INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Disentangled Representation Learning for Multilingual Speaker Recognition

Kihyun Nam[1†], *Youkyum Kim*[1†], *Jaesung Huh*[2], *Hee-Soo Heo*[3], *Jee-weon Jung*[4], *Joon Son Chung*[1]

[1]Korea Advanced Institute of Science and Technology, South Korea
[2]University of Oxford, United Kingdom
[3]Naver Corporation, South Korea
[4] Carnegie Mellon University, USA

joonsc@kaist.ac.kr

Abstract

The goal of this paper is to learn robust speaker representation for bilingual speaking scenario. The majority of the world's population speak at least two languages; however, most speaker recognition systems fail to recognise the same speaker when speaking in different languages.

Popular speaker recognition evaluation sets do not consider the bilingual scenario, making it difficult to analyse the effect of bilingual speakers on speaker recognition performance. In this paper, we publish a large-scale evaluation set named VoxCeleb1-B derived from VoxCeleb that considers bilingual scenarios.

We introduce an effective disentanglement learning strategy that combines adversarial and metric learning-based methods. This approach addresses the bilingual situation by disentangling language-related information from speaker representation while ensuring stable speaker representation learning. Our language-disentangled learning method only uses language pseudo-labels without manual information.

Index Terms: speaker recognition, real conversation, bilingual speaking, disentangled representation learning

1. Introduction

An estimated 60 to 75 percent of the world's population speaks at least two languages [1]. While somebody is speaking in a foreign language, it has been observed that the person's voice sounds different from when speaking in their mother tongue [2]. With recent trends in globalisation, it has become easier to encounter multilingual scenarios. Therefore, the focus on multilingual speaker recognition has become more important [3–7].

While the performance of speaker recognition systems has improved significantly due to recent advances in deep learning [8–15] and the availability of large-scale datasets [16, 17], the state-of-the-art systems fail easily under the language mismatch condition. The popular speaker recognition evaluation sets do not consider bilingual scenarios, making it difficult to analyse their effect on speaker recognition performance. There are a few evaluation datasets that consider bilingual scenarios; however, they are collected from controlled environments like phone-call platform [3] or contain only limited languages [6]. The recent VoxCeleb Speaker Recognition Challenge (VoxSRC) [18] contains some bilingual speakers; however, their evaluation datasets remain private. Hence, to the best of our knowledge, there is no large-scale public evaluation set that takes bilingual speakers into account.

To this end, we publish a large-scale bilingual evaluation set derived from VoxCeleb1 [16], focusing on bilingual speaking problems. We call this test protocol `VoxCeleb1-B` [1]. To increase the scale and the diversity compared to the VoxSRC challenge test set [18], we expand the number of bilingual trials and the number of languages, resulting in a total of 808,574 trials and 15 languages. Moreover, for the first time, we release the manually annotated language labels of VoxCeleb1. More details of VoxCeleb1-B and the language labels are given in Section 2.

Previous literature finds that a speaker's identity information is intertwined with various factors including accent [19], gender [20, 21], age [21], nationality [21], emotion [22, 23], and spoken language [24]. Using the proposed evaluation protocol, we observe that the existing speaker recognition models do not generalise well to bilingual speakers. We suppose that the mismatched prosodic characteristics from bilingual speakers' different languages significantly affect the performance of the speaker recognition models.

To resolve the language-dependent problem, traditional methods on multilingual speaker recognition have mostly utilised combination of probabilistic linear discriminant analysis and scoring functions based on a standard backbone system such as the i-vectors [5, 7, 25]. However, these methods do not ensure language-invariant speaker representations. Other studies [15, 26–33] have proposed two types of disentangled representation learning methods, namely *adversarial learning-based method* and *metric learning-based method*, which isolate nuisance attributes from the speaker representation. Adversarial learning-based method disturbs convergence of non-speaker discriminator, while metric learning-based method minimises distance or similarity between speaker-relevant and non-speaker representations. For adversarial learning-based method, some studies [15, 26–28] utilise the gradient reversal layer (GRL). Although GRL has shown performance improvement in disentanglement of the target information, we find through our experiments that it frequently causes unstable training and is sensitive to hyperparameters. On the other hand, some studies [29, 31–33] propose metric learning-based methods to minimise correlation between speaker representation and non-speaker representations. [29] utilises mean absolute Pearson's correlation (MAPC) minimisation and [31] uses cosine similarity (COS) minimisation. [32, 33] employ mutual information minimisation. However, since [29, 32, 33] perform domain adaptation for different domains, there is no guarantee that they will perform well in the goal of this work, namely *intra-domain disentangled representation learning*. We evaluate the existing methods and our method on the evaluation set which reflects real-world bilingual scenario unlike [31] which conducts experiments on a simulated dataset.

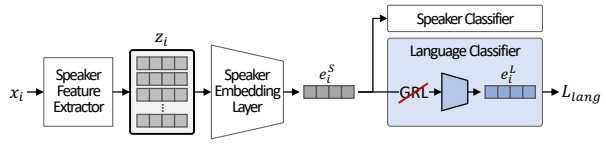
In this work, we propose an effective disentangled representation learning, which weakens the language-dependent information that resides in the speaker representation. The proposed learning strategy combines GRL and MAPC minimisation objective, which overcomes unstable learning and effectively learns language-disentangled speaker representation. The neural network consists of a main speaker recognition model and a spoken language recognition model. During training, language-disentangled learning leverages language pseudo-labels extracted from a spoken language

[†]These authors contributed equally to this work.
[1]The official website's url: https://mm.kaist.ac.kr/projects/voxceleb1-b/

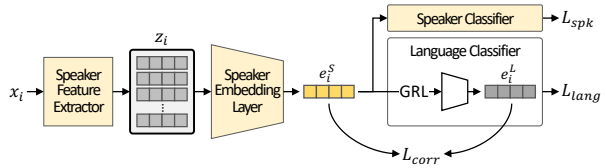
10.21437/Interspeech.2023-1603

Table 1: *Statistics of the VoxCeleb1 test sets, VoxSRC validation sets and VoxCeleb1-B.* **Pos.***: # of positive trials;* **Neg.***: # of negative trials;* **cl.***: Cleaned version;* **Cross-lingual***: Whether the test set is constructed in consideration of bilingual scenario.*

Test set	VoxCeleb1 cl.	VoxCeleb1-E cl.	VoxCeleb1-H cl.	VoxSRC 2020 Val	VoxSRC 2021 Val	VoxCeleb1-B
# of trials	37,611	579,818	550,894	263,486	60,000	808,574
(Pos. / Neg.)	(18,802 / 18,809)	(289,921 / 289,897)	(275,488 / 275,406)	(131,743 / 131,743)	(29,969 / 30,031)	(404,287 / 404,287)
Cross-lingual	✗	✗	✗	✗	✓	✓



(a) *Language discriminator training*



(b) *Speaker embedding training*

Figure 1: *Overview of the training strategy. The coloured parts of the network are updated during each training procedure. Note that the Gradient Reversal Layer (GRL) is only activated during speaker embedding training procedure.* $\boldsymbol{x_i}$*: input mel-spectrogram;* $\boldsymbol{z_i}$*: frame-level embeddings;* $\boldsymbol{e_i}^{\boldsymbol{S}}$*: speaker embedding vector;* $\boldsymbol{e_i}^{\boldsymbol{L}}$*: language feature vector.*

recognition model pre-trained on VoxLingua107 [34] dataset. To the best of our knowledge, we are the first to perform intra-domain disentangled representation learning using only pseudo-labels.

2. Bilingual speaker recognition test set

We publish a large-scale bilingual speaker recognition evaluation protocol derived from VoxCeleb1 dataset [16], which is one of the widespread benchmark evaluation datasets in the recent speaker recognition field. Most of existing evaluation sets do not focus on the bilingual scenarios. Our speaker recognition evaluation set contains 808,574 trials in total. Half of the trials are intra-speaker cross-lingual and the remaining trials are inter-speaker monolingual.

2.1. Obtaining language labels

To simulate the bilingual scenarios with VoxCeleb1 dataset, it is necessary to have the language labels of the utterances in the test set. We utilise language annotations of VoxCeleb1 dataset from VoxSRC 2021 [18] which are manually checked after obtaining language pseudo-labels of the utterances by using a Spoken Language Recognition (SLR) model pre-trained on VoxLingua107 [34] dataset. VoxLingua107 dataset contains 6,628 hours of speech that are divided into 107 languages.

Assuming that a single speaker speaks only one language in a video, one audio sample is randomly sampled for each video. 15 languages including English, French, Hindi, German, Spanish, Italian, Afrikaans, Portuguese, Dutch, Korean, Urdu, Swedish, Russian, Chinese, and Arabic are annotated by annotators of various nationalities. Out of 153,516 utterances in the VoxCeleb1 dataset, 883 utterances, whose language could not be recognised by annotators, have been excluded from the proposed evaluation list.

2.2. `VoxCeleb1-B` Evaluation list

Speaker verification evaluation protocol consists of *positive* and *negative* trials. Each trial involves an enrollment utterance and a test utterance. The trial type is decided based on whether the enrollment and the test utterances have the same speaker identity. To evaluate the robustness of speaker recognition models in the bilingual scenarios, we propose an evaluation protocol named VoxCeleb1-B, which simulates language-mismatch scenarios with a large amount of cross-lingual trials. Using the speaker and language annotations, we generate 404,287 intra-speaker cross-lingual trials and inter-speaker monolingual trials each. The number of speakers for each language and the number of samples per speaker are limited to 1,000 and 15, respectively, to avoid bias towards more frequent languages.

Table 1 shows the statistics of existing evaluation lists derived from VoxCeleb1, and the proposed VoxCeleb1-B. The three original VoxCeleb1 test sets and the VoxSRC 2020 [35] validation set are expected to contain very few cross-lingual positive trials, whereas the VoxSRC 2021 [18] contains some cross-lingual trials. VoxCeleb1-B is explicitly designed to contain a large number of cross-lingual trials.

3. Language-disentangled learning

In this section, we describe the proposed language-disentangled representation learning strategy. Our training framework is inspired by [15, 36, 37] and summarised in Figure 1. The network consists of a speaker embedding network that includes a speaker feature extractor and a speaker embedding layer, a speaker classifier, and a language classifier. The speaker embedding network follows the existing speaker models [38, 39] while the language classifier is attached for the purpose of a language discriminator.

The speaker embedding network produces frame-level embeddings z_i from the input mel-spectrogram data $x_i \in \mathcal{R}^{T \times F}$ $(1 \leq i \leq N)$, where T, F and N are the number of frames, frequency bins, and the size of mini-batch, respectively. To derive an utterance-level vector ${e_i}^S$ from the frame-level embeddings z_i, we adopt Attentive Pooling Layer (APL) which includes self-attentive pooling (SAP) [40] or attentive statistics pooling (ASP) [41] as an speaker embedding layer.

The speaker embedding vector ${e_i}^S$ is passed as an input feature to both the speaker classifier and the language classifier, which consist of one and three fully-connected layers, respectively. We obtain the language feature vector, ${e_i}^L$, from the output of the second fully-connected layer in the language classifier. For the language classifier, the GRL is placed at the front of the language classifier and is activated in speaker embedding training step.

The training process of our framework alternates between two phases for the data from the same mini-batch: (1) language discriminator training, and (2) speaker embedding training. In the first phase, we train the language discriminator to recognise the spoken language from ${e_i}^S$. In the second phase, the speaker recognition network is trained to classify speakers, while intentionally trained to poorly recognise spoken languages.

3.1. Language discriminator training

In this step, we train the language classifier, while freezing the speaker recognition network. This approach can be interpreted to train the language recognition for the latest state of the speaker representation vector ${e_i}^S$ that has been extracted by the speaker embedding layer. The objective function L_{lang} of the language classifier is a categorical cross-entropy loss. In Figure 1a, the parts of the network coloured in blue are optimised by L_{lang}.

3.2. Speaker embedding training

In this step, we train the speaker recognition network with language-disentangled representation learning. The language classifier's parameters are not updated at this stage. The total loss function to train the language-disentangled speaker recognition model can be formulated as follows.

$$L_{total} = L_{spk} + L_{de} \tag{1}$$

where L_{spk} is an objective function for the speaker recognition and L_{de} is an objective function of the disentangled representation learning. For L_{spk}, we can utilise objective functions such as softmax loss, prototypical loss, and contrastive loss, which have been employed in the previous works [38, 39]. For prototypical loss and contrastive loss, we exclude the speaker classifier since these losses are directly derived from the speaker embedding vectors rather than speaker logits. For L_{de}, we can apply objective functions from two types of learning methods, namely adversarial learning-based method and metric learning-based method. In this work, we select gradient reversal layer as an adversarial learning-based method, and metric learning-based methods include cosine similarity minimisation and mean absolute Pearson's correlation minimisation. The details of each method are as follows.

Gradient Reversal Layer (GRL). Gradient reversal layer inverts the gradient value of target loss function to opposite sign for disturbing the convergence of the target loss function. In our work, the target loss function is L_{lang}.

Cosine similarity (COS) minimisation. This method minimises cosine similarity between speaker embedding vector ${e_i}^S$ and language feature vector ${e_i}^L$.

Mean Absolute Pearson's Correlation (MAPC) minimisation. This method minimises mean absolute Pearson's correlation [29] between speaker embedding vector and language feature vector. In our work, L_{corr} can be formulated as follows.

$$L_{corr} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{F}\frac{|\mathrm{Cov}(e_{i,j}^S, e_{i,j}^L)|}{\sigma(e_{i,j}^S)\cdot\sigma(e_{i,j}^L)} \tag{2}$$

where $\mathrm{Cov}(\cdot)$ is the covariance and $\sigma(\cdot)$ is the standard deviation. F is the dimensionality of the embedding vector e_i.

Ours. We propose an effective disentangled representation learning that consists of GRL and MAPC minimisation. The total loss function L_{total} of our method can be formulated as follows.

$$L_{total} = L_{spk} + L_{corr} + \lambda L_{lang} \tag{3}$$

where λ is a weight value for summation. In Figure 1b, the parts of the network coloured in yellow are optimised by L_{total}.

4. Experiments

4.1. Input representations and model architecture

For the input representation of the neural network, we use log-mel spectrograms that are extracted with a hamming window, 25ms window size and 10ms stride size.

We focus on demonstrating the effectiveness of the proposed learning strategy and its compatibility with previous models. Thus, we employ two existing variants [38, 39] of the 34-layer residual network, and rename each variant as ResNet-S [38] and ResNet-L [39]. ResNet-S uses the SAP [40] and the angular prototypical loss, and ResNet-L uses the ASP [41] and the angular prototypical loss combined with the softmax loss, which is in line with the original papers. The output size of each classifier is equal to the number of each task's classes.

4.2. Disentangled representation learning method

We evaluate various disentangled representation learning strategies in terms of separating irrelevant information from intra-domain speaker representation rather than domain adaptation in cross-domain. We perform extensive experiments on various disentangled representation learning strategies including adversarial learning-based method using the GRL, metric learning-based method using COS minimisation or MAPC minimisation, and the proposed method, which is the combination of GRL and MAPC minimisation.

4.3. Implementation details

Datasets. We use the development partition with 5,994 speakers of the VoxCeleb2 [17] as the training dataset. In order to learn language-disentangled representation, we use the language pseudo-labels of VoxCeleb2 dataset extracted from an SLR model pre-trained on VoxLingua107 dataset. For evaluation, we use the three original test sets based on VoxCeleb1 [16], the VoxSRC validation sets [18, 35] and VoxCeleb1-B, which is the proposed large-scale bilingual speaking evaluation set.

Training. Our implementation is based on the PyTorch framework [42]. We use the Adam Optimizer [43] with initial learning rate of 0.001 decreasing by 3% every epoch. All experiments are performed on a single NVIDIA A5000 GPU with 24GB memory. We use the batch size of 500 and 300 for ResNet-S and ResNet-L, respectively. The training takes around 3 days. The λ value is set to 0.5.

4.4. Evaluation protocol

We report Equal Error Rate (EER) where the False Rejection Rate (FRR) and the False Alarm Rate (FAR) are equal, and minimum Detection Cost Function (minDCF) [44] which is a weighted sum of FRR and FAR. For each trial, we sample each utterance into ten 4-second segments and compute similarities between all possible combinations of segment pairs. We use the mean of the similarities as a score of the trial. This evaluation protocol is in line with that from [15, 17, 37, 38].

5. Results

The experimental results are summarised in Table 2. Specifically, Table 2a reports the results on the test sets that mainly consider bilingual scenario, while Table 2b contains the results on existing test sets that do not take the bilingual speakers into account during their construction phases. We train all models 3 times with different random seeds, and report the mean and the standard deviation.

Bilingual scenario in speaker recognition. As shown in Table 2a, most of the models show poor performance on the VoxSRC 2021 validation set and VoxCeleb1-B. In particular, all baselines show the lowest performance on VoxCeleb1-B which is composed entirely of bilingual trials. This implies that bilingual scenario is one of the most demanding challenges in the speaker recognition field.

Intra-domain disentangled representation learning. When compared to the baseline of ResNet-L, the use of GRL shows notable

Table 2: *Equal Error Rates (EER) and minimum Detection Cost Function (minDCF) on (a) VoxSRC 2021 validation set, VoxCeleb1-B, (b) VoxCeleb1 test sets, and VoxSRC 2020 validation set. Accuracy of Spoken Language Recognition (SLR Acc.) is computed on VoxCeleb1-B. Lower SLR means less language information. All experiments except for spoken language recognition are repeated three times, and we report the mean and the standard deviation.* **GRL***: Gradient reversal layer;* **COS min.***: Cosine similarity minimisation;* **MAPC min.***: Mean absolute Pearson's correlation minimisation;* **Ours***: Combination of GRL and MAPC minimisation.*

Model	VoxSRC 2021 Val		VoxCeleb1-B		
	EER (%)	minDCF	EER (%)	minDCF	SLR Acc. (%)
ResNet-S [38]	9.22 ± 0.15	0.503 ± 0.007	9.69 ± 0.14	0.617 ± 0.007	87.2
+ *GRL*	10.27 ± 0.88	0.541 ± 0.036	10.39 ± 1.24	0.598 ± 0.055	87.2
+ *COS min.*	9.33 ± 0.18	0.505 ± 0.002	10.21 ± 0.22	0.614 ± 0.012	87.0
+ *MAPC min.*	8.85 ± 0.12	0.486 ± 0.005	9.85 ± 0.15	0.594 ± 0.011	86.8
Ours	$\mathbf{8.35 \pm 0.05}$	$\mathbf{0.461 \pm 0.002}$	$\mathbf{8.25 \pm 0.06}$	$\mathbf{0.506 \pm 0.002}$	82.9
ResNet-L [39]	5.16 ± 0.08	0.308 ± 0.010	5.96 ± 0.23	0.397 ± 0.016	88.3
+ *GRL*	4.53 ± 0.25	0.263 ± 0.008	3.98 ± 0.27	0.268 ± 0.020	72.1
+ *COS min.*	5.21 ± 0.18	0.317 ± 0.015	5.99 ± 0.35	0.423 ± 0.016	88.8
+ *MAPC min.*	5.23 ± 0.01	0.311 ± 0.009	5.93 ± 0.19	0.411 ± 0.018	88.2
Ours	$\mathbf{4.22 \pm 0.03}$	$\mathbf{0.246 \pm 0.006}$	$\mathbf{3.69 \pm 0.12}$	$\mathbf{0.254 \pm 0.010}$	80.1

(a) *Results on VoxSRC 2021 validation set and VoxCeleb1-B.*

Model	VoxCeleb1 cl.		VoxCeleb1-E cl.		VoxCeleb1-H cl.		VoxSRC 2020 Val	
	EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF
ResNet-S	2.24 ± 0.13	0.174 ± 0.005	2.43 ± 0.04	0.175 ± 0.003	4.74 ± 0.07	0.299 ± 0.005	6.91 ± 0.07	0.393 ± 0.004
+ *GRL*	2.98 ± 0.14	0.210 ± 0.012	3.12 ± 0.19	0.222 ± 0.013	5.72 ± 0.37	0.351 ± 0.020	8.07 ± 0.47	0.457 ± 0.025
+ *COS min.*	$\mathbf{2.13 \pm 0.08}$	0.164 ± 0.006	2.45 ± 0.06	0.178 ± 0.004	4.78 ± 0.12	0.303 ± 0.006	6.83 ± 0.06	0.388 ± 0.004
+ *MAPC min.*	2.16 ± 0.07	$\mathbf{0.157 \pm 0.003}$	$\mathbf{2.33 \pm 0.01}$	$\mathbf{0.166 \pm 0.001}$	$\mathbf{4.48 \pm 0.02}$	$\mathbf{0.284 \pm 0.003}$	6.61 ± 0.04	$\mathbf{0.373 \pm 0.003}$
Ours	2.15 ± 0.01	0.172 ± 0.001	2.42 ± 0.01	0.171 ± 0.000	4.49 ± 0.02	$\mathbf{0.284 \pm 0.001}$	$\mathbf{6.54 \pm 0.02}$	0.378 ± 0.001
ResNet-L	1.17 ± 0.00	0.083 ± 0.003	1.30 ± 0.01	0.091 ± 0.001	2.58 ± 0.02	0.164 ± 0.001	4.06 ± 0.02	0.231 ± 0.005
+ *GRL*	1.22 ± 0.04	0.088 ± 0.009	1.34 ± 0.04	0.096 ± 0.002	2.58 ± 0.02	0.166 ± 0.001	4.13 ± 0.04	0.230 ± 0.003
+ *COS min.*	1.11 ± 0.01	0.084 ± 0.007	1.25 ± 0.03	0.091 ± 0.004	2.52 ± 0.02	0.164 ± 0.002	3.99 ± 0.03	0.228 ± 0.002
+ *MAPC min.*	1.10 ± 0.02	$\mathbf{0.079 \pm 0.001}$	$\mathbf{1.24 \pm 0.02}$	$\mathbf{0.088 \pm 0.002}$	2.48 ± 0.04	0.160 ± 0.002	3.99 ± 0.04	0.224 ± 0.001
Ours	$\mathbf{0.99 \pm 0.05}$	$\mathbf{0.079 \pm 0.004}$	1.25 ± 0.01	$\mathbf{0.088 \pm 0.000}$	$\mathbf{2.42 \pm 0.04}$	$\mathbf{0.154 \pm 0.003}$	$\mathbf{3.91 \pm 0.06}$	$\mathbf{0.220 \pm 0.001}$

(b) *Results on cleaned version of VoxCeleb1 test sets and VoxSRC 2020 validation set.*

performance improvements of 33% and 12% in VoxCeleb1-B and VoxSRC2021 validation set, respectively, while the performance degrades on the other evaluation sets. Furthermore, ResNet-S with GRL exhibits the lowest performance and the highest standard deviation on every evaluation set including VoxCeleb1-B. This highlights the drawback of GRL, which tends to induce unstable training despite its effectiveness in removing the language information from the speaker representation.

We observe out that the metric learning-based methods, COS and MAPC minimisation, outperform the baselines on VoxCeleb1 test sets and VoxSRC 2020 validation set. However, we observe no performance improvement on VoxCeleb1-B. This suggests that the disentangled representation learning method which utilises metric learning performs a role of regularisation, but fails to isolate the language information from the speaker representation. As a result, we verify that existing disentangled representation learning strategies applied to cross-domain tasks do not guarantee generalisation to intra-domain tasks such as bilingual scenarios.

The proposed learning method on ResNet-L outperforms the baselines and all existing methods on most of evaluation sets except VoxCeleb1-E test set. In the case of ResNet-S, our method shows the best performance on VoxSRC validation sets and VoxCeleb1-B while remaining effective on the other evaluation sets. Especially, ResNet-S and ResNet-L exhibit significant performance improvements by 15% and 38% on VoxCeleb1-B, respectively. This demonstrates that the proposed learning strategy overcomes the limitations of existing disentangled representation learning methods and facilitates robust language-disentangled speaker representation learning.

The use of pseudo-labels. Training of all experiments are performed with language pseudo-labels of VoxCeleb2. Nonetheless, the proposed learning strategy works successfully, showing significant performance improvements in bilingual scenarios. This indicates that it can be cost-effective to use pseudo-labels of specific factor in speaker recognition, which can be extended to other factors of variation that must be disentangled from speaker embeddings.

Language-disentangled speaker representation. To verify whether the language information is separated from the speaker representation, we evaluate a spoken language recognition model trained from scratch with the speaker embedding vector extracted from each model as input data. The structure of the spoken language recognition model is the same as the language classifier described in Section 3. As shown in Table 2a (SLR Acc.), our method shows lower spoken language recognition performance than the baselines of two models. This highlights that the proposed method successfully removes the language information from the speaker representation.

6. Conclusion

We have developed strategies to train speaker embeddings that are robust to bilingual speaking scenarios, and proposed an evaluation protocol that takes bilingual speakers into account. Our large-scale evaluation protocol is designed to analyse speaker recognition performance under bilingual scenarios, and we make this evaluation set publicly available. We also propose a new learning strategy to resolve the bilingual problem. Our learning strategy disentangles language information from the speaker representation in order to make the embeddings robust to cross-lingual trials. Our proposed learning strategy shows significant performance improvements under bilingual scenarios, while remaining effective on existing test sets.

7. Acknowledgements

We would like to thank Icksang Han and Bong-Jin Lee for helpful comments.

8. References

[1] G. Vince, "The amazing benefits of being bilingual," *BBC*, 2016.

[2] B. Lee and D. V. L. Sidtis, "The bilingual voice: Vocal characteristics when speaking two languages across speech tasks," *Speech, Language and Hearing*, vol. 20, no. 3, pp. 174–185, 2017.

[3] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: the mixer 3, 4 and 5 corpora." in *Proc. Interspeech*. Citeseer, 2007, pp. 950–953.

[4] M. Akbacak and J. H. Hansen, "Language normalization for bilingual speaker recognition systems," in *Proc. ICASSP*, vol. 4. IEEE, 2007, pp. IV–257.

[5] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot *et al.*, "Promoting robustness for speaker modeling in the community: the prism evaluation set," in *Proceedings of NIST 2011 workshop*. Citeseer, 2011, pp. 1–7.

[6] D. Reynolds, E. Singer, S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. Greenberg, L. Mason, and J. Hernandez-Cordero, "The 2016 nist speaker recognition evaluation," MIT Lincoln Laboratory Lexington United States, Tech. Rep., 2017.

[7] P. Matejka, O. Novotnỳ, O. Plchot, L. Burget, M. D. Sánchez, and J. Cernockỳ, "Analysis of score normalization in multilingual speaker recognition." in *Proc. Interspeech*, 2017, pp. 1567–1571.

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

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

[10] J.-W. Jung, H.-S. Heo, I.-H. Yang, H.-J. Shim, and H.-J. Yu, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *Proc. ICASSP*. IEEE, 2018, pp. 5349–5353.

[11] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *IEEE Spoken Language Technology workshop*. IEEE, 2018, pp. 1021–1028.

[12] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. ICASSP*. IEEE, 2019, pp. 5796–5800.

[13] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, vol. 30, 2017.

[14] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *NeurIPS*, vol. 33, 2020, pp. 18 661–18 673.

[15] J. Kang, J. Huh, H. S. Heo, and J. S. Chung, "Augmentation adversarial training for self-supervised speaker representation learning," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[16] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

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

[18] A. Brown, J. Huh, J. S. Chung, A. Nagrani, and A. Zisserman, "Voxsrc 2021: The third VoxCeleb speaker recognition challenge," *arXiv preprint arXiv:2201.04583*, 2022.

[19] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *IEEE Automatic Speech Recognition and Understanding workshop*. IEEE, 2019.

[20] C. Luu, P. Bell, and S. Renals, "Leveraging Speaker Attribute Information Using Multi Task Learning for Speaker Verification and Diarization," in *Proc. Interspeech*, 2021, pp. 491–495.

[21] C. Luu, S. Renals, and P. Bell, "Investigating the contribution of speaker attributes to speaker separability using disentangled speaker representations," in *Proc. Interspeech*, 2022.

[22] J. Williams and S. King, "Disentangling style factors from speaker representations." in *Proc. Interspeech*, 2019, pp. 3945–3949.

[23] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *Proc. ICASSP*. IEEE, 2020, pp. 7169–7173.

[24] S. Maiti, E. Marchi, and A. Conkie, "Generating multilingual voices using speaker space translation based on bilingual speaker data," in *Proc. ICASSP*. IEEE, 2020, pp. 7624–7628.

[25] A. Misra and J. H. Hansen, "Spoken language mismatch in speaker verification: An investigation with nist-sre and crss bi-ling corpora," in *IEEE Spoken Language Technology workshop*. IEEE, 2014, pp. 372–377.

[26] Z. Meng, Y. Zhao, J. Li, and Y. Gong, "Adversarial speaker verification," in *Proc. ICASSP*. IEEE, 2019, pp. 6216–6220.

[27] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *Proc. ICASSP*. IEEE, 2019, pp. 5816–5820.

[28] D. Xin, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, "Cross-lingual speaker adaptation using domain adaptation and speaker consistency loss for text-to-speech synthesis." in *Proc. Interspeech*, 2021, pp. 1614–1618.

[29] W. H. Kang, S. H. Mun, M. H. Han, and N. S. Kim, "Disentangled speaker and nuisance attribute embedding for robust speaker verification," *IEEE Access*, vol. 8, pp. 141 838–141 849, 2020.

[30] Y. Kwon, S. W. Chung, and H. G. Kang, "Intra-class variation reduction of speaker representation in disentanglement framework," in *Proc. Interspeech*, vol. 2020, 2020, pp. 3231–3235.

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

[32] L. Yi and M.-W. Mak, "Disentangled speaker embedding for robust speaker verification," in *Proc. ICASSP*. IEEE, 2022, pp. 7662–7666.

[33] S. H. Mun, M. H. Han, M. Kim, D. Lee, and N. S. Kim, "Disentangled speaker representation learning via mutual information minimization," in *Proc. APSIPA ASC*. IEEE, 2022.

[34] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *IEEE Spoken Language Technology workshop*. IEEE, 2021, pp. 652–658.

[35] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "Voxsrc 2020: The second VoxCeleb speaker recognition challenge," *arXiv preprint arXiv:2012.06867*, 2020.

[36] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. CVPR*, 2017, pp. 7167–7176.

[37] J. S. Chung, J. Huh, and S. Mun, "Delving into VoxCeleb: environment invariant speaker recognition," in *Proc. Speaker Odyssey*, 2020.

[38] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In Defence of Metric Learning for Speaker Recognition," in *Proc. Interspeech*, 2020, pp. 2977–2981.

[39] Y. Kwon, H.-S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: lessons from voxsrc 2020," in *Proc. ICASSP*. IEEE, 2021, pp. 5809–5813.

[40] W. Cai, J. Chen, and M. Li, "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System," in *Proc. Speaker Odyssey*, 2018, pp. 74–81.

[41] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, vol. 32, 2019.

[43] D. P. Kingma, J. Ba, Y. Bengio, and Y. LeCun, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[44] O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, "The 2018 nist speaker recognition evaluation," in *Proc. Interspeech*, 2019.