Unsupervised regularization of the embedding extractor for robust language identification

Raphaël Duroselle, Denis Jouvet, Irina Illina

Unsupervised regularization of the embedding extractor for robust language identification

Raphaël Duroselle, Denis Jouvet, Irina Illina

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
raphael.duroselle@loria.fr denis.jouvet@inria.fr irina.illina@loria.fr

Abstract

State-of-the-art spoken language identification systems are constituted of three modules: a frame-level feature extractor, a segment-level embedding extractor and a final classifier. The performance of these systems degrades when facing mismatch between training and testing data. Most domain adaptation methods focus on adaptation of the final classifier. In this article, we propose a model-based unsupervised domain adaptation of the segment-level embedding extractor. The approach consists in a modification of the loss function used for training the embedding extractor. We introduce a regularization term based on the maximum mean discrepancy loss. Experiments were performed on the RATS corpus with transmission channel mismatch between telephone and radio channels. We obtained the same language identification performance as supervised training on the target domains but without using labeled data from these domains.

1. Introduction

Mismatch between domains appeared as a limitation of the generalization ability of the language recognition systems during the NIST Language Recognition Evaluation 2017 [1, 2]. All systems suffered from a performance drop between two datasets, CTS/BNBS recordings and audio extracted from videos. Only a small proportion of the development data was drawn from the audio from videos domain, althought it has different acoustic conditions from CTS. Such mismatch issue cannot be avoided when one wants to apply a model to real world data. In this work we address the most difficult domain mismatch configuration when there is no labeled data from the target domain. This problem is called unsupervised domain adaptation.

Domain mismatch theory presented in [3] shows that the gap between domains can be controlled by a divergence between the distributions of each domain in the data space. Consequently unsupervised domain adaptation can be understood as the design of domain invariant representations.

Language identification and speaker recognition systems are constitued of three main components [4]: a frame-level feature extractor [5], a segment-level embedding extractor (i-vector [6] or x-vector [4]) and a final classifier. Domain adaptation methods have been applied to the final classifier. They can be divided between feature-based and model-based domain adaptation methods. Feature-based methods transform representations of the domains to make them more similar [7]. The transformation can be a translation (domain mean adaptation) or a matrix multiplication (correlation alignment [8], feature distribution adaptator [7]). Model-based domain adaptation methods modify parameters of the classifier in order to improve performance on the target domain. For unsupervised adaptation, self-labeling [9] is a way of using predictions of the classifier on the target domain to retrain it and improve its performance. In [10], CORAL+, an unsupervised domain adaptation of PLDA, has been proposed for speaker recognition.

In this article, we propose a new approach for model-based unsupervised domain adaptation for robust language identification. Our main contribution is to perform this adaptation for the segment-level embedding extractor. The segment-level embedding extractor of a state-of-the-art language identification system is a neural network [11], which can be used for direct classification [12], or just to extract representations [4] which are then processed by a classifier. Different model-based domain adaptation methods have been introduced in the neural network training literature. They differ by the nature of the loss function: deep CORAL [13], maximum mean discrepancy [14, 15], and domain adversarial learning [16]. Domain adversarial training for speaker recognition embeddings was introduced to tackle language [17, 18, 19, 20] and environmental condition [19, 21] mismatch.

We propose to add a regularization term to the loss function of the segment-level embedding extractor, the second component of the language identification system. This regularization is based on the maximum mean discrepancy (MMD) which is an efficient measure of divergence between domains [22], without the training instabilities of domain adversarial training. We validate this method by training a language identification system for different radio transmission channels (UHF, VHF, HF), without using labeled data from these domains. We compared this approach to the adaptation of the final classifier, with embeddings trained with telephone data. The reported experiments lead us to the argument that unsupervised domain adaptation of the segment-level embedding extractor is more efficient than adaptation of the final classifier.

In Section 2, we describe the model-based unsupervised domain adaptation of the embedding extractor. In Section 3, we define the experimental protocol and results are presented and discussed in Section 4.

2. Unsupervised domain adaptation of the embedding extractor

As mentioned before, a language identification system contains three components: a frame-level feature extractor, a segment-level embedding extractor and a final classifier. These components are displayed in Figure 1, and detailed later in Section 3. In this section, we describe our method for model-based unsupervised domain adaptation of the embedding extractor. The segment-level embedding extractor is a discriminatively trained neural network for the language identification task. In order to

enforce invariance of its activations between source and target domains, we propose to use an unsupervised domain adaptation regularization.

2.1. Definition of the learning problem

To get the embedding extractor, a neural network f_θ of parameters θ is trained. This network takes a variable number of speech frames as input and outputs a posterior probability for each language class. The extracted embedding is a fixed-sized activation map of a given layer of the network.

A labeled corpus is necessary to learn the parameters of this network. We use the notation (x_S, y_S) where x_S are speech segments sampled from a source domain, defined by its distribution $\mathcal{D}_S$, and y_S are the associated language labels. The parameters θ of the network are chosen by minimizing a classification loss function L_{class} over this dataset :

$$\min_{\theta} \mathbb{E}_{(x_S,y_S)\sim\mathcal{D}_S}[L_{class}(f_\theta, x_S, y_S)] \quad (1)$$

This method is called supervised learning on the source domain.

For a domain adaptation problem, we aim at applying this model to a target domain defined by its distribution $\mathcal{D}_T$, which is significantly different from $\mathcal{D}_S$. Because of the limited quantity of training data, supervised learning cannot be directly applied on the target domain. In this paper we address a difficult domain adaptation task where no data from the target domain is annotated, namely unsupervised domain adaptation.

The problem can be formulated as follows. We only have access to unlabeled data $x_T \sim \mathcal{D}_T$ from the target domain and labeled data $x_S, y_S \sim \mathcal{D}_S$ from the source domain. Based on this data, how to build a model that achieves a low classification loss over the target domain?

2.2. Regularization of the embedding extractor

Two main approaches have been proposed to tackle this problem: feature-based domain adaptation and model-based domain adaptation. Both of them rely on the assumption that domains share common characterics that can be used to transfer knowledge from source to target domain [23].

For feature-based domain adaptation, a transformation is learned to map target data to the source domain where the model trained on source domain can be used. This approach includes CORAL [8], a recently introduced feature-based adaptator [7] and Cycle-GAN for the input features [24].

Model-based domain adaptation aims at selecting the parameters θ of the model so that its inner representations remain invariant between domains. Following the ideas of [3], classification loss on the target domain can be controlled by the sum of the classification loss function on the source domain and a measure L_R of divergence between domains. With a well chosen regularization loss L_R, the domain adaptation optimization problem can be solved by:

$$\min_{\theta} \mathbb{E}_{(x_S,y_S)\sim\mathcal{D}_S}[L_{class}(f_\theta, x_S, y_S)] + \lambda L_R(f_\theta, \mathcal{D}_S, \mathcal{D}_T) \quad (2)$$

where λ models the compromise between the classification performance on the source domain and the invariance between domains.

The choice of the regularization L_R defines the domain adaptation strategy. For a neural network, a natural choice is a measure of similarity between the distributions of the activations of a given layer. We use the notation $\Phi_f(x)$ for the activations of a given layer of a model f with an input speech segment x. In our experiments, we used the activations of the output layer of the network.

Several regularization losses have been introduced previously. Deep CORAL [13] measures the $\mathcal{L}_2$ distance between covariance matrices of the two domains. Adversarial training [16] provides an adaptative measure of discrepancy between domains. Approaches inspired from adversarial training have been applied to the segment-level embedding extractor in speaker recognition systems [17, 18, 19, 20, 21]. Nevertheless they rely on a careful design of discriminators and have to tackle the instability of the adversarial training process.

In this work we use the Maximum Mean Discrepancy loss for unsupervised regularization of the segment-level embedding extractor.

2.3. Maximum Mean Discrepancy loss

When the similarity between activations $\Phi_f(x)$ and $\Phi_f(x')$ is measured with a positive semi-definite kernel $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, (n being the dimension of the activation maps $\Phi_f(x)$), a well-known measure of discrepancy between domains is Maximum Mean Discrepancy (MMD). The MMD loss for unsupervised regularization can be computed as:

$$L_R = \mathbb{E}_{x_S,x'_S\sim\mathcal{D}_S}[k(\Phi_f(x_S), \Phi_f(x'_S))] + \mathbb{E}_{x_T,x'_T\sim\mathcal{D}_T}[k(\Phi_f(x_T), \Phi_f(x'_T))] - 2\,\mathbb{E}_{x_S\sim\mathcal{D}_S,x_T\sim\mathcal{D}_T}[k(\Phi_f(x_S), \Phi_f(x_T))] \quad (3)$$

The MMD loss is differentiable and can be efficently estimated with finite sets of samples from both domains, even in a high dimensional space [22]. Hence, it is a commonly used measure of discrepancy for training generative models [25, 26]. It has also been used for feature-based adaptation of i-vectors in a speaker recognition system [15]. An efficient GPU implementation of the MMD loss has been provided by the authors of [27].

If k is a universal kernel, $L_R = 0$ is equivalent to equality of distributions. As a univeral kernel, we use the gaussian kernel:

$$k(\Phi_f(x), \Phi_f(x')) = \exp(-\frac{||\Phi_f(x) - \Phi_f(x')||_2^2}{2\,\sigma^2}) \quad (4)$$

The unsupervised MMD regularization of the segment-level embedding extractor depends on only two hyperparameters: the weight λ of the regularization loss and the variance σ^2 of the kernel. Furthermore, training of a neural network with the MMD regularization loss does not suffer from the instability of adversarial training. That makes it an appealing method for domain adaptation.

3. Design of the experiments

3.1. The RATS Corpus

Our spoken language identification system has been evaluated on the RATS corpus [28]. We used the releases LDC2015S02 and LDC2017S20 that contain five languages (Arabic, English, Persian, Pushto, Urdu). The original data from a telephone channel (called *src*) was recorded under eight different radio channels: *A*, *B*, *C*, *F*, *G* (UHF channels), *E* (VHF channel), *D* and *H* (HF channels). These channels have different noise and distorsion characteristics and are challenging for a classification

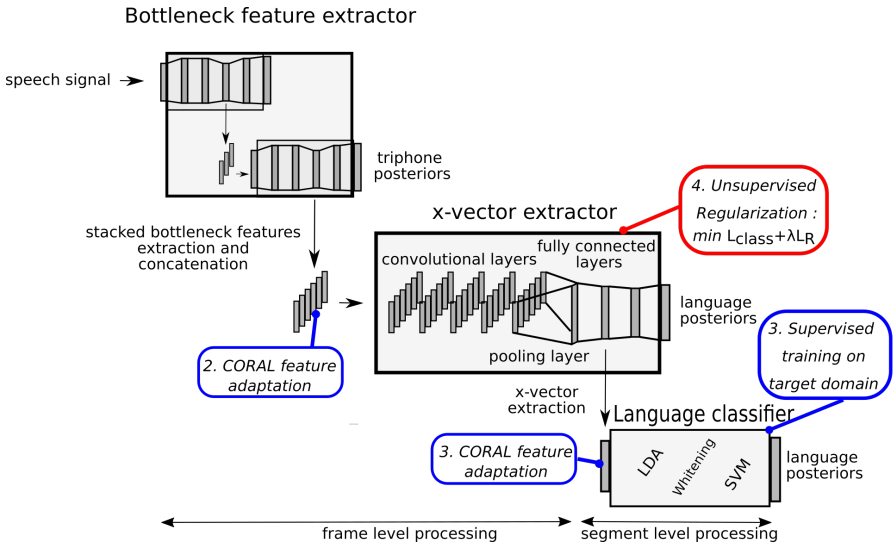

Figure 1: *Architecture of the language identification system. The frame-level feature extractor is a bottleneck feature extractor. The segment-level embedding extractor provides x-vectors that are then processed by the language classifier to produce language posteriors. Numbered blocks correspond to domain adaptation methods. Baseline methods (1. CORAL feature adaptation of the x-vector extractor, 2. CORAL feature adaptation of the classifier, 3. Supervised training of the classifier on the target domain) are described in subsection 3.3. The introduced method (4. Unsupervised Regularization of the embedding extractor) is described in section 2.*

system. They are provided with speech activity labels and only speech segments are used. We divided the corpus into three sets: training, validation and testing. We ensured that parallel utterances recorded on different channels belong to the same set (see Figure 2).

In order to avoid the possible bias of this corpus with parallel utterances, for domain adaptation experiments, we only used half of the training set and half of the validation set. The selection criteria ensures that the utterances used for adaptation to the target domains do not have the same linguistic content as the one used for the source domain (see Figure 2). In our experiments, the source domain was the telephone channel src.

In [6, 29, 30, 31], language identification on the RATS corpus was studied with the release LDC2018S10 which also contains five languages: Arabic, Persian, Dari, Pushto, Urdu. The authors of [6] achieved an average equal error rate (EER) of 9.59% for this task, using 3-second speech segments with stacked bottleneck features followed by an i-vector extractor and a neural network classifier. This is the best reported performance on RATS. It is important to note that this result was obtained using a system trained with labeled data from all channels (like all other competing systems).

In [32] domain adaptation on the RATS corpus was studied for a speaker recognition task by adapting the LDA matrix of the final classifier in order to compensate the dataset shift. The experiments reported in the following subsections aim at comparing the impact of domain adaptation applied during the training of the segment-level embedding extractor, to corrections applied to the final classifier.

3.2. Language identification system

The three-step language identification pipeline is described in Figure 1. The three components are a frame-level feature extractor, a segment-level embedding extractor and a final classifier.

The first module extracts frame-level features. These features are multilingual bottleneck features, extracted from two stacked feed-forward neural networks which were trained to recognise triphones for seventeen languages of the Babel corpus. We used the BUT/Phonexia bottleneck feature trained models [5] that exhibited good language identification performance on the NIST LRE 2017 dataset [33]. These frame-level features of dimension 80 were concatenated along the time dimension. During training, segments of 3 seconds (300 frames of 10 ms) were presented to the segment-level embedding extractor.

For the segment-level embedding extractor, we used the

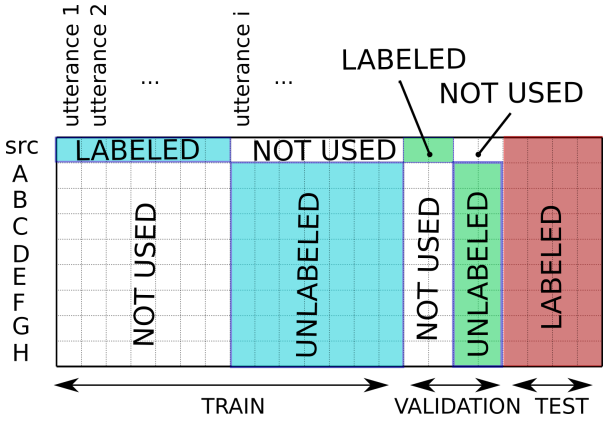

Figure 2: *Division of the RATS corpus. Colored blocks are used for training, validation and testing. Blocks are used with language identification labels (labeled blocks) or without labels (unlabeled blocks). White blocks are not used. Rows represent transmission channels and columns utterances. For instance, the recording of utterance* 1 *on channel* src *belongs to the training set with a language label, the recordings of utterance* i *on all other channels belong to the training set without labels.*

architecture of the language identification classifier developed for x-vector extraction that demonstrated state-of-the-art performance on the NIST LRE 2017 dataset [4]. It is a feed-forward neural network taking variable length speech segments as inputs and constituted of five 1d-convolutional layers followed by a statistics pooling layer and three segment-level layers. x-vectors of dimension 512 are extracted from the antepenultimate layer (the *segment 6* layer using the notations in [4]).

The segment-level embedding extractor was trained by stochastic gradient descent using the cross-entropy loss with a learning rate of 10^{-1}, which was divided by two each time the validation loss had not improved for five epochs. We used overlapping segments of 3-second speech with 1 second shift between segments, to increase the quantity of segment data for training. Minibatches of size 500 were used for training.

The final classifier takes the x-vector embeddings as inputs. We performed dimensionality reduction with Linear Discriminant Analysis (reducing the dimension of the x-vectors to 4) and performed whitening of the resulting embeddings. Then a Support Vector Machine [34] with the linear kernel was trained to predict the language label. The outputs of this classifier are used for the global evaluation of the system.

3.3. Baseline unsupervised domain adaptation methods

We compared our model-based unsupervised domain adaptation of the x-vector extractor (as described in Section 2) to three methods: 1. CORAL feature-based adaptation of the segment-level embedding extractor, 2. CORAL feature-based adaptation of the final classifier and 3. Supervised training of the final classifier on the target domain. Each method is represented in Figure 1.

CORAL [35] is an unsupervised feature-based domain adaptation method. It is a transformation (translation and matrix multiplication) with the objective of making mean and covariance matrices of both domains identical in the representation space. It has been sucessfully applied in the x-vector space [8] for feature-based adaptation of the final classifier, for a speaker verification task. We applied it for adapting our final classifier. We also applied it in the bottleneck feature space as a feature-based adaptation of the embedding extractor.

In [36], unsupervised model-based adaptation of the final classifier has been used for speaker recognition. For speaker recognition, the final classifier is a PLDA. In the following, we compare the use of domain invariant segment-level embeddings to adaptation of the final classifier. We have performed a supervised training of the final classifier on the target domain to get an upper bound of the potential performance of unsupervised adaptation of the final classifier to the target domain.

3.4. Analysis of the hyperparameters of the proposed MMD regularization

The proposed MMD based model-adaptation method depends on two hyperparameters: the weight of the regularization loss λ and the variance σ^2 of the gaussian kernel.

The variance σ^2 defines the scale of the differences between domains that are measured by the MMD loss. The segment-level embedding extractor is intialized with random weights. Consequently, at the beginning of training, distributions of activations of each domain are very similar and the MMD loss is close to 0. During training of the network, domains start to separate themselves. When the scale of the differences between domains reaches σ^2, it activates the MMD loss. As a consequence, if σ^2 is very small, the training is constrained, the domains don't separate themseleves but the classification loss is not improved. Conversely, if σ^2 is too high, the MMD loss is not activated and there is no domain adaptation at all. In practice, to get fast convergence, λ must be chosen carefully depending on the value of σ^2. We found that domain adaptation was possible with values of σ^2 between 1 and 10^4, with no clear difference within this range of values. In all experiments presented below, σ^2 is set to 10.

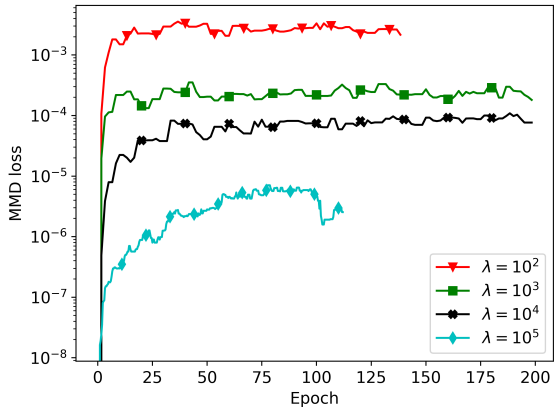

Figure 3: *MMD loss on the validation data with respect to the number of training epochs for $\sigma^2 = 10$ and different values of λ, src as source domain, D as target domain.*

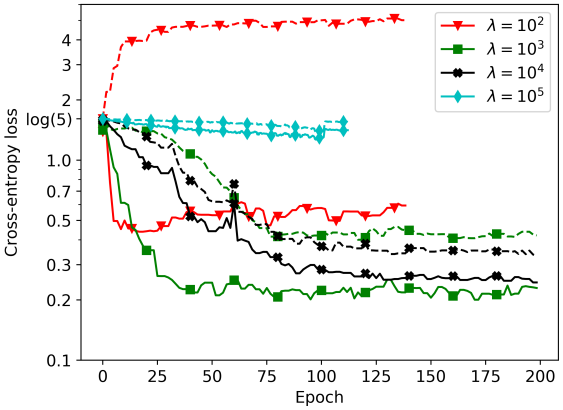

Figure 4: *Classification loss on the validation data with respect to the number of training epochs for $\sigma^2 = 10$ and different values of λ. Solid lines refer to the source domain (channel src) and dashed lines to the target domain (channel D). All curves start at $\log(5)$ because the parameters of the network are randomly initialized.*

Figures 3 and 4 report the same learning process, with different values of λ, with channel src as source domain and channel D as target domain. Trainings with different values of λ do not take the same number of epochs, because the decrease in the learning rate depends on the evolution of the validation loss.

All loss functions are measured on the validation set. Solid lines refer to loss functions that are explicitely minimized by the training process: classification loss on the source domain (Fig-

ure 4) and MMD regularization loss between domains (Figure 3). Dashed lines of Figure 4 refer to the classification loss on the target domain. The objective of model-based unsupervised domain adaptation is the minimization of this loss on the target domain by minimizing a combination of the classification loss on source domain and the regularization loss, with weight λ (cf. equation (2)).

Figure 3 shows the evolution of the MMD loss during training. For all parameters λ, the regularization loss achieves a maximum value that is controled by the value of λ. The training process reaches a configuration where the MMD loss is activated, meaning that the differences between domains remain at a scale of order σ^2.

Figure 4 displays the classification loss functions on both source domain (solid lines) and target domain (dashed lines). High values of λ slow down training on the source domain but allow to reduce the gap between classification loss on the target and source domains. With low values of λ, the classification loss on the target domain diverges. Interestingly, for $\lambda = 10^2$, unsupervised adaptation is inefficient but it also prevents convergence on the source domain. For higher values, the choice of λ is a compromise between the performance gap between domains and the duration of the convergence process on the source domain. In practice for very high values of λ, training is so constrained that convergence on the source domain is very slow.

After these preliminary experiments, we set the values of the hyperparameters to $\lambda = 10^4$ and $\sigma^2 = 10$. Experiments in Section 4 show that these hyperparameters are robust to different target channels. The unsupervised domain adaptation scenario does not allow to select the best hyperparameters values for each configuration since performance on the target domain cannot be measured. Consequently robustness of the hyperparameters is necessary for any unsupervised domain adaptation method.

4. Results

First we trained our language identification system on every domain and checked that it exhibits a domain mismatch issue. Then we applied our unsupervised adaptation method and compared it to the baseline domain adaptation approaches.

4.1. Performance evaluation

In this work, we evaluate the potential of unsupervised model-based adaptation of the segment-level embedding extractor to produce domain invariant representations for the language identification task. To evaluate directly the quality of representations, we used a metric that does not depend on calibration: the average equal error rate (EER) [6, 29, 30, 31]. One EER is computed for each language of the corpus and they are averaged. Evaluations were performed with 3-second speech segments.

4.2. Supervised performance of the system

Table 1 presents language identification performance in terms of average equal error rate (EER) for each training and testing domains. Both the segment-level embedding extractor and the final classifier are trained on the same training domain. In the last row, we train a system on the whole training set (i.e., training data from all domains). For this system trained on all channels, the average EER over all channels is 9.36%. The use of a discriminatively trained neural network instead of i-vectors, equals previous reported results on RATS [6, 29, 30, 31]. The other results in Table 1 clearly show that, when trained on only one channel, the system suffers from domain mismatch when the test differs from the training channel. Outside of the diagonal of Table 1, average equal error rates of around 50% mean that the system is totally inefficient on the test domain.

When labeled data is only available on the telephone channel (*src*), performance of the system on the other domains is very poor (first line of Table 1). These values are reported in Table 2.

Table 1: *Average EER (%) of the language identification system (3-second speech segments). Rows: training channel of the embedding extractor and of the final classifier. Columns: testing channel. 'all' means training of the system using labeled data from all the channels. Values rounded to the nearest integer.*

	Test channel								
Train	src	A	B	C	D	E	F	G	H
src	**6**	50	42	34	39	48	45	17	43
A	42	**15**	40	43	39	49	47	48	49
B	45	32	**12**	53	39	48	52	53	44
C	45	35	40	**13**	39	43	50	46	34
D	38	35	38	47	**7**	47	45	41	49
E	37	58	51	42	52	**14**	47	39	36
F	38	44	42	42	40	45	**13**	39	42
G	11	46	44	33	34	44	32	**9**	36
H	52	42	52	35	47	43	49	45	**14**
all	**5**	**10**	**10**	**10**	**7**	**12**	**13**	**7**	**12**

4.3. Unsupervised domain adaptation

We applied unsupervised domain adaptation from telephone data (channel src) to every target domain and present performance of each method in Table 2 for 3-second speech segments. We compared these results to the performance obtained using supervised training with labeled data from the source and from the target domains. Results of the eight target domains are consistent.

First, the results show that none of the baseline domain adaptation method outperforms supervised training on the target domain.

CORAL feature-based adaptation gives a small improvement when applied to the final classifier (method 1). Inspite of being a commonly used tool for adaptation of the backend of a speaker recognition x-vector system [7, 8], the limited improvement achieved by this method illustrates the difficulty of the domain adaptation task on the RATS corpus. This indicates that mismatch in the x-vector space cannot be compensated by a linear transformation

CORAL feature-based adaptation gives a more substantial improvement when applied to the input of the x-vector extractor (method 2). This supports our claim that invariance of the representations has to be enforced at the stage of the segment-level embedding extractor in a language identification system. The CORAL feature transformation has been applied separately to each frame of the input segments. Consequently a better improvement can be expected from a transformation of the whole speech segments.

When CORAL feature-based adaptation is applied to both the x-vector extractor and the final classifier (methods 1 and 2), there is no significant gain in comparison with a system where it is only applied to the x-vector extractor.

Table 2: *Performance obtained with various training approaches of the embedding extractor and of the final classifier. MMD model adaptation and CORAL feature adaptation are unsupervised approaches (i.e., they do not use labels for the training target data). Training is done using telephone source domain data (src) and different target domain data. The first two rows correspond to values already reported in Table 1: the performance of a system fully trained with labels on source domain and of systems trained on target domains (i.e. diagonal of Table 1). The second block displays performance of the domain adaptation baselines. The last two lines present performance of the introduced MMD model-based adaptation of the segment-level embedding extractor with classifiers trained on source or on target domains. The method numbers refer to Figure 1.*

Method number	Training method		Average EER on target domain (%)							
	embedding extractor	final classifier	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
	supervised on *source*	supervised on *source*	50.2	42.3	34.4	39.6	48.5	45.1	17.4	43.6
	supervised on *target*	supervised on *target*	14.6	12.5	12.6	6.7	13.6	13.5	8.6	14.2
1	supervised on *source*	*CORAL feature adaptation*	38.6	38.0	32.5	32.8	38.8	33.7	13.2	42.5
2	*CORAL feature adaptation*	supervised on *source*	40.7	34.9	32.0	27.8	33.2	30.3	13.2	32.8
1 and 2	*CORAL feature adaptation*	*CORAL feature adaptation*	39.5	35.3	31.9	28.0	32.5	29.8	13.2	32.7
3	supervised on *source*	supervised on *target*	15.8	15.0	14.1	14.3	21.8	20.5	9.8	18.8
4	*MMD model adaptation*	supervised on *source*	12.7	10.6	11.7	7.6	13.3	11.9	5.5	12.2
3 and 4	*MMD model adaptation*	supervised on *target*	**10.2**	**9.2**	**11.3**	**6.0**	**11.8**	**10.3**	**5.1**	**10.0**

Supervised training of the final classifier on the target domain (method 3) is not a domain adaptation method but it provides a lower bound on the average EER that can be achieved by a model-based adaptation of the final classifier. Comparison with this method relies on the assumption that no model-based adaptation of the final classifier could achieve better performance than supervised training on the target domain with the same data quantity. Quite logically supervised training of the classifer on the target domain with x-vectors trained on the source domain achieves worse EERs than supervised training of the whole system on the target domain, with a gap between 1.2% for the channels A and G and 7.6% for the channel D. It means that embeddings that have been trained on the source domain are not as efficient for classification on the target domain as embeddings trained on the target domain.

The proposed MMD model-based adaptation of the segment-level embedding extractor (method 4) precisely aims at reducing this gap. It is trained to produce domain invariant embeddings. The first observation is that the mismatch between domains is drastically reduced since a final classifier trained on these embeddings on the source domain achieves better performance on the target domain than any of the three domain adaptation baselines.

In fact, the performance on the target domain of this system, a final classifier trained on the source domain with embeddings extracted from a regularized network (method 4), is superior to a fully supervised training on the target domain (line 2 of Table 2), except for the channel D. We conclude that the MMD model-based adaptation allows to extract better embeddings for the language identification task.

Table 2 allows comparing two final classifiers trained on the target domain with two different embeddings: embeddings trained on the target domain (line 2 of Table 2) and embeddings trained with MMD model-based adaptation (methods 3 and 4). MMD model-based adaptation allows an absolute reduction of the average EER ranging between 0.7% on the channel D and 4.4% on the channel A. We conclude that the MMD model-based adaptation of the x-vector extractor is useful to improve the quality of the x-vectors, even when labeled data from the target domain are available. We hypothesize that MMD model-based adaptation helps removing a nuisance factor in the training dataset and consequently increases the capacity of the x-vector extractor.

Finally we observe that there is still a mismatch between domains in the x-vector space since a final classifier trained on the target domain achieves a better performance than a final classifier trained on the source domain (last two lines of Table 2). The model-based adaptation of the segment-level embedding extractor allows to transfer classification ability between domains but a greater improvement can be expected by combining it with an adaptation of the final classifier.

5. Conclusion

Using the Maximum Mean Discrepancy unsupervised domain adaptation loss, we demonstrated the superiority of model-based adaptation of the segment-level embedding extractor over adaptation of the final classifier in a language identification system. On the RATS corpus, we showed that unsupervised domain adaptation of the embedding extractor allows training of a language identification system for eight radio channels without using labeled data from these channels and with better performance than supervised training for seven of them.

In this work we achieved the fully unsupervised domain adaptation of an x-vector extractor. The method only depends on two hyperparameters: the weight λ of the MMD regularization loss and the variance σ^2 of the kernel. The simplicity of the MMD-based regularization allows it to be applied to other kinds of domain mismatch in order to evaluate its usefulness to improve the quality of x-vector embeddings.

6. Acknowledgements

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr). This work has been partly funded by the French Direction Générale de l 'Armement.

7. References

[1] Seyed Omid Sadjadi, Timothee Kheyrkhah, Audrey Tong, Craig S Greenberg, Douglas A Reynolds, Elliot Singer, Lisa P Mason, and Jaime Hernandez-Cordero, "The

2017 NIST language recognition evaluation.," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 82–89.

[2] Seyed Omid Sadjadi, Timothee Kheyrkhah, Craig S Greenberg, Elliot Singer, Douglas A Reynolds, Lisa P Mason, and Jaime Hernandez-Cordero, "Performance analysis of the 2017 NIST language recognition evaluation.," in *Proceedings of INTERSPEECH 2018 - Annual Conference of the International Speech Communication Association*, 2018, pp. 1798–1802.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan, "A theory of learning from different domains," in *Machine learning*. 2010, vol. 79, pp. 151–175, Springer.

[4] David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "Spoken language recognition using x-vectors.," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 105–111.

[5] Radek Fer, Pavel Matějka, František Grézl, Oldřich Plchot, Karel Veselý, and Jan Honza Černocký, "Multilingually trained bottleneck features in spoken language recognition," in *Computer Speech & Language*. 2017, vol. 46, pp. 252–267, Elsevier.

[6] Pavel Matějka, Le Zhang, Tim Ng, Sri Harish Mallidi, Ondřej Glembek, Jeff Ma, and Bing Zhang, "Neural network bottleneck features for language identification," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 299–304.

[7] Pierre-Michel Bousquet and Mickael Rouvier, "On robustness of unsupervised domain adaptation for speaker recognition," in *Proceedings of INTERSPEECH 2019 - Annual Conference of the International Speech Communication Association*, 2019, pp. 2958–2962.

[8] Jahangir Alam, Gautam Bhattacharya, and Patrick Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adaptation.," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 176–180.

[9] Shahan Nercessian, Pedro Torres-Carrasquillo, and Gabriel Martinez-Montes, "Approaches for language identification in mismatched environments," in *Proceedings of SLT 2016 - IEEE Spoken Language Technology Workshop*. IEEE, 2016, pp. 335–340.

[10] Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka, "The CORAL+ algorithm for unsupervised domain adaptation of PLDA," in *Proceedings of ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 5821–5825.

[11] Alicia Lozano Diez, *Bottleneck and Embedding Representation of Speech for DNN-Based Language and Speaker Recognition*, Ph.D. thesis, Universidad Autónoma de Madrid, 2018.

[12] Jesus Antonio Villalba Lopez, Niko Brümmer, and Najim Dehak, "End-to-end versus embedding neural networks for language recognition in mismatched conditions," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 112–119.

[13] Baochen Sun and Kate Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proceedings of ECCV 2016 - European Conference on Computer Vision*. Springer, 2016, pp. 443–450.

[14] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan, "Learning transferable features with deep adaptation networks," in *Proceedings of ICML 2015 - International Conference on Machine Learning*, 2015, pp. 97–105.

[15] Wei-Wei Lin, Man-Wai Mak, Longxin Li, and Jen-Tzung Chien, "Reducing domain mismatch by maximum mean discrepancy based autoencoders.," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 162–167.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," in *The Journal of Machine Learning Research*, 2016, vol. 17, pp. 2096–2030.

[17] Johan Rohdin, Themos Stafylakis, Anna Silnova, Hossein Zeinali, Lukáš Burget, and Oldřich Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *Proceedings of ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6006–6010.

[18] Gautam Bhattacharya, Joao Monteiro, Jahangir Alam, and Patrick Kenny, "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification," in *Proceedings of ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6226–6230.

[19] Gautam Bhattacharya, Jahangir Alam, and Patrick Kenny, "Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training," in *Proceedings of ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6041–6045.

[20] Wei Xia, Jing Huang, and John HL Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *Proceedings of ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 5816–5820.

[21] Zhong Meng, Yong Zhao, Jinyu Li, and Yifan Gong, "Adversarial speaker verification," in *Proceedings of ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6216–6220.

[22] Gabriel Peyré and Marco Cuturi, "Computational optimal transport," in *Foundations and Trends® in Machine Learning*. 2019, vol. 11, pp. 355–607, Now Publishers, Inc.

[23] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy, "Optimal transport for domain adaptation," in *IEEE transactions on pattern analysis and machine intelligence*, 2016, vol. 39, pp. 1853–1865.

[24] Phani Sankar Nidadavolu, Saurabh Kataria, Jesús Villalba, and Najim Dehak, "Low-resource domain adaptation for speaker recognition using Cycle-GANs," 2019.

[25] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani, "Training generative neural networks via

maximum mean discrepancy optimization," in *Proceedings of UAI 2015 - Conference on Uncertainty in Artificial Intelligence*, 2015, pp. 258–267.

[26] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos, "MMD GAN: Towards deeper understanding of moment matching network," in *Advances in Neural Information Processing Systems*, 2017, pp. 2203–2213.

[27] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré, "Interpolating between optimal transport and MMD using Sinkhorn divergences," in *Proceedings of The Twenty-second International Conference on Artificial Intelligence and Statistics*, 2019, pp. 2681–2690.

[28] Kevin Walker and Stephanie Strassel, "The RATS radio traffic collection system," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2012, pp. 291–297.

[29] Yun Lei, Luciana Ferrer, Aaron Lawson, Mitchell McLaren, and Nicolas Scheffer, "Application of convolutional neural networks to language identification in noisy conditions," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014, vol. 41, pp. 1–8.

[30] Jeff Z Ma, Bing Zhang, Spyros Matsoukas, Sri Harish Reddy Mallidi, Feipeng Li, and Hynek Hermansky, "Improvements in language identification on the RATS noisy speech corpus," in *Proceedings of INTERSPEECH 2013 - Annual Conference of the International Speech Communication Association*, 2013, pp. 69–73.

[31] Kyu Jeong Han, Sriram Ganapathy, Ming Li, Mohamed Kamal Omar, and Shrikanth Narayanan, "TRAP language identification system for RATS phase II evaluation," in *Proceedings of INTERSPEECH 2013 - Annual Conference of the International Speech Communication Association*, 2013, pp. 1502–1506.

[32] Ondřej Glembek, Jeff Ma, Pavel Matějka, Bing Zhang, Oldřich Plchot, Lukáš Burget, and Spyros Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *Proceedings of ICASSP 2014 - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4032–4036.

[33] Oldřich Plchot, Pavel Matějka, Ondřej Novotný, Sandro Cumani, Alicia Lozano-Diez, Josef Slavicek, Mireia Diez, František Grézl, Ondřej Glembek, Kamsali Veera Mounika, Anna Silnova, Lukáš Burget, Lucas Ondel, Santosh Kesiraju, and Johan Rohdin, "Analysis of BUT-PT submission for NIST LRE 2017," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 47–53.

[34] Najim Dehak, Pedro A Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proceedings of INTERSPEECH 2011 - Annual Conference of the International Speech Communication Association*, 2011, pp. 857–860.

[35] Baochen Sun, Jiashi Feng, and Kate Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 2058–2065.

[36] Daniel Garcia-Romero, Alan McCree, Stephen Shum, Niko Brümmer, and Carlos Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 260–264.