Metric learning loss functions to reduce domain mismatch in the x-vector space for language recognition

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 language recognition systems are based on discriminative embeddings called *x-vectors*. Channel and gender distortions produce mismatch in such *x-vector* space where embeddings corresponding to the same language are not grouped in an unique cluster. To control this mismatch, we propose to train the *x-vector* DNN with metric learning objective functions. Combining a classification loss with the metric learning n-pair loss allows to improve the language recognition performance. Such a system achieves a robustness comparable to a system trained with a domain adaptation loss function but without using the domain information. We also analyze the mismatch due to channel and gender, in comparison to language proximity, in the *x-vector* space. This is achieved using the Maximum Mean Discrepancy divergence measure between groups of *x-vectors*. Our analysis shows that using the metric learning loss function reduces gender and channel mismatch in the *x-vector* space, even for languages only observed on one channel in the train set.

Index Terms: language recognition, domain adaptation, domain mismatch, x-vector, embedding, metric learning

1. Introduction

Language recognition systems rely on the extraction of a fixed size embedding to represent a spoken utterance. The commonly used language recognition pipeline involves the computation of frame-level features, the extraction of an utterance-level embedding and the prediction of language scores by a backend classifier. In the recent years, the generative *i-vectors* embeddings [1] have been replaced by discriminative embeddings [2, 3, 4], most populars being the *x-vectors*. In this context, a Deep Neural Network (DNN) is trained to minimize a language recognition classification loss. Then embeddings are extracted from an hidden layer of the network and used as intputs of a backend classifier, eventually after a dimension reduction and normalization. A simple classifier such as a Gaussian Linear Classifier [2] achieves competitive performance, meaning that training of the embedding extractor allows to separate language clusters in the embedding space.

However, *x-vector* embeddings may encode information not relevant to the task. The authors of [5] have demonstrated that several classifiers can be efficiently trained to retrieve information related to channel, transcription, data augmentation strategy or segment duration from speaker identification *x-vectors*. This implies that the structure of the embedding space does not only correspond to the separation of classes enforced by the classification loss but also to other distortion factors contained in the data. Additionally a mismatch between train and test domains can further affect the embedding space and produce a drop in performance, as evidenced by the mismatch between telephone data and audio from videos in the NIST LRE 2017 campaign [6, 3, 7, 8].

One approach to reduce the impact of this mismatch is to apply domain adaptation to the backend classifier. After training of the *x-vector* DNN, a transformation of these embeddings is learned with the objective of reducing domain mismatch [9]. Alternatively parameters of the backend classifier can be adapted to each domain [10, 7]. A second approach is to design the *x-vector* DNN to reduce mismatch in the embedding space. Two components of the DNN have been modified in this approach. First, the temporal pooling layer can be replaced by a learnable dictionary encoding layer which has the ability to learn several statistical modes for each class [11]. Second, a change in the training loss function of the DNN can enforce invariance of the embeddings.

Originally the *x-vector* DNN was trained by minimization of the cross-entropy loss [2, 3, 4]. Refinements of the cross-entropy, taking into account angular proximities on the softmax layer have increased the discrimination capability [11, 12]. Alternatively, domain adaptation loss functions, enforcing invariance between embeddings of two different domains, have effectively reduced the mismatch [13, 14, 15]. Nevertheless domain adaptation loss functions require to focus on one precise mismatch which has to be annotated in the training set.

Our work supports the claim of the authors of [16] who propose metric learning loss functions in order to group *x-vectors* belonging to the same class. Metric learning loss functions have been appplied to speaker recognition *x-vector* DNNs: triplet loss [16, 17], prototypical networks [16], PLDA-like similarity [18]. For language recognition, the triplet loss has been used to train the backend classifier [19] and cosine similarity has been used during training of an LSTM-based language embedding extractor [20]. The superiority of metric learning over domain adaptation approaches is that it does not rely on the definition of a source and a target domain and can reduce mismatch between *a priori* unkwnown domains.

Our main contribution is an analysis of the effect of metric learning loss functions for training a language identification *x-vector* DNN, targetting domain mismatch reduction. In addition to the language recognition performance, we propose to directly measure mismatches in the embedding space, thanks to a divergence measure between groups of embeddings. This measure shows that a metric learning loss function reduces the mismatch between two domains without using domain labels during training. We observe that domain mismatch reduction is not restricted to language classes which have been observed on both the domains in the training set.

In Section 2, we describe a state-of-the-art language recognition system, based on *x-vectors*, and show how mismatch can be measured in the embedding space. In Section 3, metric learning loss functions are defined. Their application to the *x-vector* DNN training is discussed in Section 4.

2. Analysis of an *x-vector* based language recognition system

In this section, we describe an *x-vector* based system [3, 4] for the task of language recognition on the NIST LRE 2011 corpus [21]. Since a significant part of the errors of the system are attributed to a channel mismatch in the embedding space, we introduce language discriminability and channel mismatch as a way to evaluate the quality of language identification embeddings.

2.1. System description

The language recognition system is constituted of the following components:

- a CNN-based speech activity detector, trained on the RATS SAD corpus [22].
- a multililingual stacked bottleneck features extractor. We used the model from [23], trained on 17 languages of the Babel corpus.
- a TDNN-based embedding extractor with the architecture described in [4]. It is trained by minimization of the cross-entropy loss by stochastic gradient descent with 51 languages of our training corpus, with speech segments of 2 to 4 seconds. Embeddings of dimension 512 are extracted from the antepenultimate layer (*segment6* layer in [4]). Following [3, 4], we apply four data augmentation strategies, adding artificial reverberation, noise, music and babble noise.
- a backend classifier, constituted of an LDA for dimension reduction, normalization, and a Gaussian Linear Classifier with shared covariance matrix [2].
- a multi-class calibration strategy [24].

2.2. Corpus description

Evaluation is performed on the test set of the NIST LRE 2011 campaign, constitued of 24 languages and 2 channels: telephone data and broadcast narrow-band speech [21]. Following the evaluation protocol, we trained our model on data from the previous NIST LRE evaluations contained in the LDC releases LDC2006S31, LDC2008S05, LDC2009S05, LDC2009S04, LDC2014S06 and LDC2018S06. The training set contains 51 languages, with data coming either from both or from only one of the two target channels.

One key property of this corpus is that only 7 of the 24 languages of the test set are present in the train set in both channels (see Table 1). This implies that a channel independent class representation is necessary for good generalization to the test set.

2.3. Language recognition metrics

Standard multiclass detection cost functions are called minimum detection cost function (minDCF) and actual detection cost function (actDCF). We also compute the equal error rate (EER). In addition, following the NIST LRE 2011 protocol, we measure detection performance for language pairs. Detection costs are computed with the NIST LRE 2011 values: $P_{Target} = 0.5$ and $C_{Miss} = C_{FA} = 1$.

For every language pair, a detection cost is computed. The overall performance is the average of the costs of the 24 pairs with highest minDCF for segments of 30 seconds. Two metrics are computed: minimum average pair detection cost (minAPD) and actual average pair detection cost (actAPD), depending on whether the detection threshold is chosen to minimize the cost on the test set or chosen by the calibration module [25].

2.4. Domain mismatch in the language embedding space

In this work, we investigate the impact of the training loss of the *x-vector* extractor over the structure of the *x-vector* space. An ideal embedding extractor for language identification should group *x-vectors* according to the language class. Consequently mismatch incurred in the *x-vector* space by distortion factors should be less than the divergence between different languages. Mismatch in the *x-vector* space can be measured with a divergence D between groups of embeddings. In this work we use the Maximum Mean Discrepancy (MMD) based on the kernel $k(x, y) = -\|x - y\|_2$ as a scale free measure of divergence between clouds of points [26]. We use the GPU implementation of the authors of [27].

We will now define language discriminability as well as channel and gender mismatch. For a language L and a condition C - Telephone (T), Broadcast (B), Female (F) or Male (M) - we use the notation $X^{L,C}$ for the set of *x-vectors* corresponding to recordings of these language and condition. Then we define language discriminability $\mathcal{D}_{\text{Language}}(L, C)$ for the language L and condition C as the smallest divergence with a group of *x-vectors* from another language and the same condition:

$$\mathcal{D}_{\text{Language}}(L, C) = \min_{L' \neq L} D(X^{L,C}, X^{L',C}) \tag{1}$$

It has to be compared to the mismatch caused by the distortion factors for embeddings of the language L. Thus we define channel mismatch $\mathcal{D}_{\text{Channel}}(L)$ and gender mismatch $\mathcal{D}_{\text{Gender}}(L)$ for language L as:

$$\mathcal{D}_{\text{Channel}}(L) = D(X^{L,\text{Telephone}}, X^{L,\text{Broadcast}}) \tag{2}$$

$$\mathcal{D}_{\text{Gender}}(L) = D(X^{L,\text{Female}}, X^{L,\text{Male}}) \tag{3}$$

3. Metric learning for x-vector extractor

To reduce the impact of distortion factors in the embedding space, we investigate metric learning for training of the x-vector extractor. Metric learning enables extraction of representations which preserve meaningful distances between samples. For a classification task, it can be viewed as designing representations such that two samples belonging to the same class are closer to each other than two samples from different classes. When representations are embeddings extracted from a deep neural network, deep metric learning can be achieved by using a loss function that measures distances between embeddings.

In this work, we experiment with different loss functions to train a language identification *x-vector* DNN.

3.1. Metric learning loss functions

We evaluate two commonly used metric learning loss functions for training the *x-vector* DNN: triplet loss and n-pair loss.

Triplet loss [28] operates on pairs of samples. It aims at enforcing that pairs of samples from the same class are closer from each other than pairs from different classes. For an embedding extractor $f(.)$, an anchor x, a positive sample x^+ from the same class as x, a negative samples x^- from a different class, the triplet loss is given by:

$$\mathcal{L}_{\text{triplet}} = \max(0, \|f(x) - f(x^+)\|_2^2 - \|f(x) - f(x^-)\|_2^2 + m) \tag{4}$$

Table 1: *Distribution of the languages with respect to their presence in the train and test sets, and the channels in which they occur.*

Presence in the train set	present in the test set in both channels	present in the test set in a single channel only	not present in the test set
present on both channels in the train set	G_A: American English, Farsi, Hindi, Mandarin, Russian, Spanish, Urdu		Arabic Egyptian, Arabic (unknown dialect), Cantonese, French, Korean, Vietnamese
only telephone data in the train set	G_T: Bengali, Czech, Indian English, Lao, Panjabi, Polish, Slovak, Tamil, Thai	Arabic Iraqi, Arabic Levantine, Arabic Maghrebi	German, Indonesian, Italian, Japanese, Min Nan Chinese, Tagalog, Wu Chinese
only broadcast narrow-band data in the train set	G_B: Dari, Pashto, Turkish, Ukrainian	Arabic MSA	Amharic, Azerbaijani, Belarusian, Bosnian, Bulgarian, Croatian, Georgian, Haitian, Hausa, Portuguese, Romanian, Swahili, Tibetan, Uzbek

where m is a margin parameter experimentally selected on a validation set. We used the value $m = 1$ in our experiments.

We also evaluate a robust version of metric learning with multiple negative examples: n-pair loss [29, 30]. This loss function uses $N-1$ negative samples $\{x_i^-\}_{i=1}^{N-1}$ belonging to each of the other classes. It is given by:

$$\mathcal{L}_{\text{n-pair}} = \log(1 + \sum_{i=1}^{N-1} \mathrm{e}^{f(x)^T f(x_i^-) - f(x)^T f(x^+)}) \qquad (5)$$

where $f(a)^T f(b)$ is the canonical scalar product between two embeddings $f(a)$ and $f(b)$. We used the train corpus described in Table 1 with $N = 51$ languages.

3.2. Baseline loss functions

Metric learning loss functions are compared to two classification loss functions: the traditional cross-entropy loss (CE) and the newly introduced additive angular margin softmax loss (AAM) [31, 12].

Moreover, the mismatch reduction capability of metric learning is compared to a domain adaptation loss function which uses the domain information: cross-entropy regularized with Maximum Mean Discrepancy between telephone and broadcast channels. The efficiency of this loss function to reduce the domain mismatch has been demonstrated in [15]. Here we use the kernel $k(x, y) = -\|x - y\|_2$, to be consistent with the mismatch measure.

4. Results

Several language recognition systems have been trained on the NIST LRE2011 corpus with the recipe described in Section 2. The systems only differ by the loss function used to train the embedding extractor.

4.1. Language recognition performance

Table 2 presents performance of the trained systems over the test set of the NIST LRE 2011 corpus, for the three segment durations. Embedding extractors have been trained with segments of two to four seconds and a specific backend classifier has been trained for each segment duration. We train two systems with a metric learning loss only: triplet or n-pair loss. Then two systems are trained with classification losses: cross-entropy and AAM-softmax. Finally each classification loss is combined with the metric learning n-pair loss (by summation of the losses). We compare the trained systems with a MMD based domain adaptation regularization of the cross-entropy loss [15].

First, the baseline system trained with cross-entropy (line 'CE' in Table 2) achieves a competitive performance in comparison with the best systems submitted to the evaluation NIST LRE 2011 [32], which get values of actAPD of the order 8%, 12% and 22% for segments of respectively 30 s, 10 s and 3 s. Moreover the comparison with top-leading individual systems [25] confirms the superiority of *x-vector* systems trained with data augmentation over *i-vector* approaches. The recently introduced AAM-softmax loss improves over this baseline. Domain adaptation with MMD loss is useful but uses more information during training.

Using only a metric learning loss function (triplet or n-pair) allows to train an *x-vector* DNN, but with poorer performance than a classification loss. The use of several negative samples for the n-pair loss improves over the sampling of only one negative example for triplet loss. The best performance is achieved by combination of classification losses with n-pair loss. We hypothesize that n-pair loss reduces the impact of distortion factors in the embedding space. In the next two subsections, we measure this effect.

4.2. Channel mismatch

We evaluate the quality of embeddings by measuring and comparing language discriminability with channel mismatch. We use the MMD divergence between groups of x-vectors computed on 10 second speech segments. We average the values by groups of languages (see Table 1): languages present on both channels (G_A), languages only present on the telephone channel (G_T), languages only present on the broadcast channel (G_B) in the train set.

Figure 1 displays mismatches for three types of systems: two systems trained with a classification loss (cross-entropy, AAM-softmax), a system trained with a domain adaptation strategy (cross-entropy with MMD) and two systems trained with metric learning (cross-entropy and n-pair, AAM-softmax and n-pair). The scale of the MMD varies across different systems and only the relative values of language discriminability and channel mismatch is relevant. To allow a fair comparison, we normalize all divergences of each system by dividing them by the average language discriminability on the telephone channel $\mathcal{D}_{\text{Language}}(L, T)$ for languages of the group G_A.

For a baseline system trained with cross-entropy, channel mismatch has the same magnitude as language discriminability for languages observed on both channels and is superior for languages observed on only one channel. The same observation can be made for the AAM-softmax loss. When using the

Table 2: *Performance of the language recognition systems. The x-vectors DNN was trained with different loss functions.*

	Performance (%) for different test segment durations														
	30 seconds					10 seconds					3 seconds				
	APD		DCF		EER	APD		DCF		EER	APD		DCF		EER
Loss function	min	act	min	act		min	act	min	act		min	act	min	act	
triplet	12.5	15.9	5.7	9.3	6.1	19.9	23.5	9.6	12.2	10.2	29.7	32.4	19.1	20.5	19.6
n-pair	9.8	12.8	3.8	6.9	4.2	15.6	20.6	6.6	9.1	7.1	22.6	27.9	13.6	15.0	14.0
CE	6.5	**8.7**	2.8	5.7	3.1	11.7	15.3	5.2	7.1	5.5	**20.2**	**22.9**	**11.3**	**12.4**	**11.6**
CE and MMD	6.4	**8.7**	2.7	**4.2**	**3.0**	11.1	**14.2**	4.8	**6.2**	5.2	21.2	23.9	12.6	13.7	13.0
CE and n-pair	**6.2**	9.2	**2.6**	5.3	**3.0**	**10.8**	15.3	**4.7**	**6.2**	**5.1**	21.0	24.7	12.2	13.8	12.7
AAM	4.8	**6.6**	**2.1**	4.8	**2.4**	9.3	11.5	4.2	6.1	4.6	19.1	21.6	11.7	**13.0**	12.1
AAM and n-Pair	**4.5**	6.9	**2.1**	**4.6**	**2.4**	**8.8**	**11.4**	**4.1**	**5.8**	**4.5**	**18.2**	**21.1**	**11.6**	**13.0**	**12.0**

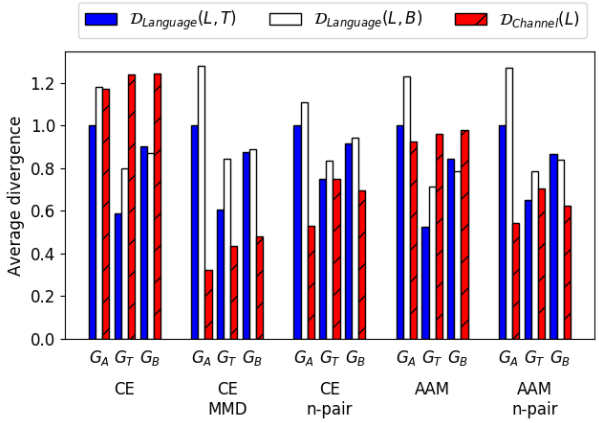

Figure 1: *Language discriminability and channel mismatch, averaged by groups of languages, see Table 1. Divergences are measured for segments of 10 seconds and normalized for each system with the average value of $\mathcal{D}_{Language}(L,T)$ on G_A.*

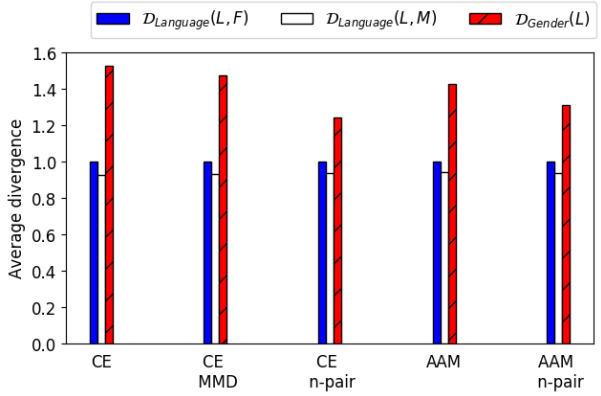

Figure 2: *Language discriminability and gender mismatch, averaged over all languages. Divergences are measured for segments of 10 seconds and normalized for each system with the average value of $\mathcal{D}_{Language}(L,F)$.*

domain information during training, the MMD loss function reduces the channel mismatch in comparison with language discriminability, especially for languages observed on both channels (G_A).

Metric learning achieves a similar effect without using domain information during training. For an *x-vector* DNN trained with the combination of cross-entropy and n-pair loss, channel mismatch is inferior to language discriminability for languages observed on both channels. Both divergences have the same order of magnitude for languages observed on only one channel.

4.3. Gender mismatch

Figure 2 displays language discriminability and gender mismatch, averaged over the 24 languages of the test set. As for channel mismatch, gender mismatch is more important than language discriminability for the systems trained with classification loss functions. N-pair loss reduces this mismatch without using gender information. As expected, the system trained with MMD loss does not reduce the gender mismatch because it has been designed only to reduce the channel mismatch. Metric learning reduces a mismatch in the embedding space, even if this mismatch is unknown to the designer of the system. This makes metric learning a fundamental tool to improve robustness of language identification embeddings.

5. Conclusion

We investigated metric learning objective for training of the *x-vector* DNN in a language recognition system. We showed that the combination of the n-pair loss with a classification loss reduces the mismatch between *x-vectors* from telephone and broadcast channels, even for languages which have only been observed on one of the two channels in the train set. Consequently the overall language recognition performance is improved compared to training only with a classification loss. Metric learning achieves similar performance as regularization of the embedding extractor with a domain adaptation loss but without using domain labels during training. It has the ability to reduce an *a priori* unknown mismatch, as evidenced for gender.

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] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in i-vectors space," in *Proc. Interspeech - Annual Conference of the International Speech Communication Association*, 2011.

[2] A. Lozano-Diez, O. Plchot, P. Matějka, and J. Gonzalez-Rodriguez, "DNN based embeddings for language recognition," in *Proc. ICASSP - International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 5184–5188.

[3] M. Mclaren, M. K. Nandwana, D. Castán, and L. Ferrer, "Approaches to multi-domain language recognition." in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 90–97.

[4] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors." in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 105–111.

[5] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," *arXiv preprint arXiv:1909.06351*, 2019.

[6] S. O. Sadjadi, T. Kheyrkhah, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, "Performance analysis of the 2017 NIST language recognition evaluation." in *Proc. Interspeech - Annual Conference of the International Speech Communication Association*, 2018, pp. 1798–1802.

[7] O. Plchot, P. Matějka, O. Novotný, S. Cumani, A. Lozano-Diez, J. Slavicek, M. Diez, F. Grézl, O. Glembek, M. Kamsali *et al.*, "Analysis of BUT-PT submission for NIST LRE 2017." in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 47–53.

[8] F. Richardson, P. A. Torres-Carrasquillo, J. Borgstrom, D. E. Sturim, Y. Gwon, J. Villalba, J. Trmal, N. Chen, R. Dehak, and N. Dehak, "The MIT Lincoln Laboratory/JHU/EPITA-LSE LRE17 system." in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 54–59.

[9] P.-M. Bousquet and M. Rouvier, "On robustness of unsupervised domain adaptation for speaker recognition," *Proc. Interspeech - Annual Conference of the International Speech Communication Association*, pp. 2958–2962, 2019.

[10] F. Verdet, D. Matrouf, J.-F. Bonastre, and J. Hennebert, "Coping with two different transmission channels in language recognition." in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2010, p. 39.

[11] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2018, pp. 74–81.

[12] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak *et al.*, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations," *Computer Speech & Language*, vol. 60, p. 101026, 2020.

[13] J. Rohdin, T. Stafylakis, A. Silnova, H. Zeinali, L. Burget, and O. Plchot, "Speaker verification using end-to-end adversarial language adaptation," in *Proc. ICASSP - International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 6006–6010.

[14] G. Bhattacharya, J. Alam, and P. Kenny, "Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training," in *Proc. ICASSP - International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 6041–6045.

[15] R. Duroselle, D. Jouvet, and I. Illina, "Unsupervised regularization of the embedding extractor for robust language identification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2020.

[16] J. Son Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," *arXiv*, pp. arXiv–2003, 2020.

[17] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances." in *Proc. Interspeech - Annual Conference of the International Speech Communication Association*, 2017, pp. 1487–1491.

[18] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Proc. Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[19] V. Mingote, D. Castan, M. McLaren, M. K. Nandwana, E. L. Ortega, and A. Miguel, "Language recognition using triplet neural networks," *Proc. Interspeech - Annual Conference of the International Speech Communication Association*, pp. 4025–4029, 2019.

[20] G. Gelly and J.-L. Gauvain, "Spoken language identification using LSTM-based angular proximity." in *Proc. Interspeech - Annual Conference of the International Speech Communication Association*, 2017, pp. 2566–2570.

[21] C. S. Greenberg, A. F. Martin, and M. A. Przybocki, "The 2011 NIST language recognition evaluation," in *Proc. Interspeech - Annual Conference of the International Speech Communication Association*, 2012.

[22] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2012, pp. 291–297.

[23] R. Fér, P. Matějka, F. Grézl, O. Plchot, K. Veselý, and J. H. Černocký, "Multilingually trained bottleneck features in spoken language recognition," *Computer Speech & Language*, vol. 46, pp. 252–267, 2017.

[24] N. Brummer and D. A. Van Leeuwen, "On calibration of language recognition scores," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[25] E. Singer, P. Torres-Carrasquillo, D. A. Reynolds, A. McCree, F. Richardson, N. Dehak, and D. Sturim, "The MITLL NIST LRE 2011 language recognition system," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2012.

[26] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[27] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouve, and G. Peyré, "Interpolating between optimal transport and MMD using Sinkhorn divergences," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2019, pp. 2681–2690.

[28] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.

[29] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in neural information processing systems*, 2016, pp. 1857–1865.

[30] A. Kulkarni, V. Colotte, and D. Jouvet, "Transfer learning of the expressivity using flow metric learning in multispeaker text-to-speech synthesis," *to appear in Proc. Interspeech - Annual Conference of the International Speech Communication Association*, 2020.

[31] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR - Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[32] "LRE 2011 results, NIST," https://www.nist.gov/itl/iad/mig/lre11-results, 2013.