Adaptation de domaine non supervisée pour la reconnaissance de la langue par régularisation d'un réseau de neurones

Raphaël Duroselle, Denis Jouvet, Irina Illina

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Adaptation de domaine non supervisée pour la reconnaissance de la langue par régularisation d'un réseau de neurones

Raphaël Duroselle[1] Denis Jouvet[1] Irina Illina [1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
raphael.duroselle@loria.fr denis.jouvet@inria.fr irina.illina@loria.fr

RÉSUMÉ

Les systèmes automatiques d'identification de la langue subissent une dégradation importante de leurs performances quand les caractéristiques acoustiques des signaux de test diffèrent fortement des caractéristiques des données d'entraînement. Dans cet article, nous étudions l'adaptation de domaine non supervisée d'un système entraîné sur des conversations téléphoniques à des transmissions radio. Nous présentons une méthode de régularisation d'un réseau de neurones consistant à ajouter à la fonction de coût un terme mesurant la divergence entre les deux domaines. Des expériences sur le corpus OpenSAD15 nous permettent de sélectionner la *Maximum Mean Discrepancy* pour réaliser cette mesure. Cette approche est ensuite appliquée à un système moderne d'identification de la langue reposant sur des *x-vectors*. Sur le corpus RATS, pour sept des huit canaux radio étudiés, l'approche permet, sans utiliser de données annotées du domaine cible, de surpasser la performance d'un système entraîné de façon supervisée avec des données annotées de ce domaine.

ABSTRACT

Unsupervised domain adaptation for language identification by regularization of a neural network

Automatic spoken language identification systems suffer from a performance drop when acoustic characterics of the test signal differ in a significative way from the characteristics of the training data. In this paper, we study the unsupervised domain adaptation of a system trained on conversational telephone speech to radio transmission channels. We present a regularization method for a neural network which consists in adding to the cost function a term that measures the discrepancy between domains. Based on experiments on the corpus OpenSAD15, we select the *Maximum Mean Discrepancy* loss to perform this measure. This approach is then applied to a state-of-the-art x-vector system. On the RATS corpus, for seven of the eight studied radio channels, our approach achieves a better performance on the target domain than a system trained in a supervised way using labelled data from this domain.

MOTS-CLÉS : adaptation de domaine non supervisée, identification de la langue, régularisation, maximum mean discrepancy, robustesse.

KEYWORDS: unsupervised domain adaptation, language identification, regularization, maximum mean discrepancy, robustness.

Volume 1 : Journées d'Études sur la Parole, pages 190–198.
HAL : hal-02798536.

1 Introduction

Un système d'identification de la langue est habituellement entraîné sur un corpus d'apprentissage spécifique à un environnement. Si les données de test ne proviennent pas de la même distribution que les données d'entraînement (et ont donc des caractéristiques différentes), les performances du système peuvent chuter significativement. Dans cet article, nous étudions l'effet du changement de canal de transmission entre les données d'entraînement et de test. Les données d'entraînement sont des conversations téléphoniques. Nous voulons appliquer un tel système à des communications radio, pour lesquelles nous ne disposons pas de données annotées. Ce problème est appelé adaptation de domaine non supervisée.

La possibilité d'adapter un système de classification fonctionnant sur un domaine source à un domaine cible repose sur l'hypothèse que les distributions des données des deux domaines partagent des caractéristiques communes pouvant être utilisées pour la classification (Ben-David *et al.*, 2010). Par conséquent l'adaptation de domaine peut être réalisée en utilisant des représentations invariantes entre les domaines. Dans ce but, deux types d'approche ont émergé (Bousquet & Rouvier, 2019) : les méthodes *feature-based*, qui transforment les représentations des données du domaine source afin de les rendre similaires au domaine cible, et les méthodes *model-based*.

Lors d'une adaptation *model-based*, les paramètres du modèle sont déterminés en prenant en compte l'objectif de généralisation au domaine cible. Nous proposons une approche *model-based* s'appliquant à un réseau de neurones dont les paramètres sont obtenus par minimisation d'une fonction de coût. Un terme de régularisation est ajouté à la fonction de coût afin de prendre en compte la contrainte d'invariance entre les domaines. Différentes fonctions de régularisation ont été proposées dans la littérature en traitement de l'image et analyse de texte : *deep CORAL* (Sun & Saenko, 2016) *Maximum Mean Discrepancy* (Long *et al.*, 2015), des fonctions de coût antagonistes (Ganin *et al.*, 2016). Jusqu'à présent aucune de ces approches n'a été appliquée à la reconnaissance de la langue.

Dans ce travail, nous comparons d'abord trois fonctions de coût pour l'adaptation à des canaux radio d'un réseau de neurones entraîné pour la tâche d'identification de la langue : la distance entre les moyennes des distributions, *deep CORAL* et la *Maximum Mean Discrepancy*. Nous montrons que cette dernière permet d'annuler la baisse de performance due à l'absence de données annotées sur le domaine cible.

Dans un second temps, nous étudions un système d'identification de la langue correspondant à l'état de l'art (Snyder *et al.*, 2018; Plchot *et al.*, 2018), constitué d'un extracteur de *features*, d'un extracteur de vecteurs représentatifs des segments audio et d'un classifieur final. Un tel système subit bien une dégradation importante de performance due au changement de canal de transmission entre les données d'entraînement et de test. Notre approche appliquée au module d'extraction de vecteurs représentatifs du segment audio permet de réduire cette dégradation et conduit même à des performances meilleures que celles d'un système entraîné de façon supervisée sur le domaine cible.

2 Méthode d'adaptation de domaine non supervisée d'un réseau de neurones

Nous nous plaçons dans le cadre d'une adaptation de domaine non supervisée. Nous disposons de données annotées (x_S, y_S) provenant d'un domaine source défini par sa distribution $\mathcal{D}_S$, et de

données non annotées x_T d'un domaine cible défini par sa distribution $\mathcal{D}_T$. Les x_S et x_T sont les données audio et les y_S les étiquettes de langue associées. L'objectif de la tâche d'adaptation de domaine est l'entraînement d'un système d'identification de langue performant sur le domaine cible.

2.1 Régularisation de la fonction de coût

Nous nous intéressons à un modèle de classification pour la tâche d'identification de la langue. C'est un réseau de neurones f_θ de paramètres θ. Ses paramètres sont appris de façon supervisée en minimisant l'entropie croisée L_{CE} sur le domaine source :

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

Fondée sur le constat que l'erreur du modèle sur le domaine cible peut être contrôlée par la somme de l'erreur sur le domaine source et d'une mesure de divergence entre les domaines (Ben-David *et al.*, 2010), notre méthode *model-based* d'adaptation de domaine non supervisée consiste à ajouter une fonction de régularisation L_R à la fonction de coût. Le problème d'optimisation devient :

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

L_R est une mesure de la divergence des représentations du réseau entre les distributions $\mathcal{D}_S$ et $\mathcal{D}_T$. λ est un paramètre représentant le compromis entre bonne performance de classification sur le domaine source et invariance des représentations entre les domaines. En pratique, on choisit une couche du réseau et la fonction de coût L_R est mesurée pour les activations de celle-ci. Nous utilisons la notation $\Phi_f(x)$ pour les valeurs des activations de cette couche pour un réseau f et une donnée d'entrée x. Dans nos expériences, nous nous plaçons sur la couche de sortie du réseau.

Différentes fonctions de régularisation L_R ont été introduites : *deep CORAL* (Sun & Saenko, 2016), *Maximum Mean Discrepancy* (Long *et al.*, 2015; Lin *et al.*, 2018), ainsi que des fonctions de coût antagonistes (*adversarial*) (Ganin *et al.*, 2016). Dans ce travail, nous comparons trois fonctions de régularisation, basées sur la distance entre les moyennes des distributions, sur la distance entre les seconds moments (*deep CORAL*) et sur la *Maximum Mean Discrepancy*.

2.2 Fonctions de régularisation

Une correction simple à appliquer à deux distributions de probabilité pour les rapprocher serait de leur faire partager la même moyenne. Par conséquent, notre première fonction de régularisation est le carré de la **distance euclidienne entre les moyennes** des distributions des deux domaines :

$$L_{moy} = || \mathbb{E}_{x_S\sim\mathcal{D}_S}[\Phi_f(x_S)] - \mathbb{E}_{x_T\sim\mathcal{D}_T}[\Phi_f(x_T)] ||_2^2 \tag{3}$$

Dans le même esprit, la fonction de coût **deep CORAL** (Sun & Saenko, 2016) vise à aligner les seconds moments des deux distributions. Elle correspond au carré de la distance euclidienne entre les matrices de covariance des distributions de chacun des deux domaines :

$$L_{CORAL} = ||C_S - C_T||_2^2 \tag{4}$$

où C_S et C_T sont les matrices de covariance des activations $\Phi_f(x)$ sur les domaines $\mathcal{D}_S$ et $\mathcal{D}_T$.

Enfin, la ***Maximum Mean Discrepancy*** (MMD) est une mesure de divergence entre les domaines basée sur une mesure de similarité entre paires d'échantillons définie par un noyau semi-défini positif k. Elle prend la valeur :

$$L_{MMD} = \mathop{\mathbb{E}}_{x_S, x'_S \sim \mathcal{D}_S}[k(\Phi_f(x_S), \Phi_f(x'_S))] + \mathop{\mathbb{E}}_{x_T, x'_T \sim \mathcal{D}_T}[k(\Phi_f(x_T), \Phi_f(x'_T))] - 2\mathop{\mathbb{E}}_{x_S \sim \mathcal{D}_S, x_T \sim \mathcal{D}_T}[k(\Phi_f(x_S), \Phi_f(x_T))] \quad (5)$$

Lorsque le noyau est le produit scalaire usuel alors L_{MMD} est équivalente à L_{moy}, présentée précédemment. Pour prendre en compte de façon plus fine l'écart entre les distributions, nous utilisons un noyau gaussien, de variance notée σ^2 :

$$k(\Phi_f(x), \Phi_f(x')) = \exp(-\frac{||\Phi_f(x) - \Phi_f(x')||_2^2}{2\,\sigma^2}) \quad (6)$$

La régularisation *MMD* est une mesure de divergence entre deux distributions de probabilité, pouvant être estimée à partir d'un nombre fini d'échantillons, y compris dans des espaces de haute dimension (Peyré & Cuturi, 2019). Dans le domaine du traitement de la parole, elle a été utilisée pour l'adaptation de domaine *feature-based* d'un système de reconnaissance du locuteur (Lin *et al.*, 2018). De plus, l'estimation de la *Maximum Mean Discrepancy* peut être réalisée de façon efficace sur un GPU (Feydy *et al.*, 2019).

Au cours de l'apprentissage, ces trois fonctions de régularisation seront simplement estimées par moyenne empirique sur chaque *minibatch*, puis ajoutées au coût de classification, voir Équation (2).

3 Sélection de la fonction de régularisation

Pour isoler l'effet de la méthode de régularisation proposée, nous comparons les trois fonctions de régularisation sur un système *end-to-end* constitué d'un seul réseau de neurones, avec le corpus OpenSAD15.

3.1 Architecture du système *end-to-end*

L'identification de la langue peut être directement réalisée avec un réseau de neurones convolutionnel (Lozano-Diez *et al.*, 2015). Nous utilisons une architecture similaire décrite dans le Tableau 1. Les *features* d'entrée de notre système sont des MFCC de dimension 12, calculés pour des trames de 10 ms. Nous réalisons la classification en utilisant directement les probabilités *a posteriori* renvoyées par la couche de sortie pour chaque langue. Le système est entraîné et évalué avec des segments de parole de trois secondes.

3.2 Le corpus OpenSAD 2015

Pour étudier l'effet du changement de canal de transmission, nous avons réalisé nos expériences préliminaires sur quatre langues du corpus OpenSAD15 (NIST, 2016) : anglais, arabe, pashto et urdu.

TABLE 1 – Architecture du réseau de neurones convolutionnel

Convolutions selon l'axe temporel			
nom de la couche	taille du noyau / du *max pooling*	nombre de filtres	fonction d'activation
conv. 1	5 / 2	1024	ReLU
conv. 2	5 / 2	1024	ReLU
conv. 3	5 / 2	128	ReLU
Agrégation statistique des moyennes et écarts-types (pooling)			
dimension de sortie : $2 \times 128 = 256$			
Couches connectées			
nom de la couche	dimension		fonction d'activation
fc. 1	256×128		ReLU
fc. 2	128×4		Softmax

Il s'agit d'un corpus créé à partir de conversations téléphoniques, canal *src* du corpus, qui ont ensuite été transmises par six systèmes radio différents : B, F, G (UHF), E (VHF), D et H (HF).

Afin d'éviter un biais lors de l'apprentissage, nous n'utilisons que la moitié des données d'entraînement. La moitié des fichiers audio d'origine sont utilisés pour le domaine source (canal *src*) et l'autre moitié, correspondant à des phrases différentes, est utilisée pour les domaines cibles (canaux radio). De cette façon, le même contenu linguistique n'est pas présent sur les deux domaines lors de l'apprentissage.

3.3 Résultats des expériences préliminaires

Nous réalisons différents entraînements du réseau de neurones convolutionnel sur le corpus OpenSAD15. La performance de chacun des systèmes sur les canaux d'intérêt est présentée dans le Tableau 2. Les performances sont mesurées avec un *Equal Error Rate* (EER) moyen pour des segments de parole de trois secondes. Un EER est calculé pour chacune des quatre langues du corpus et le score obtenu est la moyenne arithmétique des taux de chaque langue.

TABLE 2 – Résultats en taux d'égale erreur de différentes méthodes d'entraînement du réseau convolutionnel pour le corpus OpenSAD15 (segments de 3 secondes). Le domaine source est le canal téléphonique (*src*).

Méthode d'apprentissage	EER sur le domaine cible (%)					
	B	*D*	*E*	*F*	*G*	*H*
supervisé sur *source*	57	52	48	51	30	50
supervisé sur *cible*	**18**	15	19	15	14	22
distance moyennes	53	44	38	35	12	41
deep CORAL	32	32	26	18	11	20
MMD	19	**11**	**16**	**13**	**9**	**18**

Le réseau est d'abord entraîné avec des données du canal téléphonique. Ce système, qui obtient un EER moyen de 8% sur le canal *src* est totalement inopérant sur les canaux cibles. Cependant, lorsqu'un système est entraîné de façon supervisée sur chacun des canaux cibles, alors nous obtenons un EER moyen compris entre 14% et 22%.

Les trois dernières lignes du Tableau 2 présentent les performances de l'apprentissage avec chacune des fonctions de régularisation proposées, en utilisant des données annotées du canal *src* et des données non annotées sur le domaine cible. Le paramètre λ (ainsi que σ^2 pour la *MMD*) est sélectionné pour chaque fonction de coût en fonction de la performance obtenue sur un ensemble de validation. Les résultats de tous les domaines cibles sont cohérents : les méthodes de régularisation permettent d'améliorer les EER moyens sur le domaine cible par rapport à un apprentissage sur le domaine source. Une hiérarchie claire apparaît : la *MMD* avec noyau gaussien est plus efficace que *deep CORAL*, qui est elle-même supérieure à la contrainte sur la distance entre les moyennes. Ce résultat signifie que, pour supprimer la distorsion due au changement de canal, le système ne peut se limiter aux deux premiers moments des distributions mais doit prendre en compte une géométrie plus complexe.

Pour cinq des six domaines cibles testés, la régularisation avec la fonction de coût *MMD* permet d'obtenir une meilleure performance sur le domaine cible que l'apprentissage supervisé sur ce domaine, alors même que l'entraînement n'a pas utilisé de données annotées du domaine cible.

4 Application à un système à l'état de l'art

Les expériences préliminaires ont permis de sélectionner la fonction de régularisation basée sur la *Maximum Mean Discrepancy* pour l'adaptation d'un réseau de neurones convolutionnel. Nous appliquons donc cette méthode d'apprentissage à un système d'identification de la langue correspondant à l'état de l'art pour cette tâche.

4.1 Architecture du système

Un système moderne de reconnaissance de la langue (Snyder *et al.*, 2018; Plchot *et al.*, 2018) est en règle générale formé de trois modules : un extracteur de représentations pour des trames localisées dans le temps, un extracteur de représentations pour l'ensemble du segment audio et un classifieur final.

Dans notre système, le premier module extrait des *stacked multilingual bottleneck features*. Il s'agit des activations d'une couche intermédiaire (*bottleneck*) d'un réseau de neurones ayant été entraîné à reconnaître des triphones pour dix-sept langues du corpus Babel. Nous utilisons les réseaux entraînés *BUT/PHONEXIA bottleneck feature extractor* (Fer *et al.*, 2017), ayant donné de bons résultats pour l'évaluation NIST LRE 2017 (Plchot *et al.*, 2018). Ils génèrent des *bottleneck features* de dimension 80, pour chaque trame de 10 ms.

Le deuxième module extrait un vecteur représentatif par segment. C'est un réseau de neurones prenant en entrée la séquence des *bottleneck features* et entraîné de façon supervisée à prédire la langue utilisée dans le segment. Nous utilisons l'architecture du système *x-vector* (Snyder *et al.*, 2018), constituée de cinq couches procédant à des traitements par trame, suivies d'une couche d'agrégation (*pooling*) statistique et de trois couches pleines. Les *x-vectors* issus de cette architecture sont des vecteurs de dimension 512.

Enfin le classifieur final prend en entrée un *x-vector* et produit un score pour chacune des langues cibles. Notre classifieur final est composé d'une *LDA* (*Linear Discriminant Analysis*), utilisée pour réduire la dimension, d'un blanchiment par multiplication matricielle et d'un *SVM* (*Support Vector Machine*).

Nous appliquons la méthode de régularisation basée sur la *Maximum Mean Discrepancy* au réseau *x-vector*, dans le but de produire des *x-vectors* invariants au changement de canal. Pour des systèmes similaires consacrés à la tâche de reconnaissance du locuteur, l'adaptation *model-based* du réseau *x-vector* a permis de réduire la distorsion due à la langue (Rohdin *et al.*, 2019) et aux conditions acoustiques (Bhattacharya *et al.*, 2019), avec des fonctions de coût antagonistes.

4.2 Le corpus RATS

Nous entraînons ce système sur le corpus RATS (Walker & Strassel, 2012). Nous utilisons les livraisons LDC2015S02 et LDC2017S20 qui comptent cinq langues : anglais, arabe, farsi, pashto et urdu. Ce corpus présente les mêmes caractéristiques que le corpus OpenSAD15 qui en est un sous-ensemble. Il contient deux canaux UHF supplémentaires : A et C. Comme pour le corpus OpenSAD15, nous n'utilisons que la moitié du corpus afin qu'un même contenu linguistique ne soit pas présent à la fois sur les domaines source et cible.

L'identification de la langue a été étudiée sur le corpus RATS (Matějka *et al.*, 2014; Lei *et al.*, 2014; Han *et al.*, 2013) avec la livraison LDC2018S10, contenant également cinq langues : arabe, dari, farsi, pashto et urdu. Pour des segments de trois secondes et pour tous les canaux, le meilleur EER moyen obtenu est de 9.59% (Matějka *et al.*, 2014).

4.3 Expériences

Tout d'abord, nous entraînons le système sur tous les canaux, nous obtenons un EER moyen de 9.36% comparable à l'état de l'art sur le corpus RATS.

Ensuite nous procédons à l'entraînement du réseau *x-vector* de façon supervisée sur le domaine source (canal téléphonique) puis sur chacun des canaux cibles. Enfin, pour chacun des domaines cibles, nous appliquons la régularisation basée sur la *Maximum Mean Discrepancy*. Rappelons que le réseau de neurones n'est pas utilisé directement pour réaliser la classification mais pour extraire un *x-vector* représentatif du segment audio. Pour évaluer les propriétés des *x-vectors* ainsi extraits, nous réalisons plusieurs systèmes en entraînant le classifieur final avec des données annotées, soit du domaine source, soit du domaine cible. Les performances de chacun de ces systèmes sont présentées dans le Tableau 3.

TABLE 3 – Résultats en taux d'égale erreur de différentes méthodes d'entraînement du réseau *x-vector* et du classifieur final pour huit canaux radio du corpus RATS (segments de 3 secondes)

Méthode d'entraînement		EER moyen sur le domaine cible (%)							
x-vector	classifieur final	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
supervisé sur *source*	supervisé sur *source*	50,2	42,3	34,4	39,6	48,5	45,1	17,4	43,6
supervisé sur *source*	supervisé sur *cible*	15,8	15,0	14,1	14,3	21,8	20,5	9,8	18,8
supervisé sur *cible*	supervisé sur *cible*	14,6	12,5	12,6	6,7	13,6	13,5	8,6	14,2
MMD	supervisé sur *source*	12,7	10,6	11,7	7,6	13,3	11,9	5,5	12,2
MMD	supervisé sur *cible*	**10,2**	**9,2**	**11,3**	**6,0**	**11,8**	**10,3**	**5,1**	**10,0**

Notons d'abord qu'un système entraîné sur le domaine source, qui obtient pourtant un EER moyen de 6.0% sur ce domaine, atteint une très mauvaise performance sur les canaux radio. À l'opposé,

l'entraînement supervisé du système sur le domaine cible atteint des EER moyens compris entre 6.7% et 14.6%. D'autre part, nous observons que l'entraînement supervisé du classifieur final sur le domaine cible avec des *x-vectors* entraînés sur le domaine source (ligne 2 du Tableau 3) ne suffit pas à atteindre la performance d'un système totalement entraîné sur le domaine cible (ligne 3). Ce constat justifie la nécessité de développer une méthode d'adaptation de domaine pour le réseau *x-vector*.

La régularisation du réseau *x-vector* avec la *Maximum Mean Discrepancy* est un succès. Les *x-vectors* produits par ce réseau ont acquis une robustesse au changement de canal puisqu'un classifieur final entraîné sur le domaine source avec ces *x-vectors* (ligne 4 du Tableau 3) obtient une bonne performance sur le domaine cible. En fait, pour tous les canaux à l'exception du canal D, l'adaptation de domaine du réseau *x-vector* avec un classifieur final entraîné sur le domaine source est plus performante que l'entraînement de tout le système sur le domaine cible (ligne 3). Ce résultat confirme nos expériences préliminaires : non seulement les valeurs de sortie mais aussi les activations de la couche *x-vector* du réseau acquièrent une invariance au domaine grâce à la méthode de régularisation.

Finalement, l'entraînement d'un classifieur final sur le domaine cible avec les *x-vectors* obtenus par régularisation (ligne 5 du Tableau 3) conduit à des EER moyens significativement inférieurs à un système totalement entraîné sur le domaine cible (ligne 3). C'est donc que la régularisation a un intérêt pour améliorer la qualité des *x-vectors*, même lorsqu'on dispose d'étiquettes de langues sur le domaine cible. D'autre part, pour ces *x-vectors* régularisés avec la *MMD*, un entraînement du classifieur final sur le domaine cible (ligne 5) améliore la performance de classification par rapport à un entraînement sur le domaine source (ligne 4). Les *x-vectors* ne sont donc pas totalement invariants entre les domaines et notre approche pourrait être combinée avec une adaptation du classifieur final.

5 Conclusion

Nous avons introduit une méthode d'adaptation de domaine non supervisée d'un réseau de neurones pour un système d'identification de la langue. Cette méthode consiste en une modification de la fonction de coût utilisée lors de l'entraînement du réseau de neurones par l'ajout d'un terme de régularisation. Par des expériences préliminaires avec un réseau de neurones convolutionnel, nous avons sélectionné une fonction de coût basée sur la *Maximum Mean Discrepancy*. Dans un second temps, nous avons appliqué cette approche à un système récent d'identification de la langue, constitué d'un extracteur de *features*, d'un extracteur de *x-vectors* et d'un classifieur final.

Les résultats démontrent l'efficacité de la méthode proposée pour prendre en compte la distorsion due au canal de transmission du signal. Lorsque la régularisation est appliquée au réseau x-vector, elle produit des vecteurs ayant acquis une robustesse au changement de domaine et permettant donc le transfert d'un apprentissage entre le domaine source et le domaine cible. De plus, la régularisation proposée améliore notablement la capacité de discrimination des *x-vectors* ainsi produits par rapport à un apprentissage supervisé. Elle est donc pertinente même dans le cas où on disposerait de données annotées sur le domaine cible.

Références

BEN-DAVID S., BLITZER J., CRAMMER K., KULESZA A., PEREIRA F. & VAUGHAN J. W. (2010). A theory of learning from different domains. In *Machine learning*, volume 79, p. 151–175 : Springer.

BHATTACHARYA G., ALAM J. & KENNY P. (2019). Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training. In *Proc. ICASSP*, p. 6041–6045.

BOUSQUET P.-M. & ROUVIER M. (2019). On robustness of unsupervised domain adaptation for speaker recognition. In *Proc. INTERSPEECH*, p. 2958–2962.

FER R., MATĚJKA P., GRÉZL F., PLCHOT O., VESELÝ K. & ČERNOCKÝ J. H. (2017). Multilingually trained bottleneck features in spoken language recognition. In *Computer Speech & Language*, volume 46, p. 252–267 : Elsevier.

FEYDY J., SÉJOURNÉ T., VIALARD F.-X., AMARI S.-I., TROUVÉ A. & PEYRÉ G. (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proc. The Twenty-second International Conference on Artificial Intelligence and Statistics*, p. 2681–2690.

GANIN Y., USTINOVA E., AJAKAN H., GERMAIN P., LAROCHELLE H., LAVIOLETTE F., MARCHAND M. & LEMPITSKY V. (2016). Domain-adversarial training of neural networks. In *The Journal of Machine Learning Research*, volume 17, p. 2096–2030.

HAN K. J., GANAPATHY S., LI M., OMAR M. K. & NARAYANAN S. (2013). TRAP language identification system for RATS phase II evaluation. In *Proc. INTERSPEECH*, p. 1502–1506.

LEI Y., FERRER L., LAWSON A., MCLAREN M. & SCHEFFER N. (2014). Application of convolutional neural networks to language identification in noisy conditions. In *Proc. Odyssey*, volume 41, p. 1–8.

LIN W.-W., MAK M.-W., LI L. & CHIEN J.-T. (2018). Reducing domain mismatch by maximum mean discrepancy based autoencoders. In *Proc. Odyssey*, p. 162–167.

LONG M., CAO Y., WANG J. & JORDAN M. I. (2015). Learning transferable features with deep adaptation networks. In *Proc. ICML 2015*, p. 97–105.

LOZANO-DIEZ A., ZAZO CANDIL R., GONZÁLEZ DOMÍNGUEZ J., TOLEDANO D. & GONZÁLEZ-RODRÍGUEZ J. (2015). An end-to-end approach to language identification in short utterances using convolutional neural networks. In *Proc. INTERSPEECH*, p. 403–407.

MATĚJKA P., ZHANG L., NG T., MALLIDI S. H., GLEMBEK O., MA J. & ZHANG B. (2014). Neural network bottleneck features for language identification. In *Proc. Odyssey*, p. 299–304.

NIST (2016). Evaluation plan for the NIST open evaluation of speech activity detection (OpenSAD15). In *www.nist.gov/itl/iad/mig/nist-open-speech-activity-detection-evaluation*.

PEYRÉ G. & CUTURI M. (2019). Computational optimal transport. In *Foundations and Trends® in Machine Learning*, volume 11, p. 355–607 : Now Publishers, Inc.

PLCHOT O., MATĚJKA P., NOVOTNÝ O., CUMANI S., LOZANO-DIEZ A., SLAVICEK J., DIEZ M., GRÉZL F., GLEMBEK O., MOUNIKA K. V., SILNOVA A., BURGET L., ONDEL L., KESIRAJU S. & ROHDIN J. (2018). Analysis of BUT-PT submission for NIST LRE 2017. In *Proc. Odyssey*, p. 47–53.

ROHDIN J., STAFYLAKIS T., SILNOVA A., ZEINALI H., BURGET L. & PLCHOT O. (2019). Speaker verification using end-to-end adversarial language adaptation. In *Proc. of ICASSP*, p. 6006–6010.

SNYDER D., GARCIA-ROMERO D., MCCREE A., SELL G., POVEY D. & KHUDANPUR S. (2018). Spoken language recognition using x-vectors. In *Proc. Odyssey*, p. 105–111.

SUN B. & SAENKO K. (2016). Deep CORAL : Correlation alignment for deep domain adaptation. In *Proc. ECCV 2016*, p. 443–450 : Springer.

WALKER K. & STRASSEL S. (2012). The RATS radio traffic collection system. In *Proc. Odyssey*, p. 291–297.