Cross-Domain Adaptation of Spoken Language Identification for Related Languages: The Curious Case of Slavic Languages

Badr M. Abdullah[1,2], Tania Avgustinova[2], Bernd Möbius[2], Dietrich Klakow[1,2]

[1]Spoken Language Systems (LSV) [2]Department of Language Science and Technology
Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany
{babdullah|dietrich}@lsv.uni-saarland.de, {avgustinova|moebius}@lst.uni-saarland.de

Abstract

State-of-the-art spoken language identification (LID) systems, which are based on end-to-end deep neural networks, have shown remarkable success not only in discriminating between distant languages but also between closely-related languages or even different spoken varieties of the same language. However, it is still unclear to what extent neural LID models generalize to speech samples with different acoustic conditions due to domain shift. In this paper, we present a set of experiments to investigate the impact of domain mismatch on the performance of neural LID systems for a subset of six Slavic languages across two domains (read speech and radio broadcast) and examine two low-level signal descriptors (spectral and cepstral features) for this task. Our experiments show that (1) out-of-domain speech samples severely hinder the performance of neural LID models, and (2) while both spectral and cepstral features show comparable performance within-domain, spectral features show more robustness under domain mismatch. Moreover, we apply unsupervised domain adaptation to minimize the discrepancy between the two domains in our study. We achieve relative accuracy improvements that range from 9% to 77% depending on the diversity of acoustic conditions in the source domain.

Index Terms: spoken language identification, Slavic languages, deep neural networks, unsupervised domain adaptation

1. Introduction

Spoken language identification, henceforth LID, is the problem of determining the identity of the language in a spoken utterance [1]. In today's globalized world, LID systems can facilitate a wide range of cross-lingual speech and communication technologies such as spoken language translation [2, 3, 4] and multilingual spoken document retrieval [5]. Furthermore, LID-aware transfer of language resources has been shown to be effective for multilingual ASR in low-resource settings [6, 7, 8, 9].

Earlier work has addressed the LID task using the so-called phonotactic approach. In this paradigm, the acoustic signal is first transduced into a sequence of discrete symbols (e.g., phones), then probabilistic models are utilized to obtain language likelihoods [10, 11]. This approach has been outperformed by acoustic approaches that are based on Gaussian Mixture Models (GMMs) and the i-vector framework which has been applied to speaker and language identification [12, 13, 14, 15]. Currently, end-to-end deep neural networks (DNNs) are predominant for LID and outperform GMMs, especially for short utterances [16, 17, 18, 19, 20].

The findings of the popular language guessing game, the Great Language Game [21], have shown that discriminating between closely-related languages is a difficult task for humans. On the other hand, neural LID models have shown striking performance discriminating between spoken varieties of Arabic [22, 23, 18], Slavic languages [16], and languages in accented speech samples from multilingual speakers [24]. For instance, the best neural LID model in [16] has reported an error rate as low as 1.2% when discriminating between 11 Slavic languages. Generally speaking, the impressive performance of DNN-based LID reported in the literature gives the impression that LID is almost a solved problem.

However, previous works have developed their models using disjoint splits of the same dataset where the training and evaluation samples have similar, if not identical, acoustic conditions (i.e., same domain). The impact of *dataset-bias* [25] on LID robustness has not yet been investigated with a systematic evaluation across datasets. In this paper we aim to fill this gap and focus on the challenging case of LID for short utterances of related languages (i.e., Slavic languages) in a cross-domain setting. We investigate the following questions:

- **RQ1** To what degree do neural LID models for related languages generalize to another domain with different acoustic conditions?

- **RQ2** Are different low-level speech features equally robust under domain mismatch?

- **RQ3** Can we adapt LID models to a new domain without using labelled data in the new domain? If yes, what are the factors that affect the adaptability of the model?

To address these research questions, we conduct a series of LID experiments with datasets from two domains: (1) Read speech recordings from the Slavic subset of the GlobalPhone speech database [26], and (2) Slavic broadcast recordings collected and distributed in [16, 27] for LID (**RQ1**). We also compare the performance of spectral (MFSCs) and cepstral (MFCCs) speech features within- and across-domain (**RQ2**). Finally, we apply adversarial domain confusion [28] to adapt our model to a target domain, analyze predictions from the adapted model, and visualize its representations compared to the baseline (**RQ3**).

2. LID with Deep Neural Networks

2.1. Problem Definition

We define the LID task as a discriminative sequence classification problem. First, a variable-length utterance is transformed by an acoustic front-end into a sequence of acoustic observations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_t \in \mathbb{R}^k$ is a low-level feature vector at timestep t. Given a sequence $\mathbf{X}$, the goal is to predict the spoken language $\hat{y}$. Using a deep neural network as a classification model, the LID problem can be defined as

$$\hat{y} = \underset{y \in \mathcal{Y}}{\arg\max}\, P(y \mid \mathbf{X};\, \boldsymbol{\theta}) \qquad (1)$$

where $\mathcal{Y}$ is a finite set of languages, $\boldsymbol{\theta}$ is the model's parameters learned in a supervised approach, and $P(y|\mathbf{X};\boldsymbol{\theta})$ represents a posterior probability of the language label y.

2.2. LID Model Overview

Our LID model consists of a 1D 3-layer convolutional network followed by 2-layer fully-connected feed-forward network as schematized in Fig. 1(a). We refer to the convolutional block as a high-level feature extractor G_f that transforms the input sequence $\mathbf{X}$ into a D-dimensional feature vector $\mathbf{f} \in \mathbb{R}^D$, i.e. $\mathbf{f} = G_f(\mathbf{X};\boldsymbol{\theta}_f)$. Then, the feed-forward layers transform $\mathbf{f}$ into a logit vector $\hat{\mathbf{y}} \in \mathbb{R}^{|\mathcal{Y}|}$ via a series of non-linear transformations, i.e. $\hat{\mathbf{y}} = G_y(\mathbf{f};\boldsymbol{\theta}_y)$, followed by a softmax function that maps $\hat{\mathbf{y}}$ into a probability distribution over the language space. We refer to the fully-connected block of the model G_y as a language classifier. The parameters of the network $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_y$ are learned jointly in an end-to-end approach given a dataset $\mathcal{D}_\mathcal{S} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N_\mathcal{S}}$ of $N_\mathcal{S}$ labelled samples in one domain. The objective function is to minimize

$$J(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y) = \sum_{(\mathbf{X}_i, y_i) \in \mathcal{D}_\mathcal{S}} L_y\Big(G_y\big(G_f(\mathbf{X}_i;\boldsymbol{\theta}_f);\boldsymbol{\theta}_y\big), y_i\Big) \quad (2)$$

where L_y is the loss of the language classifier.

2.3. Domain Adaptive LID

In this paper, we explore a well-established domain adaptation technique that has been successfully applied to many vision and speech recognition problems [28, 29, 30]. This technique aims to minimize the discrepancy between two domains given a dataset $\mathcal{D}_\mathcal{T} = \{\mathbf{X}_i\}_{i=1}^{N_\mathcal{T}}$ of $N_\mathcal{T}$ unlabelled samples in the target domain, in addition to the source labelled samples $\mathcal{D}_\mathcal{S}$.

To improve the LID model's out-of-domain generalization, the feature representations emerging from the model should be both language-discriminative and domain-invariant. This objective can be achieved if the model is encouraged during training to build up representations that are good predictors of the spoken language but do not encode domain-related information. To this end, a fully-connected feed-forward block G_d is added to the network to predict the domain given $\mathbf{f}$ (see Fig. 1(b)). We view G_d as a domain classifier with a separate set of parameters $\boldsymbol{\theta}_d$ which are learned by exploiting the domain labels of source and target samples. That is, each training sample in the source domain $(\mathbf{X}_i, y_i)$ is augmented with a domain label $d_i = 0$, while each training sample in the target domain $\mathbf{X}_j$ is augmented with a domain label $d_j = 1$. We seek the parameters $\boldsymbol{\theta}_d$ that minimize the loss of the domain classifier. On the other hand, the feature extractor G_f is trained such that $\mathbf{f}$ is uninformative for the domain classifier. Thus, we seek the parameters $\boldsymbol{\theta}_f$ that maximize the domain classifier loss. This procedure is an instance of adversarial learning where different blocks in the network are trained with competing objectives. The overall objective function is to minimize

$$J(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y, \boldsymbol{\theta}_d) = \sum_{(\mathbf{X}_i, y_i) \in \mathcal{D}_\mathcal{S}} L_y\Big(G_y\big(G_f(\mathbf{X}_i;\boldsymbol{\theta}_f);\boldsymbol{\theta}_y\big), y_i\Big)$$
$$- \lambda \sum_{(\mathbf{X}_i, d_i) \in (\mathcal{D}_\mathcal{S} \cup \mathcal{D}_\mathcal{T})} L_d\Big(G_d\big(G_f(\mathbf{X}_i;\boldsymbol{\theta}_f);\boldsymbol{\theta}_d\big), d_i\Big) \quad (3)$$

where L_y is the loss of the language classifier, L_d is the loss of the domain classifier, and λ is a parameter that controls the
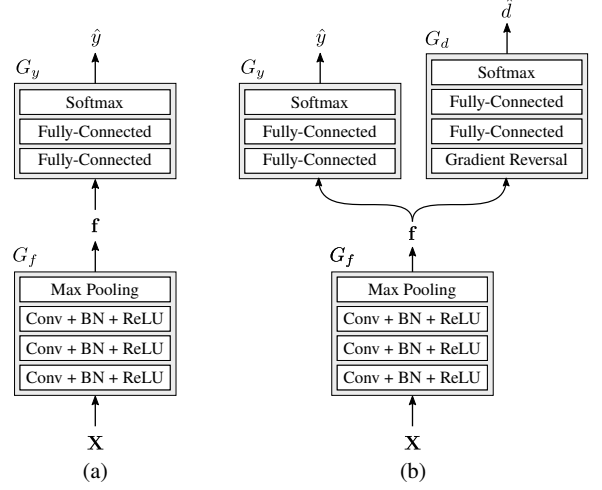

Figure 1: *A schematic view of our models: (a) Non-adaptive, and (b) Domain adaptive LID with adversarial classifier G_d.*

contribution of the domain classifier's loss to the overall loss. In practice, this adversarial loss is realized with a special layer that reverses the direction of the gradient signal coming from the domain classifier's loss into the feature extractor during back-propagation, which is referred to as a gradient reversal layer. We refer the reader to the original paper for a detailed overview of the training procedure [28].

3. Experimental Setup

3.1. Datasets for Slavic LID

GlobalPhone Read Speech (GRS) We use the Slavic portion of the multilingual GlobalPhone speech database [26] which includes read speech recordings from native speakers of six Slavic languages: Bulgarian (BUL), Croatian (HRV), Czech (CZE), Polish (POL), Russian (RUS), and Ukrainian (UKR). The utterances vary in length and quality across languages. We set the minimum utterance length to 3 seconds and segment longer utterances into non-overlapping 3-second speech segments. Our final training subset consists of 8,000 utterances per language. We use the same splits as in [31].

Radio Broadcast Speech (RBS) A large collection of Slavic recordings were collected by harvesting online radio broadcasts in [16, 27]. The original dataset contains recordings for 11 Slavic languages. We use the same subset of six languages as in the GRS dataset. The extracted utterances are either segments of professional news reports or of spontaneous speech during discussions. Occasionally, the utterances include background music and different sorts of acoustic noise. We sample 8,000 and 500 utterances per language from the training split as our training and validation sets, respectively. This dataset does not include any speaker IDs. Thus, we cannot confirm whether training and evaluation speakers are disjoint.

3.2. Low-level Feature Extraction

In our experiments, we use the first 13 coefficients of MFSCs and MFCCs, with the zeroth coefficient being the average frame energy, as low-level speech features. While previous works usually refer to MFSCs as mel-filterbanks [18], we use the term MFSCs to refer to mel-frequency spectral features that are cor-

related [32]. Since both datasets in our study are sampled at 16 kHz, we extract frames of 400 samples with 160 samples overlap, which corresponds to 25 ms and 10 ms, respectively. We normalize the features to have utterance-level zero mean and unit variance.

3.3. Model Architecture and Hyperparameters

CNN Architecture We use 1D 3-layer convolution over the temporal dimension with 128, 256, and 512 filters and widths of 5, 10, and 10 for each layer and keep stride step at 1. We apply batch normalization and ReLU non-linearity following each convolutional operation. We apply max pooling to downsample the representation only at the end of the convolution block. For the language classifier, we use 2 fully-connected layers (512 $\rightarrow$ 512 $\rightarrow$ 6) before the softmax for both the non-adapted and the adapted LID models.

Domain-Adaptive Model For our adapted LID models, we use a 3-layer feed-forward network (512 $\rightarrow$ 1024 $\rightarrow$ 1024 $\rightarrow$ 2) as the domain classifier. For the adaptation factor λ, we use a gradually increasing value $\in [0, 1]$ to suppress the noise from the feature extractor during the initial phase of the training procedure. We experiment with two variants of the domain-adaptive model: (1) DA-LID I: an identical configuration to [28], where the convolutional block of the model is considered as the feature extractor, and (2) DA-LID II: we consider the feature extractor as the convolutional block as well as the first layer of the fully-connected block; thus, the reversed gradient signal from the domain classifier is back-propagated into all layers of the network except the final layer before the softmax of the language classifier.

Training Details We use cross-entropy loss for both L_y and L_d. The ADAM optimizer is used with learning rate of 0.001. We train our models with a batch size of 256 for 50 epochs and observe the validation performance during training.

Implementation We use PyTorch to implement the LID models and make our code publicly available.[1]

4. Experimental Results

We now present and discuss the results of our experiments. To make the results comparable across datasets and prevent undesirable effects due to utterance length mismatch, we train and evaluate each of our LID models on 3-second utterances. Since the GRS evaluation data is imbalanced, we use balanced accuracy [33] as our evaluation metric to obtain a better estimate of the model performance. We observe that balanced accuracy scores highly correlate with equal error rate (EER) and average cost (C_{avg}), which we do not report for the sake of conciseness.

4.1. Cross-Domain Evaluation

Table 1 presents the results of the cross-domain evaluation on both datasets without adaptation. Even though our LID models are not heavily regularized, the in-domain performance is always above 95%, while MFSC and MFCC features yield a comparable performance. On the other hand, out-of-domain (OOD) evaluation shows a considerable drop in accuracy in each cross-domain setting. It is interesting to observe that the drop in accuracy is more pronounced for MFCC features, and MFSCs seem to be more robust under domain shift. The impact of domain shift is more pronounced in the GRS $\rightarrow$ RBS direction.

---

[1]https://github.com/uds-lsv/da-lang-id

Table 1: *Cross-domain evaluation of LID models in acc. (%).*

Dataset		Evaluation		
		In-domain	OOD	Δ
GRS	MFSCs	95.27	**43.27**	-54.55
	MFCCs	95.81	37.80	-60.54
RBS	MFSCs	95.34	**54.01**	-43.35
	MFCCs	95.00	50.97	-46.35

4.2. Adaptation Results

In our adaptation experiments, we investigate two transfer tasks; GRS $\rightarrow$ RBS and RBS $\rightarrow$ GRS. The results are shown in Table 2. The adapted models consistently improve the accuracy compared to the source-only non-adapted baseline with both features and in both directions. Our DA-LID II model yields the best results, which suggests that the domain discrepancy is present not only in the convolutional layers, but also in the fully-connected layers that are more distant from the input. We present and discuss the results for both directions.

RBS $\rightarrow$ GRS Both adapted models yield significant improvements over non-adapted models. The MFCC-based DA-LID II boosts OOD accuracy from 50.94% to 90.56% with a relative accuracy gain of 77.7%.

GRS $\rightarrow$ RBS Even though adapted models improve over the baseline, the improvements in this direction are less impressive than what is observed in the RBS $\rightarrow$ GRS direction. Our MFSC-based DA-LID II model performs best and improves the accuracy by 16.6% compared to the baseline.

4.3. Discussion

The performance gap between the two directions in our experiments seemed surprising at the beginning. In retrospective, this should not be surprising as the two directions are not equally challenging. The RBS dataset is more diverse in terms of the number of unique speakers and background noise. An LID model trained on the RBS dataset has learned to extract language ID features from noisy speech signals, thus it is expected to be more generic and perform well on clean speech signals even under domain shift. This finding is consistent with what has been reported in the domain adaptation literature on how source domain diversity affects adaptability of the model to new domains [28]. On the other hand, if the model has not been exposed to noisy speech signals during training, it is unlikely to perform well on noisy signals even if the representation discrepancy has been minimized, which is the case in the GRS $\rightarrow$ RBS direction. This suggests that alternative adversarial training procedures that add noise to the input representation could

Table 2: *OOD performance of adapted models in acc. (%).*

Direction		Adaptation Method		
		None	DA-LID I	DA-LID II
GRS $\rightarrow$ RBS	MFSCs	43.27	47.22 (+09.1)	**50.56** (+16.6)
	MFCCs	37.80	41.77 (+12.6)	44.55 (+17.9)
RBS $\rightarrow$ GRS	MFSCs	54.01	72.12 (+33.5)	86.99 (+61.1)
	MFCCs	50.97	66.50 (+30.5)	**90.56** (+77.7)
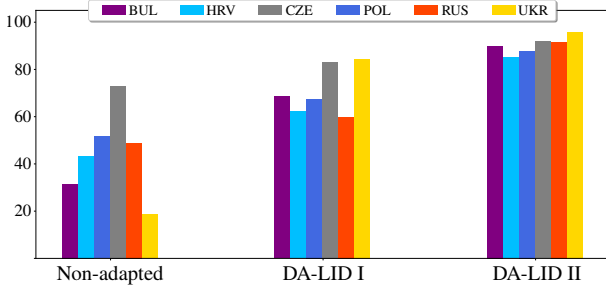
Figure 2: *OOD F_1 score (%) per language of our MFCC-based models in the RBS $\rightarrow$ GRS direction.*

(a) Non-adapted (b) DA-LID I (c) DA-LID II

Figure 3: *t-SNE visualization: (Top) data points are marked by domain, red points correspond to source samples while green points corresponds to target samples, and (Bottom) data points are marked by language.*

be explored to encourage the model to transform the noisy input signals into noise-robust representations. Moreover, our experiments show that MFCCs are more sensitive to input variations due to domain shift, thus MFCC-based models in both directions tend to benefit more from adaptation in terms of relative accuracy gain compared to their MFSC-based counterparts, with only one exception case.

5. Adaptive Model Analysis

In this section, we seek to understand why unsupervised adaptation with adversarial training improves OOD performance. We analyze the results of the RBS $\rightarrow$ GRS transfer task to get insights into the factors that lead to the significant improvement.

5.1. Fine-grained Performance Analysis

Fig. 2 shows the performance per language measured by F_1 score. In the non-adapted case, we observe a much higher variance between languages compared to the adapted models. For example, while the non-adapted model achieves up to 70% F_1 on Czech, it drops to 18.6% on Ukrainian, which is slightly better than the chance-level F_1 (16.7%). We inspected the performance on Ukrainian in the other direction and found that the F_1 is even worse than chance-level. We hypothesized that the acoustic conditions of the Czech recordings in the two domains are similar, while the discrepancy is maximal in the case of Ukrainian. To validate this hypothesis, we manually inspected several Ukrainian utterances from the GRS dataset. We found that most utterances are characterized by unnatural pauses and hesitations that distort the speech signal and are uniformly distributed across Ukrainian training and evaluation speakers in the GRS dataset. This effect adds to the discrepancy due to domain shift since RBS utterances are more naturally flowing speech than the read speech from the GRS dataset, despite the occasional background noise. In particular, this effect creates abnormal patterns that hinder non-adapted LID performance in two ways: (1) if these patterns are not uniformly distributed across languages and observed during training, the network exploits them as shortcuts to discriminate between languages, and (2) if these patterns are encountered during OOD inference, the distorted signal causes a failure because the model has not been exposed to such patterns during training. Both cases lead to poor OOD generalization when training on a single-domain dataset. However, since these patterns are only present in one dataset, they are good predictors of the domain. Therefore, adversarial training with domain confusion prevents the models from exploiting such dataset-specific artifacts, which consistently yields a better OOD generalization. The advantage of ad-
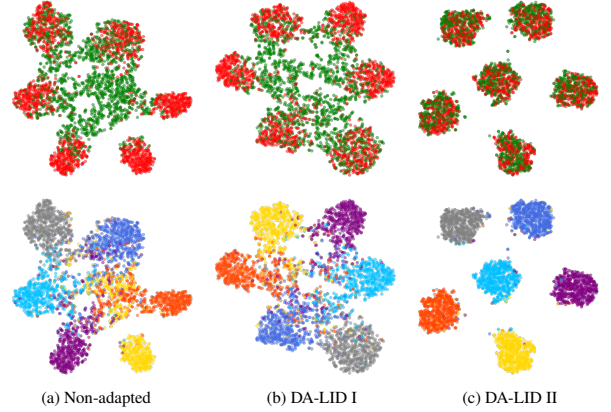
versarial training is demonstrated in Fig. 2. Our adapted model boosts the F_1 score on Ukrainian from 18.6% to 96.0%, which is surprisingly the highest in this direction.

5.2. Visualizing the Representations

In Fig. 3, we visualize the representations using the t-SNE algorithm [34]. We sample a set of 1800 data points from each domain and obtain the representations from the last hidden layer of the MFCC-based LID models: (a) source-only non-adapted LID, (b) DA-LID I, and (c) DA-LID II. Fig. 3 shows how adaptation aligns the distributions of the extracted representations from the two domains.

6. Conclusions

We have investigated the problem of spoken language identification for closely-related languages in a cross-domain setting, using deep convolutional neural networks as discriminative models. While our experiments have confirmed that they perform very well within-domain, our cross-domain evaluation has revealed that neural models poorly generalize to a novel dataset with acoustic conditions that differ from those that have been observed during training. To improve the robustness of our models against domain mismatch, we have applied unsupervised domain adaptation with gradient reversal and shown that our adaptive models generalize better across domains. Our analysis has shown that adversarial training prevents the model from exploiting dataset-specific artifacts, thus leading to better out-of-domain generalization. We have identified the diversity of the speech samples in the source domain as the major factor that affects the adaptability of the model to a new domain. Given a diverse source dataset, our adaptive models achieved relative accuracy improvements of up to 77.7%.

7. Acknowledgements

8. References

[1] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[2] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313, 2000.

[3] C. Fügen, A. Waibel, and M. Kolss, "Simultaneous translation of lectures and speeches," *Machine translation*, vol. 21, no. 4, pp. 209–252, 2007.

[4] S. Bangalore, V. K. Rangarajan Sridhar, P. Kolan, L. Golipour, and A. Jimenez, "Real-time incremental speech-to-speech translation of dialogs," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 437–445.

[5] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 39–49, 2008.

[6] M. Müller, S. Stüker, and A. Waibel, "Language adaptive dnns for improved low resource speech recognition." in *Interspeech*, 2016, pp. 3878–3882.

[7] M. Müller and A. Waibel, "Using language adaptive deep neural networks for improved multilingual speech recognition," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, 2015.

[8] Q. B. Nguyen, J. Gehring, M. Müller, S. Stüker, and A. Waibel, "Multilingual shifting deep bottleneck features for low-resource asr," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5607–5611.

[9] A. Cutler, Y. Zhang, E. Chuangsuwanich, and J. R. Glass, "Language id-based training of multilingual stacked bottleneck features," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[10] L. F. Lamel and J.-L. Gauvain, "Language identification using phone-based acoustic likelihoods," in *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. IEEE, 1994, pp. I–293.

[11] H. Li and B. Ma, "A phonotactic language model for spoken language identification," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 2005, pp. 515–522.

[12] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[13] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth annual conference of the international speech communication association*, 2011.

[14] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Twelfth annual conference of the international speech communication association*, 2011.

[15] H. Su and S. Wegmann, "Factor analysis based speaker verification using asr." in *Interspeech*, 2016, pp. 2223–2227.

[16] L. Mateju, P. Cerva, J. Zdánský, and R. Safarik, "Using deep neural networks for identification of slavic languages from acoustic signal." in *Interspeech*, 2018, pp. 1803–1807.

[17] P. Shen, X. Lu, S. Li, and H. Kawai, "Feature representation of short utterances based on knowledge distillation for spoken language identification." in *Interspeech*, 2018, pp. 1813–1817.

[18] S. Shon, A. Ali, and J. Glass, "Convolutional neural network and language embeddings for end-to-end dialect recognition," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 98–104.

[19] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 5337–5341.

[20] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[21] H. Skirgård, S. G. Roberts, and L. Yencken, "Why are some languages confused for others? investigating data from the great language game," *PloS one*, vol. 12, no. 4, 2017.

[22] A. E. Bulut, Q. Zhang, C. Zhang, F. Bahmaninezhad, and J. H. Hansen, "Utd-crss submission for mgb-3 arabic dialect identification: Front-end and back-end advancements on broadcast speech," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 360–367.

[23] G. Gelly, J.-L. Gauvain, L. Lamel, A. Laurent, V. B. Le, and A. Messaoudi, "Language recognition for dialects and closely related languages." in *Odyssey*, 2016, pp. 124–131.

[24] A. Titus, J. Silovsky, N. Chen, R. Hsiao, M. Young, and A. Ghoshal, "Improving language identification for multilingual speakers," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8284–8288.

[25] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.

[26] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8126–8130.

[27] J. Nouza, R. Safarik, and P. Cerva, "Asr for south slavic languages developed in almost automated way," *Interspeech 2016*, pp. 3868–3872, 2016.

[28] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015, pp. 1180–1189.

[29] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition." in *Interspeech*. San Francisco, CA, USA, 2016, pp. 2369–2372.

[30] Z. Meng, Z. Chen, V. Mazalov, J. Li, and Y. Gong, "Unsupervised adaptation with domain separation networks for robust speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 214–221.

[31] M. Y. Tachbelie, A. Abulimiti, S. T. Abate, and T. Schultz, "Dnn-based speech recognition for globalphone languages," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8269–8273.

[32] A. Mohamed, "Deep neural network acoustic models for asr," Ph.D. dissertation, University of Toronto, 2014.

[33] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3121–3124.

[34] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.