

# SPEAKER VERIFICATION USING END-TO-END ADVERSARIAL LANGUAGE ADAPTATION

Johan Rohdin<sup>1\*</sup>, Themis Stafylakis<sup>2\*</sup>, Anna Silnova<sup>1</sup>, Hossein Zeinali<sup>1</sup>, Lukáš Burget<sup>1</sup>, Oldřich Plchot<sup>1</sup>

<sup>1</sup>Brno University of Technology, Faculty of Information Technology, IT4I Centre of Excellence, Czechia  
{rohdin, isilnova, zeinali, burget, iplchot}@fit.vutbr.cz

<sup>2</sup>Omilia - Conversational Intelligence, Athens, Greece  
tstafylakis@omilia.com

## ABSTRACT

In this paper we investigate the use of adversarial domain adaptation for addressing the problem of language mismatch between speaker recognition corpora. In the context of speaker verification, adversarial domain adaptation methods aim at minimizing certain divergences between the distribution that the utterance-level features follow (i.e. speaker embeddings) when drawn from source and target domains (i.e. languages), while preserving their capacity in recognizing speakers. Neural architectures for extracting utterance-level representations enable us to apply adversarial adaptation methods in an end-to-end fashion and train the network jointly with the standard cross-entropy loss. We examine several configurations, such as the use of (pseudo-)labels on the target domain as well as domain labels in the feature extractor, and we demonstrate the effectiveness of our method on the challenging NIST SRE16 and SRE18 benchmarks.

**Index Terms**— Speaker recognition, domain adaptation

## 1. INTRODUCTION

The need for domain adaptation (DA) arises in cases when the target domain data is insufficient (and possibly unlabeled) for training a model from scratch and therefore source domain data (assumed labeled and sufficient for training a model) should be leveraged as well. The core idea behind DA is that the knowledge distilled from the source domain can be transferred to the target domain, despite the differences in the marginal distributions of the two domains. Conventional approaches to DA, such as fine-tuning a source domain model to the target domain data may fail in many settings due to the target data being weakly-labeled or even unlabeled.

\*Equal Contribution.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie and it is co-financed by the South Moravian Region under grant agreement No. 665860. The project was also supported by the Czech Science Foundation under project No. GJ17-23870Y.

DA methods for speaker verification are of particular interest, as for many real-world applications large amounts of target domain labeled data are rarely available. Hence, for training state-of-the-art models which require several thousand of training utterances, one should resort to large out-of-domain corpora and use the small and possibly unlabeled target domain datasets for language, channel or other types of adaptation. In order to promote further research in DA methods, MIT-LL and NIST has organized 3 evaluations (namely the MIT-LL DA challenge, DAC-2013, NIST SRE16, and the recent NIST SRE18) with the two latter focused primarily on language adaptation. Several DA methods were introduced as part of those evaluations, the majority of which approach the problem as a transformation of fixed utterance-level representations, such as i-vectors.

In this article we examine the use of the recently emerged adversarial DA methods. Adversarial DA methods employ Generative Adversarial Networks (GANs) as a means to reduce the mismatch between source and target domains [1]. Different from [2], where an adversarial architecture is proposed for i-vector adaptation and tested for mildly mismatched domains (DAC-2013, between Switchboard and NIST data both of which telephone data and English) (a) we propose an end-to-end DA method by adding the adversarial loss to the cross-entropy loss of the x-vector architecture, and (b) we evaluate our method on the challenging task of language adaptation. Moreover, we use Wasserstein GANs, a recently proposed version of GANs which addresses the vanishing gradient problem of GANs by replacing the domain discrepancy measure with Wasserstein distance [3]. We also explore the uses of speaker labels in the adaptation data, even in the form of pseudo-labels for cases where the set is unlabeled, and we show that they are very helpful in attaining good performance. Finally, we examine the use of domain labels, which we concatenate to the layers of the network to learn domain-dependent transforms using a single network. To the best of our knowledge, we are the first to utilize domain labels in such a way.

## 2. DOMAIN ADAPTATION IN SPEAKER RECOGNITION

During the past few years, several DA methods for speaker recognition have been proposed. In the case of unsupervised adaptation several methods apply a clustering algorithm in order to estimate speaker labels, with which a target-domain PLDA model is trained, while interpolation between the parameters of the source and target-domain PLDA models is applied to obtain the adapted PLDA [4, 5, 6]. The standard speaker recognition recipe in the Kaldi toolkit utilizes a simpler method for unsupervised adaptation that does not require clustering [7]. This method aims at adjusting the covariance matrices of the PLDA model so that its total covariance better matches the total covariance of the adaptation data. An alternative approach is to compensate for dataset shift in the i-vector space by modelling the subspace of dataset shift and removing those direction from the i-vectors [8]. Other approaches do not attempt to cluster the utterances and perform DA simply by matching first and second order statistics of the i-vectors between source and target domains [9].

Closer to the spirit of our work, two methods based on domain-adversarial adaptation and maximum mean discrepancy were recently introduced. In [2], a domain-adversarial neural networks (DANN) is employed in order to transform i-vectors to a domain-invariant representation space. The authors follow the recipe introduced in [1] without using or estimating speaker labels for the target domain training data. They evaluate their method on DAC 2013 and demonstrate significant gains over other DA methods. In [10], DA is performed using maximum mean discrepancy (MMD) as a means to reduce the mismatch between the two distribution. The main differences compared to the DANN-based architecture in [2] are (a) the use of MMD which makes training easier compared to GANs, (b) the use of reconstruction loss instead of cross entropy in the main network (i.e. an autoencoder architecture instead of a classifier over speakers [11]), and (c) the application field, which is the language adaptation task of NIST SRE16 instead of DAC 2013. The method yields slightly better results compared to inter-dataset variability compensation [8].

## 3. ADVERSARIAL ADAPTATION ALGORITHM

### 3.1. Notation and Wasserstein distance

We assume a labeled source dataset  $X^s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$  from the source domain  $D_s$ , and target dataset  $X^t = \{(x_i^t, y_i^t)\}_{i=1}^{n_t}$  from the target domain  $D_t$ , where  $x \in \mathbb{R}^{m,\cdot}$  denotes utterances and the labels  $y_i^t$  may be given, estimated (e.g. using clustering) or not used at all. The two domains (i.e. languages) have different marginal data distributions,  $P_{x^s}$  and  $P_{x^t}$  respectively. The embedding extractor, which is a standard TDNN x-vector architecture up to the embedding layer

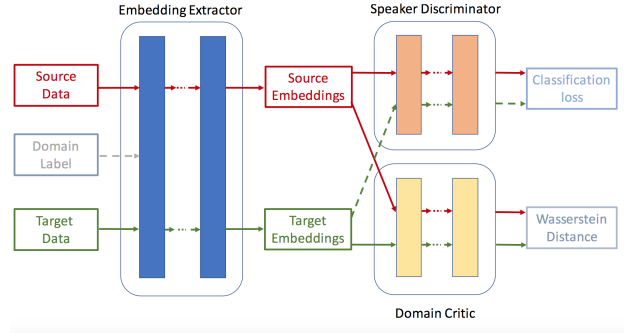


Fig. 1. Block-diagram of the architecture.

implements a function  $f_g : \mathbb{R}^{m,\cdot} \mapsto \mathbb{R}^d$  parametrized by  $\theta_g$ , where  $d$  is the size of the embedding. The additional structure, useful only during training is called the domain critic and is a feed-forward neural network implementing a function  $f_w : \mathbb{R}^d \mapsto \mathbb{R}$  parametrized by  $\theta_w$ . The Wasserstein distance between two representation distributions  $P_{x^s}$  and  $P_{x^t}$ , where  $h^s = f_g(x^s)$  and  $h^t = f_g(x^t)$  is approximated by

$$\mathcal{L}_{wd}(x^s, x^t) = \frac{1}{n^s} \sum_{x^s \in X^s} f_w(f_g(x^s)) - \frac{1}{n^t} \sum_{x^t \in X^t} f_w(f_g(x^t)). \quad (1)$$

As proposed in [3], an improved method (compared to [12]) for constraining  $g_w$  to be 1-Lipschitz function (a necessary conditions so that  $\mathcal{L}_{wd}$  is an approximation of the Wasserstein distance) is to introduce a gradient penalty loss

$$\mathcal{L}_{grad}(\hat{h}) = \left( \left\| \nabla_{\hat{h}} f_w(\hat{h}) \right\|_2 - 1 \right)^2, \quad (2)$$

where  $\hat{h}$  is a set of features created by randomly pairing and linearly combining features from  $h^s$  and  $h^t$  [3].

The speaker discriminator (i.e. the part of the x-vector network after the embedding layer) implements a function  $f_c : \mathbb{R}^d \mapsto \mathbb{R}^l$ , i.e. it maps the embeddings  $h$  to the space of posterior probabilities over training speakers (either from source or from target domain) and it is parametrized by  $\theta_c$  (separate linear and softmax layers are assumed for source and target domains). The classification loss  $\mathcal{L}_c(x, y)$  is the standard cross-entropy over speakers. The architecture during training is illustrated in Fig. 1.

### 3.2. Training algorithm

The training algorithm is given in Algorithm 1 (see [13] for more information). As we observe, the critic  $\theta_w$  tries to maximize  $\mathcal{L}_{wd}$ , i.e. to approximate the Wasserstein distance between the two domains while the feature extractor  $\theta_g$  tries to minimize it, yielding the usual minimax optimization problem of GANs. In the inner loop of the algorithm, a set of points  $\hat{h}$  are randomly chosen as linear combinations between

---

**Algorithm 1** Domain adaptation algorithm

---

```
1: Initialize feature extractor, domain critic, speaker dis-
   crimator  $\theta_g, \theta_w, \theta_c$ .
2: if supervised  $d \leftarrow (s, t)$  else  $d \leftarrow s$  end
3: repeat
4:   Sample minibatch  $\{(x_i^s, y_i^s)\}, \{(x_i^t, y_i^t)\}$ 
5:   for  $t = 1, \dots, n$  do
6:      $h^s \leftarrow f_g(x^s), h^t \leftarrow f_g(x^t)$ 
7:     Sample  $h$  as the random points along straight lines
       between  $h^s$  and  $h^t$  pairs.
8:      $\hat{h} \leftarrow \{h^s, h^t, h\}$ 
9:      $\theta_w \leftarrow \theta_w + \alpha_1 \nabla_{\theta_w} [\mathcal{L}_{wd}(x^s, x^t) - \gamma \mathcal{L}_{grad}(\hat{h})]$ 
10:  end for
11:   $\theta_c \leftarrow \theta_c - \alpha_2 \nabla_{\theta_c} \mathcal{L}_c(x^d, y^d)$ 
12:   $\theta_g \leftarrow \theta_g - \alpha_2 \nabla_{\theta_g} [\mathcal{L}_c(x^d, y^d) + \delta \mathcal{L}_{wd}(x^s, x^t)]$ 
13: until  $\theta_g, \theta_w, \theta_c$  converge.
```

---

randomly paired  $h^s$  and  $h^t$ , on which the gradient penalization is applied, constraining  $f_w$  to be a 1-Lipschitz function [3]. Finally, when labels are used in the target domain, the classification loss is backpropagated for both sets.

### 3.3. Architecture and implementation details

We use Tensorflow [14] for implementing the adversarial adaptation. We follow the standard Kaldi x-vector architecture [7], i.e. 5 TDNN layers with ReLU activation functions followed by batch normalization, then a pooling layer that accumulates mean and standard deviations, then two feed-forward layers with ReLU and batch normalization, then finally a softmax layer for classifying speakers. For the critic network we use two feed-forward layers with 512 units and leaky ReLU activation functions. The critic network takes the x-vector, i.e. the output of the first affine layer after pooling, as input and returns a scalar as output. The domain label is passed to the feature extractor as a binary variable which is concatenated to the input of every affine layer (0 corresponds to the source domain). This is equivalent to having domain-dependent biases, enabling the network to learn domain-dependent transforms. Based on light tuning to make the training stable, we set the parameters of the adversarial training to  $\gamma = 10.0$ ,  $\delta = 0.1$ ,  $\alpha_1 = 0.001$ ,  $\alpha_2 = 1.0$ , and  $n = 10$ . For supervised adaptation, we add a second softmax layer to the x-vector network, i.e. the source- and target-domain classifiers share all model parameters except those of their softmax layer. In order to better balance the source- and target-domain classification losses, we normalize them with the logarithm of their number of classes so that the loss of random prediction is approximately equal to one. After that, we set the weight for the target domain classification loss to 0.2 and the weight for the source domain classification loss to 0.8. In the experiments, we use minibatches containing 150 segments of the target domain data and 150 segments of the

(labeled) source domain data. The lengths of the segments are 2-4s. We use stochastic gradient descent, starting with a learning rate of 1.0 which we then half every 5 epochs, where an epoch is defined to be 400 minibatches. We stop the training after 85 epochs. During the first 3 epochs, we trained only the critic and the network for source domain classification.

## 4. EXPERIMENTS

We conduct experiments on two databases, the NIST SRE 2018 *cmn2* evaluation set using the unlabeled *cmn2* development set for adaptation, and the NIST SRE16 evaluation set [15] using the unlabeled major data set for adaptation. It should be noted that the SRE16 evaluation data as well as the unlabeled major data contains utterances from two languages (Cantonese and Tagalog) which is not ideal for our proposed methods, since language labels are not provided (and are not estimated by our algorithm). Hence, we treat two distinct languages as one, which is clearly suboptimal. Further, for the SRE18 unlabeled development set, we have access to the telephone numbers which should help the supervised training compared to just using utterance IDs. The adaptation data is augmented similarly to the training data [16], i.e. with babble, noise, music and reverberated versions of the utterances.

The baseline x-vector model is trained by the Kaldi toolkit [17] using the Kaldi SRE16 x-vector recipe [16] but with additional training data from voxceleb2, resulting in 12170 training speakers. We apply the adversarial DA on this model rather than training a model with adversarial DA from the beginning. We used an identical backend to the one in the Kaldi x-vector recipe [16]. This backend involves a preprocessing step which first reduces the x-vector dimension by LDA from 512 to 150, and then applies an unusual variant of length-norm<sup>1</sup>. The backend was implemented in python based on our in-house toolkit *Pytel*. In all experiments, we center the evaluation data at the mean of the unlabeled development set instead of the mean of the PLDA training set. This can be seen as a light form of adaptation.

### 4.1. Results with adversarial adaptation

In this subsection, we evaluate the different variants of adversarial DA detailed in Section 3 with Gaussian PLDA backend. In summary, three observations can be made. First, adversarial adaptation is effective when it is combined with supervised training on the target data. Without supervised training on the target data adversarial adaptation degrades the performance. Second, including language labels as side-information to the TDNN layers helps for SRE18 but not SRE16. This is not too surprising, considering that SRE16 contains two languages which we treat as one.

---

<sup>1</sup>See <https://github.com/kaldi-asr/kaldi/blob/master/src/ivector/plda.cc>

**Table 1.** Results with adversarial adaptation. DCF refers to the average minDCF at the operating points 0.01 and 0.005. EER refers to equal error rate in %. Cantonese and Tagalog are denoted by yue and tgl respectively.

		SRE18		SRE16_all		SRE16_yue		SRE16_tgl	
		DCF	EER	DCF	EER	DCF	EER	DCF	EER
PLDA	Baseline	0.664	10.01	0.897	11.55	0.553	7.28	0.976	15.87
	sup	0.652	9.59	0.859	10.94	0.536	6.79	0.950	15.19
	adv	0.658	10.35	0.899	13.25	0.561	7.39	0.968	19.12
	adv+sup	0.630	8.89	<b>0.737</b>	9.88	0.501	5.73	<b>0.855</b>	14.18
	adv+lan+sup	<b>0.619</b>	<b>8.88</b>	0.746	<b>9.59</b>	<b>0.497</b>	<b>5.59</b>	0.880	<b>13.70</b>

## 4.2. Interaction with backend adaptation

We present result for the standard Kaldi-style unsupervised PLDA DA. This method essentially estimates the excess variance in the adaptation data and distributes a portion,  $\xi$ , of it to the PLDA between-class covariance matrix and a portion,  $\eta$ , to the PLDA within-class covariance matrix<sup>2</sup>. (Our experimentation with supervised PLDA adaptation using clustering methods were unsuccessful for both for the baseline and the adversarial DA model.) In Kaldi  $\xi = 1 - \eta$  and in the SRE16 x-vector recipe [16],  $\xi = 0.25$ . The results for these settings are shown in the first part of Table 2 (PLDA adp 1). As can be seen, adversarial DA degrades the performance. However, the Kaldi settings of  $\xi$  and  $\eta$  may not be optimal when adversarial DA is applied. For example, if the adversarial DA manages to remove between-class variability much better than within-class variability, the balance between  $\xi$  and  $\eta$  should be different. Moreover, in these experiments we use the same adaptation data for both the adversarial DA and the PLDA adaptation. After being used for adversarial DA, the adaptation data is most likely closer to the source data than what unseen (test) data is, meaning that the PLDA adaptation will not be strong enough. To mitigate this, we tune  $\xi$  and  $\eta$  on the SRE18 labeled development set (without requiring that they sum to 1). The results are shown in the second part of Table 2 (PLDA adp 2). As we observe, tuning  $\xi$  and  $\eta$  helps both the baseline and the models from adversarial training for SRE18. In terms of EER, the adversarial training now complements PLDA DA. For SRE16 using these values of  $\xi$  and  $\eta$  is worse than using the original Kaldi settings. Probably, the SRE18 development set is not similar enough to SRE16 for  $\xi$  and  $\eta$  to be properly estimated.

## 4.3. Adapting only the first layer after pooling

In the main experiment we adapted all layers of the network in an end-to-end fashion. In Table 3, we show results of adapting only the first layer after pooling. This is similar to i-vector AD approach in [2], although, here the transformation we learn is an affine dimensionality reduction of x-vectors.

Two observations can be made. First, when adapting only this layer there is no clear advantage of neither supervised adaptation nor including language labels. Second, in SRE16

**Table 2.** Results with adversarial DA and PLDA DA.

		SRE18		SRE16_all	
		DCF	EER	DCF	EER
PLDA adp 1	Baseline	0.580	9.05	0.613	8.01
	sup	<b>0.576</b>	<b>8.90</b>	<b>0.611</b>	<b>7.74</b>
	adv	0.616	9.70	0.664	9.42
	adv+sup	0.591	9.15	0.630	8.10
	adv+lan+sup	0.588	9.03	0.615	7.92
PLDA adp 2	Baseline	0.572	8.51	0.656	7.98
	sup	<b>0.567</b>	8.67	<b>0.624</b>	<b>7.66</b>
	adv	0.602	9.21	0.677	9.15
	adv+sup	0.578	8.28	0.649	8.00
	adv+lan+sup	0.576	<b>8.25</b>	0.651	8.00

**Table 3.** Results with adversarial DA after pooling.

		SRE18		SRE16_all	
		DCF	EER	DCF	EER
PLDA	Baseline	0.664	10.01	0.897	11.55
	sup	0.667	9.92	0.887	11.53
	adv	<b>0.615</b>	9.03	0.751	9.91
	adv+sup	0.629	<b>8.92</b>	0.741	9.79
	adv+lan+sup	0.620	9.01	<b>0.726</b>	<b>9.63</b>

adapting only this layer performs similar to adapting all layers. This can possibly be mitigated by some form of regularization in the earlier layers.

## 5. CONCLUSION AND FUTURE WORK

In this paper we introduced an end-to-end DA method for x-vectors based on Wasserstein GANs. We examined several configurations, especially with respect to the use of speaker and domain labels. We provided a detailed evaluation on NIST SRE16 and SRE18, and a fair comparison with state-of-the-art DA methods. The results show the effectiveness of the method in certain experiments, but also emphasize the need for further experimentation. To this end, we consider training the system from scratch with adversarial loss, applying the method to other DA tasks such as gender and channels, as well as addressing overfitting to the target domain data.

<sup>2</sup>See <https://github.com/kaldi-asr/kaldi/blob/master/src/ivector/plda.cc>

## 6. REFERENCES

- [1] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, Jan. 2016.
- [2] Qing Wang, Wei Rao, Sining Sun, Leib Xie, Eng Siong Chng, and Haizhou Li, “Unsupervised domain adaptation via domain adversarial training for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4889–4893.
- [3] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, “Improved training of wasserstein gans,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.
- [4] Daniel Garcia-Romero, Alan McCree, Stephen Shum, Niko Brummer, and Carlos Vaquero, “Unsupervised domain adaptation for i-vector speaker recognition,” in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014.
- [5] Stephen H Shum, Douglas A Reynolds, Daniel Garcia-Romero, and Alan McCree, “Unsupervised clustering approaches for domain adaptation in speaker recognition systems,” 2014.
- [6] Sergey Novoselov, Timur Pekhovsky, Konstantin Simonchik, and Andrey Shulipa, “Rbm-plda subsystem for the nist i-vector challenge,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [7] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in *Interspeech 2017*, Aug 2017.
- [8] Hagai Aronowitz, “Inter dataset variability compensation for speaker recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4002–4006.
- [9] Md Jahangir Alam, Gautam Bhattacharya, and Patrick Kenny, “Speaker verification in mismatched conditions with frustratingly easy domain adaptation,” in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 176–180.
- [10] Weiwei Lin, Man-Wai Mak, Longxin Li, and Jen-Tzung Chien, “Reducing domain mismatch by maximum mean discrepancy based autoencoders,” in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 162–167.
- [11] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey, “Adversarial autoencoders,” *arXiv preprint arXiv:1511.05644*, 2015.
- [12] Martín Arjovsky, Soumith Chintala, and Léon Bottou, “Wasserstein generative adversarial networks,” in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017, pp. 214–223.
- [13] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu, “Wasserstein distance guided representation learning for domain adaptation,” in *AAAI*. 2018, pp. 4058–4065, AAAI Press.
- [14] Martín Abadi et al., “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, Software available from tensorflow.org.
- [15] NIST, “Nist 2016 speaker recognition evaluation plan,” [https://www.nist.gov/sites/default/files/documents/2016/10/07/sre16\\_eval\\_plan\\_v1.3.pdf](https://www.nist.gov/sites/default/files/documents/2016/10/07/sre16_eval_plan_v1.3.pdf), 2018.
- [16] David Snyder et al., “Kaldi sre16 x-vector recipe,” <https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2>, 2017, Accessed: 2017-11.
- [17] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hanemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., “The kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.