**ABSTRACT**

Entanglement of speaker features and redundant features may lead to poor performance when evaluating speaker verification systems on an unseen domain. To address this issue, we propose an InfoMax domain separation and adaptation network (InfoMax–DSAN) to disentangle the domain-specific features and domain-invariant speaker features based on domain adaptation techniques. A frame-based mutual information neural estimator is proposed to maximize the mutual information between frame-level features and input acoustic features, which can help retain more useful information. Furthermore, we propose adopting triplet loss based on the idea of self-supervised learning to overcome the label mismatch problem. Experimental results on VOiCES Challenge 2019 demonstrate that our proposed method can help learn more discriminative and robust speaker embeddings.

**Index Terms**— Speaker verification, domain adaptation, mutual information, self-supervised learning.

1. INTRODUCTION

Speaker verification (SV) has recently attracted increasing attention, and it has been widely used in real-world biometric applications. Traditional state-of-the-art SV systems use factor analysis to extract i-vectors as speaker embeddings [1]. With the development of deep learning, more and more researchers are committed to using deep neural models to improve speaker verification performance. Among them, the x-vector [2] proposed by Snyder et al. has achieved great success and become state-of-the-art. However, deep neural network (DNN) embeddings require a large amount of training data to achieve accurate predictions or classifications. Unfortunately, acquiring sufficient labelled training data is a great challenge because the data collection process is time consuming and labour intensive. The performance will also be degraded when deploying an existing system to an unseen environment because of the domain mismatch between the trained and deployed environments.

Domain mismatch occurs when the speech is collected from different acoustic environments. While data augmentation [2] can increase the number of possible acoustic environments a system may encounter, it can only partially alleviate the domain mismatch problem. A better solution is domain adaptation (DA). Domain adaptation has been applied to either adapt PLDA models to fit the target data [3] or find a shared embedding space in which the features contain more speaker information but less unrelated information. Finding a common embedding space is a more general approach in that it is independent of the backend. Previous methods tried to align the features by minimizing the cosine distance and mean squared error [4], or by minimizing the maximum mean discrepancy (MMD) [5] at the embedding layer. More recent research is based on adversarial learning [6, 7, 8]. The essence of these methods is the same: find a shared space where the feature distributions have low discrepancy across multiple domains. However, most of these studies treat DA as a transformation of speaker embedding vectors instead of using DA to extract domain-invariant embedding directly.

Although invariant representations can be obtained [5], transformation-based DA may not perform well on the target domain because of the difference in label distributions [9]. For speaker verification, the speaker IDs of different domains are different, which increases the difficulty of DA. Therefore, it is hard to train a speaker embedding network with domain-invariant property. In addition, insufficient data from the target domain would also limit the model to find a common embedding space for all possible domains.

To overcome these difficulties, we propose a novel framework to perform domain adaptation. The framework (1) incorporates domain adaptation directly into the training of speaker embedding extractor, (2) uses self-supervised learning to overcome the label mismatch problem without using labels from the target domain, and (3) introduces a frame-based mutual information neural estimator to maximize the mutual information between the frame-level features and input acoustic features. These actions are beneficial for the network to learn informative representations. Experimental results on the VOiCES 2019 dataset demonstrate the effectiveness of maximizing frame-based mutual information in speaker embedding networks.

This work was supported by the RGC of Hong Kong SAR, Grant No. PolyU 152137/17E and National Natural Science Foundation of China (NSFC), Grant No. 61971371.
2. INFORMATIO DOMAIN SEPARATION AND ADAPTATION NETWORK

2.1. Overview

Traditional domain adaptation approaches utilize data from two domains: labelled source domain and unlabeled target domain. These methods generally require a large number of samples from the target domain; otherwise, the performance may degrade. To overcome this issue, our proposed framework, as Fig. 1 shows, considers three domains: the source domain $D^s$, the augmentation of the source domain $D^{aug}$, and the target domain $D^t$.

Denote a labelled dataset from the source domain as $X^s = \{(x^s_i, y_i)\}_{i=1}^{N_s}$, a labelled augmented set of the source domain as $X^{aug} = \{(x^{aug}_i, y_i)\}_{i=1}^{N^{aug}}$, and an unlabelled dataset from the target domain as $X^t = \{x^t_i\}_{i=1}^{N_t}$. In addition to finding the domain-specific features from two totally different domains, we create a new domain $D^{aug}$ that shares some properties with the source domain. The new domain consists of augmented samples that are created by introducing some background noise and reverberation to the labelled data. Therefore, background noise and reverberation are the factors causing the domain mismatch between $D^s$ and $D^{aug}$. In contrast, the domain mismatch between $D^t$ and other two domains are caused by various factors, e.g., microphone variability, different types and levels of noises, and different room acoustics. Because the factors are unknown and speakers in the target domain are different from those in the source domain, constructing a domain with known factors can help training a feature extractor to extract domain-specific features.

As shown in Fig. 1, the desired speaker embeddings are extracted from the speaker-feature extractor $G_{spk}$. A speaker classifier $D_{spk}$ is connected to $G_{spk}$ to ensure that the embeddings $\mu^s$ and $\mu^{aug}$ are speaker discriminative. This is achieved by minimizing the speaker identification loss:

$$ \mathcal{L}_{spk} = \mathbb{E}_{x^s \sim D^s} \left[ - \sum_{k=1}^{K} y_{spk}^{(k)}(x^s) \log D_{spk}(G_{spk}(x^s))_k \right] $$

$$ + \mathbb{E}_{x^{aug} \sim D^{aug}} \left[ - \sum_{k=1}^{K} y_{spk}^{(k)}(x^{aug}) \log D_{spk}(G_{spk}(x^{aug}))_k \right], $$

where $k$ denotes the $k$-th output of the speaker discriminator and $y_{spk}^{(k)}$ is a one-hot encoded speaker label.

Meanwhile, $G_{spk}$ is trained adversarially to extract the common features $\mu^s$ and $\mu^t$ from the source and target domains. The adaptation network $D_{adt}$ is trained to indicate whether the common features come from the source domain or the target domain. Therefore, $D_{adt}$ and $G_{spk}$ are trained with the following minimax loss:

$$ \min_{\mathcal{G}_{spk}} \max_{\mathcal{D}_{adt}} \mathbb{E}_{x \sim D^s} \left[ \log D_{adt}(G_{spk}(x^s)) \right] $$

$$ + \mathbb{E}_{x^{aug} \sim D^{aug}} \left[ \log \left[ 1 - D_{adt}(G_{spk}(x^{aug})) \right] \right]. $$

The domain-feature extractor $G_{dom}$ extracts domain-specific features ($d^s$, $d^{aug}$, $d^t$) from the acoustic features. A domain discriminator $D_{dom}$ disentangles the domain-specific features by predicting to which of the three domains the domain-specific features should belong. To this end, the loss function for $D_{dom}$ is set to

$$ \mathcal{L}_{dom} = \mathbb{E}_{x^{s} \sim D^{s}} \left[ - \log D_{dom}(G_{dom}(x^{s}))_0 \right] $$

$$ + \mathbb{E}_{x^{aug} \sim D^{aug}} \left[ - \log D_{dom}(G_{dom}(x^{aug}))_1 \right] $$

$$ + \mathbb{E}_{x^{t} \sim D^{t}} \left[ - \log D_{dom}(G_{dom}(x^{t}))_2 \right]. $$

where $(\cdot)_k$ denotes the $k$-th output of $D_{dom}$ and $k = 0, 1, 2$ correspond to the target, source, and augmented-source, respectively.

To further ensure that the domain features $d$’s contain domain information from the inputs, a frame-based mutual information neural estimator (MINE) is applied. Details will be given in Section 2.2.

To disentangle the domain-specific features from the speaker features, we introduce a constraint ($\mathcal{L}_{diff}$ in Fig. 1) to make the speaker embeddings ($\mu$) and the domain embeddings ($d$) different. In this study, we computed the difference loss in two ways: (1) making $d$ and $\mu$ orthonormal as suggested in [10] and (2) minimizing their mutual information as proposed in [11].

2.2. Mutual Information Neural Estimation

Feature learning based on mutual information has attracted increased interest in recent years. In speaker verification, many researchers applied MINE to compute the mutual information between the embeddings and frame-level features [12, 13], or between multiple embeddings [14]. Studies seldom attempt to employ MINE on frame-level features. However, when computing the mutual information between the frame-level and segment-level features, the problem of dimensionality imbalance or information loss would occur. To address these issues
Mutual information (MI) is a measure of information shared between random variables. The value of mutual information is not necessary, we may use a MINE that is based on Jensen-Shannon divergence [16] instead of KL-divergence. Then, the mutual information loss becomes

$$I_{\text{MINE}}(X, Z) = -\frac{1}{\alpha} \log (1 + e^{XZ})$$

where $\alpha$ is a hyperparameter that controls the contribution of the sub-losses to the whole loss. In our experiments, we set $\alpha = 1.0, \beta = 0.0, \gamma = 0.5, \lambda = 0.5$.

### 3. EXPERIMENTAL SETTING

#### 3.1. Data Preparation

The source domain comprises utterances from the development sets of VoxCeleb1 [18] and VoxCeleb2 [19]. We followed the data augmentation procedure in Kaldi to create the augmented set by adding noise, babble, and music from MUSAN [20] and reverberation from the RIR dataset [21]. The target domain comprises utterances from the VOiCES Challenge 2019 [22], but we only used the development set for training. We followed the Kaldi recipe to extract 40-dimensional filter bank features with a frame length of 25ms at 10ms shift and used Kaldi energy-based voice activity detection (VAD) to remove silence frames.

#### 3.2. Implementation Details

In this study, both feature extractors $G_{\text{dom}}$ and $G_{\text{spk}}$ follow the x-vector framework [2], except that they only have one segment-level layer after the pooling layer. The speaker classifier $D_{\text{spk}}$ comprises a linear layer and a softmax output layer.
Table 1. Performance on the original VOiCES 2019 development set and evaluation set. $D^s$ comprises the development sets of VoxCeleb1&2, $D^{aug}$ corresponds to the augmented set of VoxCeleb1&2, and $D^t$ comprises utterances from the development set of VOiCES 2019. Two metrics were used for computing $L_{diff}$: mutual information (MI) and orthogonality (ort).

<table>
<thead>
<tr>
<th>Row</th>
<th>System</th>
<th>Source domain</th>
<th>Target domain</th>
<th>MINE</th>
<th>$L_{triplet}$</th>
<th>$D_{adt}$</th>
<th>$L_{diff}$</th>
<th>VOiCES dev.</th>
<th>VOiCES eval.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TDNN [4]</td>
<td>$D^s$</td>
<td>$D^{aug}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.18</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>TDNN (G_{spk})</td>
<td>${D^s, D^{aug}}$</td>
<td>$D^t$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2.35</td>
<td>0.2596</td>
</tr>
<tr>
<td>3</td>
<td>TDNN (G_{spk})</td>
<td>${D^s, D^{aug}}$</td>
<td>$D^t$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2.16</td>
<td>0.2627</td>
</tr>
<tr>
<td>4</td>
<td>TDNN (G_{spk})</td>
<td>${D^s, D^{aug}}$</td>
<td>$D^t$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2.29</td>
<td>0.2453</td>
</tr>
<tr>
<td>5</td>
<td>TDNN (G_{spk})</td>
<td>${D^s, D^{aug}}$</td>
<td>$D^t$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2.31</td>
<td>0.2645</td>
</tr>
<tr>
<td>6</td>
<td>TDNN (G_{spk})</td>
<td>${D^s, D^{aug}}$</td>
<td>$D^t$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2.54</td>
<td>0.2671</td>
</tr>
<tr>
<td>7</td>
<td>InfoMax–DSAN</td>
<td>$D^s$</td>
<td>$D^{aug}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>MI</td>
<td>2.17</td>
<td>0.2418</td>
</tr>
<tr>
<td>8</td>
<td>InfoMax–DSAN</td>
<td>$D^s$</td>
<td>$D^{aug}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>MI</td>
<td>2.33</td>
<td>0.2577</td>
</tr>
<tr>
<td>9</td>
<td>InfoMax–DSAN</td>
<td>$G_{dom}$ &amp; $G_{spk}$</td>
<td>$D^s$</td>
<td>$D^{aug}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>MI</td>
<td>3.29</td>
</tr>
<tr>
<td>10</td>
<td>InfoMax–DSAN</td>
<td>$G_{dom}$ &amp; $G_{spk}$</td>
<td>$D^s$</td>
<td>$D^{aug}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>MI</td>
<td>2.31</td>
</tr>
<tr>
<td>11</td>
<td>InfoMax–DSAN</td>
<td>$G_{dom}$ &amp; $G_{spk}$</td>
<td>$D^s$</td>
<td>$D^{aug}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>MI</td>
<td>2.06</td>
</tr>
</tbody>
</table>

The domain discriminator $D_{dom}$ has two fully connected layers with 128 nodes in each layer, while $D_{adt}$ is a two-layer fully connected network with 512 nodes in each layer. The mutual information neural estimator (MINE) is a three-layer fully connected network with 512 nodes in each layer. Also, $G_{dom}$ uses statistics pooling while $G_{spk}$ uses attentive statistics pooling [23]. All models were trained using small chunks of acoustic sequences with a chunk length of 200 frames.

The proposed InfoMax domain separation and adaptation network (InfoMax–DSAN) was trained using the RAdam optimizer [24] with an initial learning rate of 0.001. The dimension of domain-specific embeddings $d^s$ is 64, while that of the speaker embeddings $\mu$’s is 256. Probabilistic linear discriminant analysis (PLDA) [25] was used as the backend for scoring. Before PLDA training, the extracted speaker embeddings were projected onto a 150-dimensional space by LDA, followed by whitening and length normalization. We also performed the symmetric score normalization (S-norm) [26] to normalize the scores of the VOiCES evaluation set, using the speech from the VOiCES development set as the cohort.

4. RESULTS AND DISCUSSIONS

Table 1 shows the performance of different systems on the VOiCES 2019 dataset. A “x” in the fourth column means that we used the development sets of VoxCeleb1&2 and their augmented sets to train the speaker embedding networks $G_{spk}$ without domain adaptation. Rows 2–6 present the performance of $G_{spk}$-based systems (Fig. 1 without $G_{dom}$ and $D_{dom}$). It can be observed that the performance can be improved when applying the MINE to maximize the mutual information between the speaker features and the input acoustic features (Row 4). This indicates that the proposed frame-based MINE can effectively help to extract informative features.

The lower part of Table 1 (Rows 7–11) shows the performance of InfoMax–DSAN based systems. The first two systems were trained by considering that the target domain can be perfectly simulated by the augmented data, i.e., we can replace $x^t$ in Fig. 1 by $x^{aug}$. Two metrics were used to evaluate $L_{diff}$: mutual information (MI) and orthogonality (ort). Overall, minimizing the mutual information between domain features and speaker features performs better. The improvement of these two systems suggests that the extracted speaker embeddings are more robust. We also compared the performance of systems with and without the triplet loss (Rows 9 and 11). Without $L_{triplet}$, the performance of the domain adaptation network performs even worse than the baseline. On the contrary, the best performance can be obtained when training the model with the triplet loss. This result demonstrates the importance of considering the label distribution of the target domain when applying domain adaptation. Ignoring the label mismatch problem would degrade performance. Rows 10 and 11 suggest that the MINE can help disentangle redundant features from speaker features. Another observation is that the TDNN system could not benefit from learning domain-invariant features only (Row 6). This implies that disentangling the domain-specific features is more effective.

5. CONCLUSIONS

In this paper, we propose a domain adaptation framework to disentangle the task-related and task-irrelevant features in the front-end of speaker verification systems. A frame-based mutual information neural estimation is proposed to learn informative features by maximizing the mutual information between the input acoustic features and frame-level features from a TDNN. We also introduce the triplet loss to address the label mismatch problem in domain adaptation. The experiments on the VOiCES 2019 corpus demonstrate that our proposed approaches can effectively help extract more robust speaker features.
6. REFERENCES


