\mathcal{M}}) + \mathcal{L}_{\text{GAN},\mathcal{T}}(\mathcal{G}_{\mathcal{M}\to\mathcal{T}},\mathcal{D}_{\mathcal{T}}) \\ &+ \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}(\mathcal{G}_{\mathcal{M}\to\mathcal{T}},\mathcal{G}_{\mathcal{T}\to\mathcal{M}}) + \lambda_{\text{id}}\mathcal{L}_{\text{id}}(\mathcal{G}_{\mathcal{M}\to\mathcal{T}},\mathcal{G}_{\mathcal{T}\to\mathcal{M}})\end{aligned} \tag{2}$$

To define the terms in (2), we introduce functions $(h^{\mathcal{M}}_{\text{cyc}}, h^{\mathcal{T}}_{\text{cyc}}, h^{\mathcal{M}}_{\text{id}}, h^{\mathcal{T}}_{\text{id}}, h^{\mathcal{M}}_{\text{disc}}, h^{\mathcal{T}}_{\text{disc}})$, which, for the vanilla Cycle-GAN in this section, are equal to identity ($\mathrm{x} = h^b_a(\mathrm{x})$). Their utility will be revealed in the next section. λ_{cyc} and λ_{id} are the weights for the cycle-consistency and identity loss constraints respectively. Identity loss constraint is a commonly used regularizer in CycleGANs generator to 1) be capable of identity functionality and 2) be conservative when presented unexpected (surprise) input. Let $\mathbf{m}$ and $\mathbf{t}$ denote the 2D Time-Frequency (TF) features from microphone and telephone domain respectively. Also, let $p_{\mathcal{M}}$ and $p_{\mathcal{T}}$ denote the real microphone and telephone training database respectively. The cycle-consistency loss constraint is

$$\begin{aligned}&\mathcal{L}_{\text{cyc}}(\mathcal{G}_{\mathcal{M}\to\mathcal{T}},\mathcal{G}_{\mathcal{T}\to\mathcal{M}},h^{\mathcal{M}}_{\text{cyc}},h^{\mathcal{T}}_{\text{cyc}}) = \\ &\mathbb{E}_{\mathbf{m}\sim p_{\mathcal{M}}}[||h^{\mathcal{M}}_{\text{cyc}}(\mathbf{m}) - \mathcal{G}_{\mathcal{T}\to\mathcal{M}}(\mathcal{G}_{\mathcal{M}\to\mathcal{T}}(h^{\mathcal{M}}_{\text{cyc}}(\mathbf{m})))||_1] \\ &+ \mathbb{E}_{\mathbf{t}\sim p_{\mathcal{T}}}[||h^{\mathcal{T}}_{\text{cyc}}(\mathbf{t}) - \mathcal{G}_{\mathcal{M}\to\mathcal{T}}(\mathcal{G}_{\mathcal{T}\to\mathcal{M}}(h^{\mathcal{T}}_{\text{cyc}}(\mathbf{t})))||_1],\end{aligned} \tag{3}$$

the identity loss constraint is

$$\begin{aligned}&\mathcal{L}_{\text{id}}(\mathcal{G}_{\mathcal{M}\to\mathcal{T}},\mathcal{G}_{\mathcal{T}\to\mathcal{M}},h^{\mathcal{M}}_{\text{id}},h^{\mathcal{T}}_{\text{id}}) = \\ &\mathbb{E}_{\mathbf{m}\sim p_{\mathcal{M}}}[||h^{\mathcal{M}}_{\text{id}}(\mathbf{m}) - \mathcal{G}_{\mathcal{T}\to\mathcal{M}}(h^{\mathcal{M}}_{\text{id}}(\mathbf{m}))||_1] \\ &+ \mathbb{E}_{\mathbf{t}\sim p_{\mathcal{T}}}[||h^{\mathcal{T}}_{\text{id}}(\mathbf{t}) - \mathcal{G}_{\mathcal{M}\to\mathcal{T}}(h^{\mathcal{T}}_{\text{id}}(\mathbf{t}))||_1],\end{aligned} \tag{4}$$

and the Least Squares GAN (LSGAN) [29] based (two) adversarial loss terms are

$$\begin{aligned}&\mathcal{L}_{\text{GAN},\mathcal{M}}(\mathcal{G}_{\mathcal{T}\to\mathcal{M}},\mathcal{D}_{\mathcal{M}},h^{\mathcal{M}}_{\text{disc}},h^{\mathcal{T}}_{\text{disc}}) = \\ &\mathbb{E}_{\mathbf{m}\sim p_{\mathcal{M}}}[(1-\mathcal{D}_{\mathcal{M}}(h^{\mathcal{M}}_{\text{disc}}(\mathbf{m})))^2] \\ &+ \mathbb{E}_{\mathbf{t}\sim p_{\mathcal{T}}}[(\mathcal{D}_{\mathcal{M}}(\mathcal{G}_{\mathcal{T}\to\mathcal{M}}(h^{\mathcal{T}}_{\text{disc}}(\mathbf{t}))))^2], \\ &\mathcal{L}_{\text{GAN},\mathcal{T}}(\mathcal{G}_{\mathcal{M}\to\mathcal{T}},\mathcal{D}_{\mathcal{T}},h^{\mathcal{M}}_{\text{disc}},h^{\mathcal{T}}_{\text{disc}}) = \\ &\mathbb{E}_{\mathbf{t}\sim p_{\mathcal{T}}}[(1-\mathcal{D}_{\mathcal{T}}(h^{\mathcal{T}}_{\text{disc}}(\mathbf{t})))^2] \\ &+ \mathbb{E}_{\mathbf{m}\sim p_{\mathcal{M}}}[(\mathcal{D}_{\mathcal{T}}(\mathcal{G}_{\mathcal{M}\to\mathcal{T}}(h^{\mathcal{M}}_{\text{disc}}(\mathbf{m})))^2].\end{aligned} \tag{5}$$

$\mathcal{G}_{\mathcal{M}\to\mathcal{T}}$ and $\mathcal{G}_{\mathcal{T}\to\mathcal{M}}$ denote the *forward* and *backward* cycle of CycleGAN. We employ LSGAN because it alleviates the vanishing gradient problem of the vanilla GAN [29]. To optimize Eq. 1, we use Alternating Gradient Descent (AGD) algorithm [28] in which discriminators and generators are updated for $n_{\mathcal{D}}$ and $n_{\mathcal{G}}$ steps alternatively.

3. Deep feature CycleGAN

As alluded in Sec. 1, using *deep features* is the core ingredient of our proposal. Given a pre-trained auxiliary network $\mathcal{A}$, with $\mathcal{A}_{[:i]}$ denoting the network till i^{th} layer, $\mathcal{A}_{[:i]}$ returns the hidden activations at layer i, and the corresponding deep features for input $\mathbf{x}$ using the first q layers is given by

$$\begin{aligned}&\texttt{DF}(\mathbf{x},\mathcal{A},q) = \\ &\texttt{Concat}(\texttt{Expand}(\{\mathcal{A}_{[:1]}(\mathbf{x}),\mathcal{A}_{[:2]}(\mathbf{x}),\ldots,\mathcal{A}_{[:q]}(\mathbf{x})\}))\end{aligned} \tag{6}$$

where the operation `Expand` means expanding each dimension of matrices to the corresponding maximum dimension seen in the set. The newly created entries in matrices are filled with zeros. `Concat` operation concatenates the matrices along a newly created dimension. These operations are done to convert a set of different shape matrices to a single rectangular matrix. For our task, the output of $\texttt{DF}(\mathbf{x},\mathcal{A},q)$ is five-dimensional: (batch size, q, channels, height, width), which makes it suitable for processing with Convolutional Neural Network (CNN) based architectures (Sec. 4).

Let $\mathcal{A}_T$ and $\mathcal{A}_M$ denote the auxiliary (speaker classification) networks pre-trained with the original (unadapted) telephone and microphone training data respectively. *DF-CycleGAN* uses the following choice of h functions in Eqs. 3-5:

$$\begin{aligned}h^{\mathcal{M}}_{\text{cyc}} &:= \texttt{DF}(\cdot,\mathcal{A}_M,q_{\text{cyc}}),\ h^{\mathcal{T}}_{\text{cyc}} := \texttt{DF}(\cdot,\mathcal{A}_T,q_{\text{cyc}}) \\ h^{\mathcal{M}}_{\text{id}} &:= \texttt{DF}(\cdot,\mathcal{A}_M,q_{\text{id}}),\ h^{\mathcal{T}}_{\text{id}} := \texttt{DF}(\cdot,\mathcal{A}_T,q_{\text{id}}) \\ h^{\mathcal{M}}_{\text{disc}} &:= \texttt{DF}(\cdot,\mathcal{A}_M,q_{\text{disc}}),\ h^{\mathcal{T}}_{\text{disc}} := \texttt{DF}(\cdot,\mathcal{A}_T,q_{\text{disc}})\end{aligned} \tag{7}$$

q_{cyc}, q_{id}, and q_{disc} are the number of layers used for deep feature extraction for cycle-consistency loss, identity loss, and adversarial loss respectively. When modifying only one out of these three terms, we refer to the corresponding CycleGANs as *DF-Cycle*, *DF-Identity*, and *DF-Disc.* respectively. Note from Eqs. 3-5 that $\texttt{Concat} \circ \texttt{Expand}$ operation requires generators and discriminators to process 5D inputs. For the generators, due to GPU memory limitations, we fuse the last two dimensions of the input tensor and simply use 2D CNNs to process the resultant 4D data. Inspired from [30], for (Deep Feature) Discriminators (DFD), we *retain* the complete 5D data structure and use 3D CNN layers to process it. Note that $\mathcal{A}_T$ and $\mathcal{A}_M$ are frozen during *DF-CycleGAN* training.

4. Experimental Setup

We use *VoxCeleb2* [31] and SRE Conversational Telephone Speech (CTS) 2012-18 English telephone sets (referred to as *SRE12-18*) as the training datasets for the source (microphone) and target (telephone) domains of CycleGAN respectively. *SRE12-18* is chosen to be representative of modern telephone speech. We apply a simple energy-based Voice Activity Detection (VAD) pre-processing to include only voiced frames for adaptation training. Subsequently, *VoxCeleb2* consists of 5994 speakers and 2247 hrs of speech, while *SRE12-18* consists of 2895 speakers and 4917 hrs of speech.

The CycleGAN is learned on 2.5 sec long 64-D Log Mel-Filter Bank (LMFB) features with 25ms Fast Fourier Transform (FFT) window and 10ms shift. The batch size is 8, the number of epochs is 15 (1 epoch = 100 hrs data here), the Learning Rate (LR) for generators is 0.0002, the LR for discriminators is 0.0001, the LR scheduler is a linearly decaying function with a minimum LR of 1e-7, the optimizer is Adam-based

with $(\beta_1, \beta_2) = (0.5, 0.999)$. Data augmentation is not performed during CycleGAN training and *VoxCeleb2* is downsampled a priori to 8 kHz to match the telephone sampling rate. We use the generator and discriminator architectures from the Voice Conversion work of [32]. The generators are 14-layer encoder-decoder style deep CNNs with Instance Normalization (IN), residual blocks, and Gated Linear Units (GLU) activations. The discriminators are simpler 5-layer CNNs. More details on architecture can be found in [32]. $n_{\mathcal{D}}$ and $n_{\mathcal{G}}$ are fixed to 1 and 2 respectively, while λ_{id} and λ_{cyc} are set to 5 and 10 respectively.

For *deep feature* based CycleGANs, $q_{\text{cyc}} = q_{\text{id}} = 5$. A higher value of q_{disc} implies longer training time and memory requirements so we choose a nominal value of $q_{\text{disc}} = 2$. Auxiliary networks $\mathcal{A}_M$ and $\mathcal{A}_T$ are (pre-)trained for speaker classification task with *VoxCeleb2* and *SRE12-18* respectively. The architecture used is Light Resnet-34, which is obtained from the popular Resnet-34 architecture by reducing the number of channels in all CNN layers by half. Refer to [2] for more details. Optimization is Adam-based Stochastic Gradient Descent (SGD), the number of epochs is 90 (1 epoch here means entire dataset), the batch size is 128, the loss function is AM-Softmax [33] (margin=0.3), the sequence length is 4 sec, the LR is 0.05, and the LR scheduler reduces the LR by half when validation loss does not decrease for three epochs. Sliding-window Mean-Variance Normalization (MVN) is done on-the-fly on the input features (64-D LMFB) with left and right context equal to 150 frames each. Final adaptation is performed with $\mathcal{G}_{\mathcal{M}\to\mathcal{T}}$.

The x-vector network for Speaker Verification is also chosen as a Light ResNet-34 with the x-vector dimension as 256. Its training data is different for each experiment (Sec. 5): *VoxCeleb2*, *SRE-Mix-SWBD*, or their combination. *SRE-Mix-SWBD* refers to the combination of SRE CTS datasets from 2004-12, MIXER6, and Switchboard, which is a standard training set used in prior work [2]. For experiments with data augmentation, we corrupt speech utterances with noise files from (downsampled) MUSAN [34] with Signal-to-Noise Ratio (SNR) range of $[3, 18]$ dB and corruption probability of 0.7.

For evaluation, we choose PLDA as the backend for target/non-target decision making. The dimension for Linear Discriminant Analysis (LDA) pre-processing (on x-vectors) is 150, PLDA dimension is 125, and its training data is the combination of train and (60% of) dev portions of CTS SRE 2016-18. We experiment with three challenging language-mismatched real test sets from SRE 2016 and 2019: *SRE-19-eval* (Tunisian Arabic), (rest 40%) *SRE-16-yue* (Cantonese), and (rest 40%) *SRE-16-tgl* (Tagalog). The x-vectors are extracted from a random chunk of 10-60s of speech. Results are reported on Equal Error Rate (EER) (in %) and Minimum Decision Cost Function (minDCF) ($p_{\text{tar}} = 0.05$). The lower the EER and minDCF, the better. All systems are implemented with Pytorch.

5. Results

5.1. Baseline CycleGAN adaptation

In Table 1, we demonstrate the benefit of adapting microphone data to the telephone domain using vanilla CycleGAN. Results are reported in the format: *EER / minDCF*. Improvement is significant in EER as well as minDCF for all three test sets. As a motivational illustration, Fig. 1 shows the working of adaptation using 2-D t-distributed Stochastic Neighbor Embedding (t-SNE) [35] visualization. We can observe that the centroid of microphone embeddings has shifted considerably close to the telephone centroid. We perform various runs of t-SNE to find

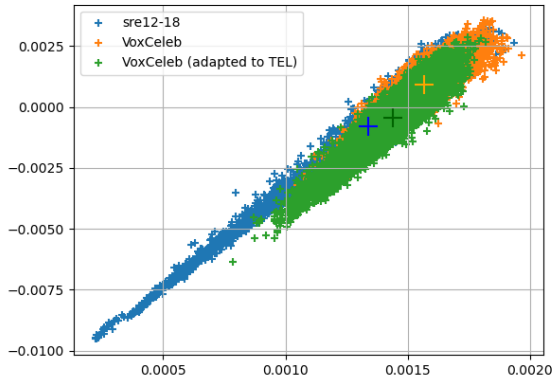

Figure 1: *t-SNE visualization of embeddings for telephone (SRE12-18), microphone (VoxCeleb2), and microphone adapted to telephone. Respective centroids are denoted by big "+".*

Table 1: *Comparison of adaptation with vanilla CycleGAN vs no adaptation when using microphone data (VoxCeleb) for x-vector training and no data augmentation. Bold typeface denotes the best metric value obtained.*

Adaptation ↓	SRE19-eval	SRE16-yue	SRE16-tgl
None	7.33 / 0.357	6.79 / 0.394	14.81 / 0.669
CycleGAN	**7.02 / 0.351**	**6.33 / 0.354**	**13.79 / 0.634**

that this structure is highly robust to the choice of perplexity value (5-50) and metric type (Euclidean/Cosine). We produce the illustration for perplexity value of 30 and Euclidean metric. To justify the significance of the relative shift in centroid locations, we preserve global structure in addition to local via Principal Component Analysis (PCA) initialization in open t-SNE [36]. Additionally, we tried Uniform Manifold Approximation and Projection (UMAP) visualization [37] to arrive at identical observations.

5.2. Deep feature-based CycleGAN adaptation

To improve CycleGANs, we introduced three modifications (Eq. 7) to the original formulation. In Table 2, we study the effect of those modifications independently as well as combined. We make three observations. One, modifying cycle-consistency, identity, and adversarial loss independently im-

Table 2: *Effect of modifying cycle-consistency loss, identity loss, and adversarial loss independently as well as in combinations on top of baseline adaptation system of Sec. 5.1.*

Adaptation ↓	SRE19-eval	SRE16-yue	SRE16-tgl
CycleGAN	7.02 / 0.351	6.33 / 0.354	13.79 / 0.634
DF-Cycle	6.71 / 0.344	6.46 / 0.357	13.40 / 0.614
DF-Identity	6.73 / 0.336	6.39 / 0.344	14.66 / 0.644
DF-Disc. (2D)	7.33 / 0.368	6.49 / 0.370	14.33 / 0.646
DF-Disc. (3D)	6.92 / 0.341	5.87 / **0.340**	**13.18** / 0.622
DF-Identity + DF-Cycle	6.99 / 0.342	6.18 / 0.368	13.66 / 0.622
DF-CycleGAN	**6.70 / 0.332**	**5.81** / 0.351	13.68 / **0.612**

Table 3: *Studying the effect of pooling together telephone and (adapted) microphone data for the training of the x-vector network. p_{adapt} is the probability of adaptation of microphone samples seen during SGD based training of x-vector network. Results are provided in format (EER (in %) / minDCF) for various values of p_{adapt} in the presence or absence of telephone codecs and data augmentation.*

x-vector training data type ↓	p_{adapt}	Tel codecs	λ_{id}	Data Augmentation	SRE19-eval	SRE16-yue	SRE16-tgl
Microphone	0	✓	-	✗	7.33 / **0.357**	6.79 / 0.394	14.81 / **0.669**
Telephone	-	-	-	✗	**6.85** / 0.381	**5.9** / **0.348**	**14.48** / 0.699
Telephone + Microphone	0	✓	-	✗	5.64 / 0.311	4.95 / 0.284	13.09 / 0.624
Telephone + Microphone	1	✗	5	✗	5.52 / 0.304	4.40 / 0.265	12.60 / 0.615
Telephone + Microphone	1	✓	5	✗	5.79 / 0.315	5.14 / 0.291	12.12 / 0.608
Telephone + Microphone	0.2	✓	5	✗	5.40 / 0.294	4.41 / 0.259	12.01 / 0.598
Telephone + Microphone	0.5	✓	5	✗	**5.26** / **0.288**	**4.37** / 0.264	11.83 / 0.587
Telephone + Microphone	0.8	✓	5	✗	5.34 / 0.299	4.79 / **0.257**	**11.79** / **0.584**
Telephone + Microphone	0	✗	-	✓	4.63 / 0.261	4.10 / 0.271	10.41 / 0.530
Telephone + Microphone	0	✓	-	✓	4.65 / 0.261	4.44 / 0.276	10.72 / 0.539
Telephone + Microphone	1	✗	10	✓	4.74 / 0.264	**4.08** / 0.254	9.81 / 0.522
Telephone + Microphone	1	✓	10	✓	4.98 / 0.282	4.80 / 0.273	9.50 / **0.497**
Telephone + Microphone	0.5	✗	10	✓	4.70 / 0.264	4.27 / 0.277	10.17 / 0.536
Telephone + Microphone	0.5	✓	5	✓	4.92 / 0.275	4.37 / 0.262	9.69 / 0.505
Telephone + Microphone	0.5	✓	10	✓	**4.60** / **0.250**	4.22 / **0.252**	**9.34** / 0.499

proves results on all six metrics with certain exceptions. Two, using 3D convolutions in discriminator (*DF-Disc. (3D)*) is vastly superior to 2D convolutions due to the preservation of 5D data structure (Sec. 3). Three, the adaptation model with all three modifications (*DF-CycleGAN*) is the best while the one with only modified identity and cycle-consistency loss (*DF-Identity + DF-Cycle*) suggests the sensitivity to the choice of hyper-parameters like λ_{id}, λ_{cyc}. It is important to note that modification of loss functions changes their dynamic range and, to obtain best results, re-tuning of weights $(\lambda_{\text{cyc}}, \lambda_{\text{id}})$ is required, which is an expensive procedure for CycleGANs. We do not pursue an extensive weight search and hence, the results can be potentially improved.

In *deep feature* based CycleGANs, the computational complexity and memory requirements increases due to the usage of auxiliary networks ($\mathcal{A}_T$, $\mathcal{A}_M$), 3D convolutions, and increased dimensionality of input features to generators and discriminators. As a trade-off between complexity and performance, we look for minimal modifications to CycleGAN and, thus use *DF-Disc. (3D)* for further detailed experiments.

5.3. Pooling telephone and adapted microphone data with data augmentation and telephone codecs

In Sec. 5.1 and Sec. 5.2, we demonstrated the benefit of baseline and proposed CycleGAN adaptation of microphone data for the purpose of x-vector training. It is imperative to study the pooling together of this adapted data with available telephone training data. Moreover, in practice, for best performance, it is standard to apply noise augmentation on all data and treat microphone speech (after downsampling) with telephone codecs like GSM, AMR-NB, etc. These operations are done in the time-domain and LMFB features are extracted subsequently.

We tabulate the corresponding results in Table 3. We first demonstrate that our choice of the telephone (*SRE-Mix-SWBD*) and the microphone data (*VoxCeleb2*) gives a good performance, and therefore, can be combined for potentially further gains. When pooling together telephone and (adapted) microphone data, we first experiment without data augmentation. A parameter p_{adapt} is introduced which refers to the probability of adapting microphone features on the fly when training the x-vector network with SGD. We introduce this parameter due to the observation that, in the case of no augmentation, adaptation does not bring consistent improvements. Comparing results with and without codecs, we observe that the role of codecs varies with test sets also. For instance, with adaptation and codecs, EER on *SRE16-tgl* is improved from 13.09 % to 12.12 % but results are worse on other test sets. Using codecs along with a convenient value of $p_{\text{adapt}} = 0.5$ alleviates these issues and we obtain a decent performance overall.

For the systems with augmentation, we find the role of codecs to be inconsistent again. We can note from the rows with $p_{\text{adapt}} = 1$ that adaptation helps regardless of other factors. Parameter $p_{\text{adapt}} = 0.5$ further boosts the performance possibly due to the increased diversity in training samples. Recall that the CycleGAN is trained with LMFB features without augmentation and codecs. Using such a system for speech with augmentation and/or codecs requires the adaptation mapping to be robust to these operations. A simple trick to achieve this is to increase the value of identity loss as it makes the generators more conservative 2. The significance of the higher value of λ_{id} can be noted from the last two rows. Our overall best system, thus, requires data augmentation, telephone codecs, stochastic adaptation ($p_{\text{adapt}} = 0.5$), and a higher identity loss ($\lambda_{\text{id}} = 10$).

6. Conclusion

We study if reducing the *domain shift* of microphone data w.r.t. telephone domain is beneficial for the x-vector training for Telephony Speaker Verification. Using a generative unpaired translation technique called CycleGAN, we observe relative improvements of 5-10% in EER, minDCF on three challenging language-mismatched test sets. Then, we propose three modifications to the formulation of CycleGAN based on *deep features*. We find this improves cycle-consistency constraint and preserves speaker identity during adaptation. Finally, we study a challenging scenario where we pool adapted microphone data with telephone data and gain insights into the interplay of various hyper-parameters $(p_{\text{adapt}}, \lambda_{\text{id}})$, codecs, and augmentation. Our approach demonstrates great potential for *task-specific* domain adaptation. In the future, we can explore the adaptation of non-English training data, learn many-to-many CycleGAN mappings, and explore the latest telephone codecs [5].

7. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. 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] J. Villalba, D. Garcia-Romero, N. Chen, G. Sell, J. Borgstrom, A. McCree, L. Garcia-Perera, S. Kataria, P. S. Nidadavolu, P. A. Torres-Carrasquillo *et al.*, "Advances in speaker recognition for telephone and audio-visual data: the jhu-mit submission for nist sre19," in *Proceedings of Odyssey*, 2020.

[3] J. Villalba, D. Garcia-Romero, N. Chen, G. Sell, J. Borgstrom, A. McCree, D. Snyder, S. Kataria, L. P. Garcıa-Perera, F. Richardson *et al.*, "The jhu-mit system description for nist sre19 av," in *NIST SRE19 Workshop*, 2019.

[4] R. Bittner, E. Humphrey, and J. Bello, "Pysox: Leveraging the audio signal processing power of sox in python," in *Proceedings of the International Society for Music Information Retrieval Conference Late Breaking and Demo Papers*, 2016.

[5] G. Sivaraman, A. Vidwans, and E. Khoury, "Speech bandwidth expansion for speaker recognition on telephony audio," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 440–445.

[6] F. Köster, *Multidimensional analysis of conversational telephone speech*. Springer, 2018.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. 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.

[8] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[9] P. S. Nidadavolu, S. Kataria, J. Villalba, P. Garcia-Perera, and N. Dehak, "Unsupervised feature enhancement for speaker verification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7599–7603.

[10] P. S. Nidadavolu, S. Kataria, P. García-Perera, J. Villalba, and N. Dehak, "Single channel far field feature enhancement for speaker verification in the wild," *arXiv preprint arXiv:2005.08331*, 2020.

[11] P. S. Nidadavolu, J. Villalba, and N. Dehak, "Cycle-gans for domain adaptation of acoustic features for speaker recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6206–6210.

[12] P. S. Nidadavolu, S. Kataria, J. Villalba, and N. Dehak, "Low-resource domain adaptation for speaker recognition using cycle-gans," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 710–717.

[13] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "A multi-discriminator cyclegan for unsupervised non-parallel speech domain adaptation," *arXiv preprint arXiv:1804.00522*, 2018.

[14] S. Zhao, X. Chen, X. Yue, C. Lin, P. Xu, R. Krishna, J. Yang, G. Ding, A. L. Sangiovanni-Vincentelli, and K. Keutzer, "Emotional semantics-preserved and feature-aligned cyclegan for visual emotion adaptation," *arXiv preprint arXiv:2011.12470*, 2020.

[15] Y. Zhao, R. Wu, and H. Dong, "Unpaired image-to-image translation using adversarial consistency loss," in *European Conference on Computer Vision*. Springer, 2020, pp. 800–815.

[16] J.-H. Park, M. Oh, and H.-M. Park, "Unsupervised speech domain adaptation based on disentangled representation learning for robust speech recognition," *arXiv preprint arXiv:1904.06086*, 2019.

[17] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "Augmented cyclic adversarial learning for low resource domain adaptation," *arXiv preprint arXiv:1807.00374*, 2018.

[18] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9465–9474.

[19] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *arXiv preprint arXiv:1910.06711*, 2019.

[20] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[21] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," *arXiv preprint arXiv:2010.11860*, 2020.

[22] Z. Zhang, R. Zhang, Z. Li, Y. Bengio, and L. Paull, "Perceptual generative autoencoders," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 298–11 306.

[23] S. Kataria, P. S. Nidadavolu, J. Villalba, N. Chen, P. Garcia-Perera, and N. Dehak, "Feature enhancement with deep feature losses for speaker verification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7584–7588.

[24] S. Kataria, P. S. Nidadavolu, J. Villalba, and N. Dehak, "Analysis of deep feature loss based enhancement for speaker verification," *arXiv preprint arXiv:2002.00139*, 2020.

[25] H.-T. Luong and J. Yamagishi, "Nautilus: a versatile voice cloning system," *arXiv preprint arXiv:2005.11004*, 2020.

[26] H. Du, X. Tian, L. Xie, and H. Li, "Optimizing voice conversion network with cycle consistency loss of speaker identity," *arXiv preprint arXiv:2011.08548*, 2020.

[27] J. You, G. Nam, D. Kim, and G. Chae, "Axial residual networks for cyclegan-based voice conversion," 2021.

[28] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[29] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[30] D. Cai, X. Qin, and M. Li, "Multi-channel training for end-to-end speaker recognition under reverberant and noisy environment." in *INTERSPEECH*, 2019, pp. 4365–4369.

[31] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[32] T. Kaneko, H. Kameoka, K. Tanaka, and N. 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.

[33] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[34] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[35] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[36] P. G. Poličar, M. Stražar, and B. Zupan, "opentsne: a modular python library for t-sne dimensionality reduction and embedding," *bioRxiv*, 2019. [Online]. Available: https://www.biorxiv.org/content/early/2019/08/13/731877

[37] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.