INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Mutual Information-based Embedding Decoupling for Generalizable Speaker Verification

Jianchen Li[1], *Jiqing Han*[1], *Shiwen Deng*[2], *Tieran Zheng*[1], *Yongjun He*[1], *Guibin Zheng*[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]School of Mathematical Science, Harbin Normal University, Harbin, China

lijianchen@hit.edu.cn

Abstract

Domain shift is a challenging problem in speaker verification, especially when dealing with unseen target domains. Recently, embedding decoupling-based methods have shown their effectiveness. Typically, domain information is extracted by a domain classification loss and then decoupled from speaker embeddings. However, the domain classification loss fails to ensure that only domain information is encoded in domain embeddings. This paper proposes a novel mutual information-based embedding decoupling framework, in which the domain information is extracted by maximizing the mutual information between different speaker sample pairs in the same domain. Then the domain information is removed from speaker embeddings by minimizing mutual information between speaker and domain embeddings. Experiments indicate that our method can improve the generalization and outperform domain classification-based decoupling methods.

Index Terms: speaker verification, domain generalization, embedding decoupling, mutual information

1. Introduction

Automatic Speaker Verification (ASV) is the task that determines whether a test utterance belongs to the enrollment speaker [1]. Currently, speaker embedding models [2, 3] have become state-of-the-art ASV models. However, speaker embedding models will suffer significant performance degradation when training and test utterances are not independent and identically distributed (*i.i.d*). Such a problem with different distributions is known as domain shift [4], where the training and test distributions are called the source and target domains, respectively. A straightforward way to alleviate domain shift is to collect some target data to transfer a source-domain-trained speaker embedding model to the target domain, i.e., the domain adaptation method [5]. However, in some scenarios, the target domain data is difficult to collect or even unknown before deployment, which limits the applicability of domain adaptation.

In the absence of target domain data, domain generalization methods [6] have emerged to learn domain-invariant speaker embeddings from multiple source domains (e.g., vlog, interview, and other genre utterances) that generalize well in unseen target domains. For ASV models, existing domain generalization methods consist of domain alignment, model-agnostic meta-learning, and embedding decoupling. Domain alignment [7] is an effective method for learning domain-invariant speaker embeddings, which minimizes certain distances (e.g., cosine distance [8], MMD [9], and JS divergence in domain adversarial training [10]) among multiple source speaker embedding distributions to remove domain information. In order to distinguish speaker information from irrelevant domain information, domain alignment relies on data from different domains for the same speaker, i.e., same-speaker multi-domain utterances. Unfortunately, it is challenging to collect large-scale same-speaker multi-domain corpus in practice. Model-agnostic meta-learning (MAML) [11] learns domain-invariant speaker embedding by simulating training and testing domain shift during training [12]. While this method achieves well generalization, it requires the calculation of second-order gradients, which leads to considerable complexity. Embedding decoupling-based methods [13, 14, 15, 16, 17] typically extract domain-related information by learning domain embeddings, and then strip this information from speaker embeddings to learn domain-invariant speaker embeddings. The embedding decoupling methods have great interpretability and show an excellent capacity for solving domain shift problems in ASV [5, 18]. Thus we focus on the embedding decoupling methods in this paper.

Existing embedding decoupling methods typically learn domain embeddings via a domain classification loss [13, 17, 19]. This causes two issues. Firstly, training with the domain classification loss requires multi-domain utterances for each speaker. Otherwise, if each domain consists of different speakers, the domain classification loss may take advantage of the differences in speaker distributions to distinguish among domains. Thus, the speaker-related information will be encoded in domain embeddings and then stripped from speaker embeddings, which weakens the discriminability of the speaker embeddings. Secondly, the domain classification loss can only learn domain embeddings that contain information that distinguishes among domains, rather than all domain-related information. Thus, this paper aims to propose a new embedding decoupling method that solves both issues.

In this paper, we propose a mutual information-based embedding decoupling method to decompose the original embeddings into the speaker-invariant domain embeddings and the domain-invariant speaker embeddings. Specifically, since the only shared information for data from different speakers in the same domain is domain-related information, we propose a cross-speaker mutual information maximization method to learn the speaker-invariant domain embeddings, which maximizes the mutual information between the sample pairs from different speakers in the same domain. Furthermore, we minimize the mutual information between speaker and domain embeddings to reduce their dependencies and obtain the domain-invariant speaker embeddings. Our method does not require multi-domain utterances for each speaker to learn domain-invariant speaker embeddings. Thus it is more conducive to alleviating domain shift in practice. Experiments on the CN-Celeb [20] corpus indicate that the proposed method can effectively improve the generalization in unseen target domains and outperform domain classification-based decoupling methods.

10.21437/Interspeech.2023-1314

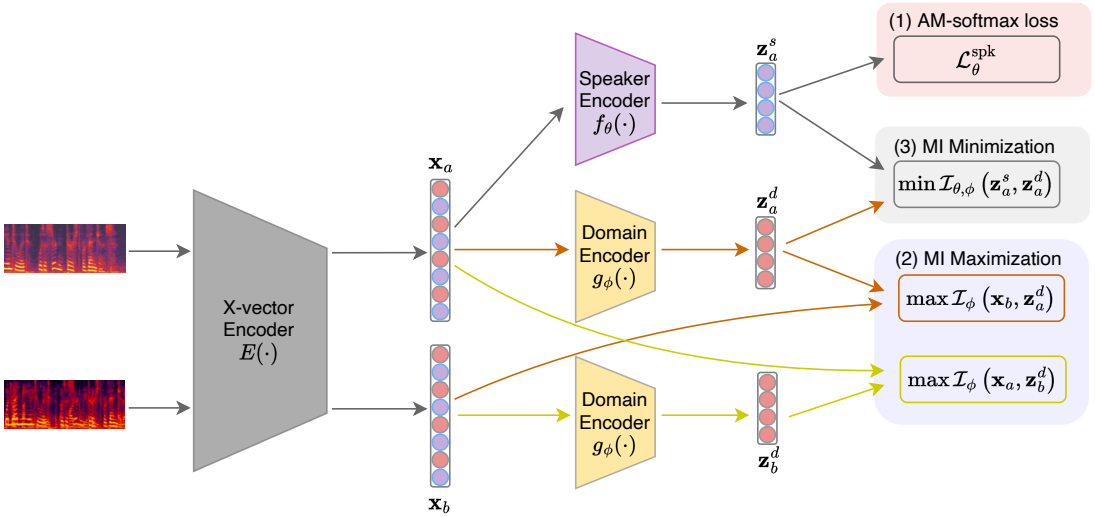


Figure 1: *Overview of the proposed method.* $\mathbf{x}_a$ *and* $\mathbf{x}_b$ *are a pair of embeddings from different speakers in the same domain, and their shared information is domain-related information. (1) AM-softmax is used to learn speaker-related information. (2) Mutual information (MI) maximization is used to learn speaker-invariant domain embeddings* $\mathbf{z}_a^d$ *and* $\mathbf{z}_b^d$*. (3) MI minimization is used to remove irrelevant domain information from speaker embeddings and learn domain-invariant speaker embeddings* $\mathbf{z}_a^s$*.*

2. Methodology

Given a training set consisting of M source domains $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$, where $\mathcal{D}_j$ denotes the j-th source domain. In most cases, each source domain typically contains different speakers. As shown in Figure 1, each utterance is transformed into a raw speaker embedding $\mathbf{x}$ via a pre-trained and frozen encoder $E(\cdot)$ called the x-vector [2] model. To decouple the domain and speaker information in $\mathbf{x}$, the domain encoder $g_\phi(\cdot)$ with parameter ϕ is trained to extract all domain-related information from $\mathbf{x}$. Further, the domain information is removed from the mapped speaker embeddings extracted by the speaker encoder $f_\theta(\cdot)$ with parameter θ, yielding the domain-invariant speaker embeddings.

2.1. Speaker Embedding Learning

The mapped speaker embeddings should maintain the raw speaker-related information. To this end, AM-softmax [21] loss is adopted to learn the mapped speaker embeddings:

$$\mathcal{L}_\theta^{\text{spk}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cdot\left(\cos\delta_{y_i,i}-m\right)}}{e^{s\cdot\left(\cos\delta_{y_i,i}-m\right)}+\sum_{c=1,c\neq y_i}^{C}e^{s\cdot\cos\delta_{c,i}}}, \tag{1}$$

where $y_i \in \{1, 2, \ldots, C\}$ is the ground-truth speaker label of $\mathbf{x}_{a_i}$, C denotes the number of speakers from all source domains, $\delta_{c,i}$ is the angle between the weight vector $\mathbf{w}_c$ and the speaker embedding $\mathbf{z}_{a_i}^s = f_\theta\left(\mathbf{x}_{a_i}\right)$, s and m denote the scaling factor and the margin, respectively.

2.2. Mutual Information Maximization

Suppose $\mathbf{x}_a$ and $\mathbf{x}_b$ are samples from different speakers in the same domain, and their domain embeddings are $\mathbf{z}_a^d = g_\phi\left(\mathbf{x}_a\right)$ and $\mathbf{z}_b^d = g_\phi\left(\mathbf{x}_b\right)$. Intuitively, the shared information between $\mathbf{x}_a$ $(\mathbf{x}_b)$ and $\mathbf{z}_b^d$ $(\mathbf{z}_a^d)$ can be any speaker-independent domain information, such as genre, background noise, and language category. Thus, we train the domain encoder to extract all shared domain information except speaker information. Since mutual information (MI) [22] can measure and quantify the information shared between two random variables and larger MI indicates more shared information, we maximize the MI between $\mathbf{x}_a$ $(\mathbf{x}_b)$ and $\mathbf{z}_b^d$ $(\mathbf{z}_a^d)$ to learn the domain embeddings:

$$\max_\phi\left(\mathcal{I}\left(\mathbf{x}_a, \mathbf{z}_b^d\right) + \mathcal{I}\left(\mathbf{x}_b, \mathbf{z}_a^d\right)\right), \tag{2}$$

where:

$$\mathcal{I}\left(\mathbf{x}_a, \mathbf{z}_b^d\right) = \mathbb{E}_{p\left(\mathbf{x}_a, \mathbf{z}_b^d\right)}\left[\log\frac{p(\mathbf{x}_a, \mathbf{z}_b^d)}{p(\mathbf{x}_a)p(\mathbf{z}_b^d)}\right]. \tag{3}$$

In practice, it is challenging to directly compute MI as the joint and marginal distributions are unknown and intractable. Thus we focus on the MI estimator Deep InfoMax [23], which can obtain a lower bound of MI. Take $\mathcal{I}\left(\mathbf{x}_a, \mathbf{z}_b^d\right)$ as an example, suppose we have N sample pairs $\{(\mathbf{x}_{a_i}, \mathbf{x}_{b_i})\}_{i=1}^N$ drawn from $p\left(\mathbf{x}_a, \mathbf{x}_b\right)$, the lower bound of $\mathcal{I}\left(\mathbf{x}_a, \mathbf{z}_b^d\right)$ is defined as:

$$\begin{aligned}\hat{\mathcal{I}}\left(\mathbf{x}_a, \mathbf{z}_b^d\right) = &-\frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{-T_\eta\left(\mathbf{x}_{a_i}, \mathbf{z}_{b_i}^d\right)}\right)\\ &-\frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{T_\eta\left(\tilde{\mathbf{x}}_{a_i}, \mathbf{z}_{b_i}^d\right)}\right),\end{aligned} \tag{4}$$

where $T_\eta : \mathcal{X}\times\mathcal{Z} \to \mathbb{R}$ denotes a neural network with parameters η called the statistics network, the pairs $\{(\tilde{\mathbf{x}}_{a_i}, \mathbf{z}_{b_i}^d)\}_{i=1}^N$ are sampled from the marginal distributions $p\left(\mathbf{x}_a\right)$ and $p(\mathbf{z}_b^d)$. We obtain such data pairs by simply shuffling $\{(\mathbf{x}_{a_i}, \mathbf{z}_{b_i}^d)\}_{i=1}^N$ along the $\mathbf{x}_{a_i}$ axis, where $\{(\mathbf{x}_{a_i}, \mathbf{z}_{b_i}^d)\}_{i=1}^N$ are sampled from the joint distribution $p(\mathbf{x}_a, \mathbf{z}_b^d)$. The lower bound $\hat{\mathcal{I}}\left(\mathbf{x}_b, \mathbf{z}_a^d\right)$ can be calculated in the same way.

The objective function of learning domain embeddings is to maximize the lower bound of MI, which is equivalent to minimizing the following loss:

$$\mathcal{L}_{\phi,\eta}^{\text{dom}} = -\hat{\mathcal{I}}\left(\mathbf{x}_a, \mathbf{z}_b^d\right) - \hat{\mathcal{I}}\left(\mathbf{x}_b, \mathbf{z}_a^d\right). \tag{5}$$

Table 1: *Statistics of the CN-Celeb dataset after custom division. Note that utterances of less than 2s within the same genre and speaker are concatenated to form long utterances rather than discarded.*

Genre	Vlog	Recitation	Speech	Live Broadcast	Interview	Entertainment
# of training speakers	462	224	299	466	1,108	928
# of training utterances	123,504	59,020	41,311	167,070	68,627	38,749
# of evaluation utterances	1,592	1,280	2,184	2,559	7,733	4,049
# of trials	12,736	10,240	17,472	20,472	61,864	32,392

2.3. Mutual Information Minimization

To remove domain information from the mapped speaker embeddings, we minimize the MI between the domain and mapped speaker embeddings to learn domain-invariant speaker embeddings. Specifically, given an input $\mathbf{x}_a$, the objective is defined as:

$$\min_{\theta,\phi} \mathcal{I}\left(\mathbf{z}_a^s, \mathbf{z}_a^d\right), \tag{6}$$

where $\mathbf{z}_a^s = f_\theta(\mathbf{x}_a)$, $\mathbf{z}_a^d = g_\phi(\mathbf{x}_a)$. The minimization of MI involves the calculation of its upper bound. We adopt Contrastive Log-ratio Upper Bound (CLUB) [24] to estimate the MI upper bound. Specifically, given N embedding pairs $\{(\mathbf{z}_{a_i}^s, \mathbf{z}_{a_i}^d)\}_{i=1}^N$, the CLUB is defined as:

$$\begin{aligned}\tilde{\mathcal{I}}\left(\mathbf{z}_a^s, \mathbf{z}_a^d\right) =& \frac{1}{N}\sum_{i=1}^{N} \log q_\gamma\left(\mathbf{z}_{a_i}^d | \mathbf{z}_{a_i}^s\right) \\ &- \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \log q_\gamma\left(\mathbf{z}_{a_j}^d | \mathbf{z}_{a_i}^s\right),\end{aligned} \tag{7}$$

where $q_\gamma(\mathbf{z}_a^d|\mathbf{z}_a^s)$ is a variational distribution (e.g., Gaussian distribution) with parameters γ, which can be learned by maximizing the log-likelihood loss $\mathcal{L}_\gamma^{\text{lld}} = \frac{1}{N}\sum_{i=1}^{N} \log q_\gamma\left(\mathbf{z}_{a_i}^d | \mathbf{z}_{a_i}^s\right)$. Subsequently, the loss function for learning decoupled embeddings is defined as:

$$\mathcal{L}_{\theta,\phi}^{\text{dec}} = \tilde{\mathcal{I}}\left(\mathbf{z}_a^s, \mathbf{z}_a^d\right). \tag{8}$$

2.4. Overall Optimization

The overall objective function is the weighted sum of all proposed loss functions:

$$\mathcal{L}^{\text{total}} = \lambda^{\text{dom}}\mathcal{L}_{\phi,\eta}^{\text{dom}} + \lambda^{\text{spk}}\mathcal{L}_{\theta}^{\text{spk}} + \lambda_t^{\text{dec}}\mathcal{L}_{\theta,\phi}^{\text{dec}}, \tag{9}$$

where λ^{dom} and λ^{spk} are fixed trade-off parameters, λ_t^{dec} is a dynamic parameter w.r.t. the t-training iteration. Since $f_\theta(\cdot)$ and $g_\phi(\cdot)$ are under-fitted in the early training iterations, λ_t^{dec} is set to a small initial value to avoid CLUB disrupting the training of $f_\theta(\cdot)$ and $g_\phi(\cdot)$, and is continuously increased during training. We implement λ_t^{dec} as a exponentially function:

$$\lambda_t^{\text{dec}} = \lambda_T^{\text{dec}}\left(\frac{2}{1+e^{-10t/T}} - 1\right), \tag{10}$$

where λ_T^{dec} is the parameter at the final iteration T.

In practice, $\mathcal{L}_{\theta,\phi}^{\text{dec}}$ relies on the accurate approximation of $q_\gamma(\mathbf{z}_a^d|\mathbf{z}_a^s)$ to the conditional distribution $p(\mathbf{z}_a^d|\mathbf{z}_a^s)$. To this end, $\mathcal{L}_\gamma^{\text{lld}}$ and $\mathcal{L}^{\text{total}}$ are updated alternately during the training. Specifically, at each training iteration, we first update $q_\gamma(\mathbf{z}_a^d|\mathbf{z}_a^s)$ by maximizing $\mathcal{L}_\gamma^{\text{lld}}$, then freeze γ and compute $\mathcal{L}^{\text{total}}$. Finally, the gradient is back-propagated to $f_\theta(\cdot)$ and $g_\phi(\cdot)$.

3. Experiments

3.1. Datasets

VoxCeleb2 [25] is adopted to train the x-vector model, which contains 1,092,009 utterances from 5,994 speakers. Online data augmentation [26] is performed with MUSAN [27] and RIRs [28]. The RIRs are limited to small and medium rooms.

CN-Celeb is adopted to train the speaker and domain encoders and evaluate the proposed method. This database contains two datasets: CN-Celeb1 [29] and CN-Celeb2 [20]. The former contains 126,532 utterances from 997 speakers, and the latter contains 524,787 utterances from 1,996 speakers. Both datasets contain 11 genres of utterances, which can be considered as 11 domains. We choose 6 domains containing large amounts of training data to conduct our experiments: Vlog, Recitation, Speech, Live Broadcast, Interview, and Entertainment. It should be noted that a large proportion of speakers in CN-Celeb contain utterances in only one genre, making it more challenging to address the domain shift problem.

To evaluate the performance of domain generalization, we sequentially select one genre as the unseen target domain and the remaining five as the seen source domains ($M = 5$), forming six sets of experiments. We train the speaker and domain encoders on the 5-source training set, and evaluate on the target evaluation set. The officially released evaluation set consists of utterances from each genre, but contains very few Vlog and Recitation utterances. To make the evaluation more convincing, we divide some training speakers with Vlog and Recitation utterances into the evaluation set and create the new trials file for each genre. The trials files are created following the setting of the VoxCeleb evaluation set to form balanced target and non-target pairs. Specifically, non-target pairs are constructed within the same genre, while target pairs are in the cross-genre scenario to evaluate the performance in two cases: the domain shift between enrollment and test utterances and the domain shift between the training and evaluation sets. The statistics of CN-Celeb after the custom division are shown in Table 1.

3.2. Implementation Details

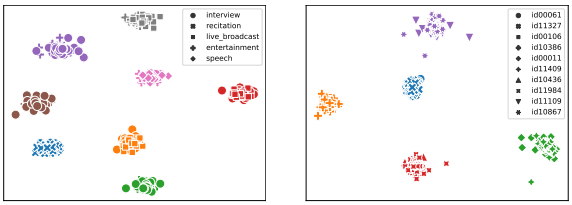
ResNet-34 [30] is adopted as the x-vector model to extract the raw speaker embeddings. The inputs acoustic features are 80-dimensional FBanks. Cepstral mean normalization (CMN) is applied, and each training utterance is cut into 200-frame chunks to create the same length inputs. For the network architecture, multi-head attention pooling [31] is adopted to convert the frame-level embeddings to the segment-level embeddings. For training, the network is optimized by the AM-softmax loss with a scaling factor of 35 and a margin of 0.2. Stochastic gradient descent (SGD) is adopted as the network optimizer, where the initial learning rate is 0.02, and the weight decay is 5e-4. The ReduceLROnPlateau scheduler is applied to update the learning rate. Once trained, the 256-dimensional embeddings

Table 2: *Results in various unseen target domains. Live. and Enter. denote live broadcast and entertainment, respectively. ResNet-34 refers to the results of raw speaker embeddings. + spk encoder refers to training the speaker encoder with only the AM-softmax loss $\mathcal{L}_{\theta}^{spk}$. MAML is the state-of-the-art domain generalization method in speaker verification. EER refers to Equal Error Rate in %, and mDCF refers to Minimum Detection Cost with P_{target} = 0.01. * denotes the results of re-implementation. Best in bold.*

Model	Vlog		Recitation		Speech		Live.		Interview		Enter.		Average	
	EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF
ResNet-34	14.07	0.632	11.78	0.681	7.66	0.501	12.02	0.687	11.98	0.682	12.51	0.601	11.67	0.631
+ spk encoder	13.95	0.697	10.31	0.519	6.91	0.403	10.51	0.590	12.02	0.601	11.60	0.682	10.88	0.582
Our proposed	**13.19**	**0.647**	**9.78**	**0.497**	**6.18**	**0.377**	**9.82**	**0.549**	**11.13**	**0.579**	**11.29**	**0.625**	**10.23**	**0.546**
*MAML [12]	13.31	0.663	10.05	0.511	6.29	0.381	10.13	0.571	11.45	0.588	11.41	0.661	10.44	0.563

Table 3: *Average results of decoupled speaker embeddings (Speaker) and domain embeddings (Domain) in 6 unseen target domains, where the domain embeddings are learned via the domain classification loss or MI maximization.*

Domain Loss	Speaker		Domain	
	EER	mDCF	EER	mDCF
Classification	10.69	0.571	41.12	0.991
MI Maximization	10.23	0.546	47.19	0.997



(a) Speaker Embeddings (b) Domain Embeddings

Figure 2: *t-SNE plots of decoupled (a) speaker embeddings and (b) domain embeddings. Embedding points are colored by speaker labels in (a) while colored by domain labels in (b).*

are extracted as inputs to the speaker and domain encoders.

For embedding decoupling, the speaker encoder is implemented by 2 fully-connected (FC) layers with 256, 128 neurons. The domain encoder and the statistic network in Deep InfoMax are implemented by 3 fully-connected (FC) layers with 512, 512, 128, and 512, 512, 1 neurons, respectively. 128-dimensional domain and mapped speaker embeddings are extracted by the last layer of the encoders. The variational distribution $q_\gamma(\mathbf{z}_a^d|\mathbf{z}_a^s)$ in CLUB is parameterized by a Gaussian distribution. To obtain a more accurate approximation to the conditional distribution $p(\mathbf{z}_a^d|\mathbf{z}_a^s)$, the mean and variance vectors of the gaussian distribution are obtained by a larger 4-layer FC network, and each hidden layer consists of 512 neurons. All networks are optimized by the Adam optimizer [32] with the learning rate of 1e-4 and the weight decay of 5e-4. The batch size of input pairs is 128. The scaling factor and margin of AM-softmax loss are 30 and 0.2, respectively. The trade-off parameters λ^{dom} and λ^{spk} are 20 and 1, respectively. The final iteration T is 16,000 and λ_T^{dec} is 0.002.

3.3. Results

Table 2 shows the performance of the various models in the 6 unseen target domains and their average results. Training the speaker encoder with only the AM-softmax loss boosts the performance due to the incorporation of the CN-Celeb training set. Our proposed embedding decoupling method prevents the speaker encoder from learning all domain-related information, thus achieving the best performance. Compared with ResNet-34 and + spk encoder, the relative reductions of the average results in EER are 12.3% and 6.0%, and the relative reductions in minimum DCF are 13.5% and 6.2%, respectively. Our method also outperforms MAML [12], which is the state-of-the-art domain generalization method in ASV, and our method does not require calculating complex second-order gradients.

In order to demonstrate the superiority of MI maximization over the domain classification in extracting domain embeddings, we replace the MI maximization loss in Eq. (9) with a domain classification loss and show the average results of the speaker and domain embeddings in Table 3. For the domain embeddings, the average result of MI maximization is worse than that of the domain classification loss, which indicates that the domain embeddings obtained by MI maximization contain less speaker information. For the speaker embeddings, the average result of MI maximization outperforms that of the domain classification loss, which indicates that learning all domain-related information by MI maximization is more conducive to embedding decoupling.

To intuitively illustrate the effectiveness of embedding decoupling, we show the t-SNE plots of speaker embeddings from 8 random speakers with multi-domain data and domain embeddings from 5 source domains (when the target domain is Vlog). As can be seen, the speaker embeddings of each speaker cluster in Figure 2(a) are indistinguishable based on the domain labels, and the domain embeddings of each domain cluster in Figure 2(b) are indistinguishable based on the speaker labels, indicating that our method can obtain domain-invariant speaker embeddings and speaker-invariant domain embeddings.

4. Conclusions

In this paper, we proposed a mutual information-based embedding decoupling method to improve domain generalization capabilities. Compared with the domain classification, our mutual information maximization method was able to avoid the presence of speaker information in domain embeddings. Then domain-invariant speaker embeddings were obtained by minimizing mutual information. In addition, our method did not require collecting multi-domain utterances for each speaker. Experiments indicated that our method learned a generalizable speaker embedding model and outperformed the previous domain classification-based decoupling methods.

5. References

[1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[3] X. Qin, N. Li, C. Weng, D. Su, and M. Li, "Simple attention module based speaker verification with iterative noisy label detection," in *Proc. ICASSP*, 2022, pp. 6722–6726.

[4] J. Li, J. Han, and H. Song, "Cdma: Cross-domain distance metric adaptation for speaker verification," in *Proc. ICASSP*, 2022, pp. 7197–7201.

[5] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[6] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2022.

[7] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2022.

[8] D. Cai, W. Cai, and M. Li, "Within-sample variability-invariant loss for robust speaker recognition under noisy environments," in *Proc. ICASSP*, 2020, pp. 6469–6473.

[9] Z. Wang, W. Xia, and J. H. Hansen, "Cross-Domain Adaptation with Discrepancy Minimization for Text-Independent Forensic Speaker Verification," in *Proc. Interspeech*, 2020, pp. 2257–2261.

[10] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang, and B. Xia, "Training multi-task adversarial network for extracting noise-robust speaker embedding," in *Proc. ICASSP*, 2019, pp. 6196–6200.

[11] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 1126–1135.

[12] J. Kang, R. Liu, L. Li, Y. Cai, D. Wang, and T. F. Zheng, "Domain-Invariant Speaker Vector Projection by Model-Agnostic Meta-Learning," in *Proc. Interspeech*, 2020, pp. 3825–3829.

[13] X. Qin, N. Li, W. Chao, D. Su, and M. Li, "Cross-Age Speaker Verification: Learning Age-Invariant Speaker Embeddings," in *Proc. Interspeech*, 2022, pp. 1436–1440.

[14] P.-G. Noé, M. Mohammadamini, D. Matrouf, T. Parcollet, A. Nautsch, and J.-F. Bonastre, "Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation," in *Proc. Interspeech*, 2021, pp. 1902–1906.

[15] W. Kang, M. J. Alam, and A. Fathan, "MIM-DG: Mutual information minimization-based domain generalization for speaker verification," in *Proc. Interspeech 2022*, 2022, pp. 3674–3678.

[16] W. H. Kang, J. Alam, and A. Fathan, "Domain Generalized Speaker Embedding Learning via Mutual Information Minimization," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 178–184.

[17] S. H. Mun, M. H. Han, M. Kim, D. Lee, and N. S. Kim, "Disentangled speaker representation learning via mutual information minimization," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 89–96.

[18] M. Sang, W. Xia, and J. H. Hansen, "Deaan: Disentangled embedding and adversarial adaptation network for robust speaker representation learning," in *Proc. ICASSP*, 2021, pp. 6169–6173.

[19] F. Tong, S. Zheng, H. Zhou, X. Xie, Q. Hong, and L. Li, "Deep Representation Decomposition for Rate-Invariant Speaker Verification," in *Proc. The Speaker and Language Recognition Workshop (Odyssey)*, 2022, pp. 228–232.

[20] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "Cn-celeb: Multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[21] 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.

[22] J. B. Kinney and G. S. Atwal, "Equitability, mutual information, and the maximal information coefficient," *Proceedings of the National Academy of Sciences*, vol. 111(9), pp. 3354–3359, 2014.

[23] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations*, 2019.

[24] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "CLUB: A contrastive log-ratio upper bound of mutual information," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, 2020, pp. 1779–1788.

[25] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[26] W. Cai, J. Chen, J. Zhang, and M. Li, "On-the-fly data loader and utterance-level aggregation for speaker and language recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1038–1051, 2020.

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

[28] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[29] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: A challenging chinese speaker recognition dataset," in *Proc. ICASSP*, 2020, pp. 7604–7608.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[31] M. India, P. Safari, and J. Hernando, "Self Multi-Head Attention for Speaker Recognition," in *Proc. Interspeech*, 2019, pp. 4305–4309.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR)*, 2015.