INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Triplet Ranking-based Neural Network for Speaker Diarization and Linking

Gaël Le Lan[1,2], *Delphine Charlet*[1], *Anthony Larcher*[2], *Sylvain Meignier*[2]

[1]Orange Labs, France
[2]LIUM, University of Le Mans, France

first.lastname@orange.com, first.lastname@lium.univ-lemans.fr

Abstract

This paper investigates a novel neural scoring method, based on conventional *i-vectors*, to perform speaker diarization and linking of large collections of recordings. Using triplet loss for training, the network projects *i-vectors* in a space that better separates speakers in terms of cosine similarity. Experiments are run on two French TV collections built from REPERE [1] and ETAPE [2] campaigns corpora, the system being trained on French Radio data. Results indicate that the proposed approach outperforms conventional cosine and Probabilistic Linear Discriminant Analysis scoring methods on both within- and cross-recording diarization tasks, with a Diarization Error Rate reduction of 14% in average.

Index Terms: speaker diarization, neural network, triplet loss

1. Introduction

The increasing volume of audio and video data daily produced by social or traditional media, conferences, meetings or MOOCs requires powerful tools to automatically index topics, languages or speakers. In that context, the task of speaker diarization and linking aims at uniquely label speakers across a collection of recordings, without a priori knowledge about the speakers.

In the literature, variable terminologies are used to describe the task (*Speaker Linking* in [3][4][5][6], *Cross-Show Speaker Diarization* for [7][8][9]), but recently, the term *Speaker Diarization and Linking* is preferred [10][11]. Each recording is usually processed separately (within-recording diarization) before estimated speaker segments are linked across the collection (cross-recording speaker linking). In this paper, we use the terms *diarization* for within-recording diarization, and *linking* for cross-recording linking.

Speaker Diarization and Linking is about differentiating speakers. State-of-the-art approaches combine the *i-vector* paradigm [12] to represent speech segments, and within- and between-speaker variability compensation to discriminate them in terms of speaker. Within- and between-speaker variabilities are estimated over a speaker labeled dataset, which must include multiple examples of a same speaker in various acoustic conditions. *I-vectors* can be compared using similarity scores (cosine, with or without speaker variability compensation like Within Class Covariance Normalization (WCCN [12])) or likelihood ratios (through Probabilistic Linear Discriminant Analysis (PLDA [13])).

In this paper, we propose a novel scoring method for diarization and linking, by replacing cosine or PLDA with a neural-based approach. Some neural-based *i-vector* scoring methods for speaker verification were introduced by [14] and proved to be competitive with PLDA. Our proposal is inspired by [15] and [16], who proposed face/speaker neural embeddings optimized for face/speaker recognition and clustering, using the triplet loss [17] for training. The main difference of our proposal is that we decide to directly work with *i-vectors* as an input, instead of raw image/acoustic features. This explains why we see the proposed method as an alternative to conventional scoring method for *i-vectors*, such as in [14].

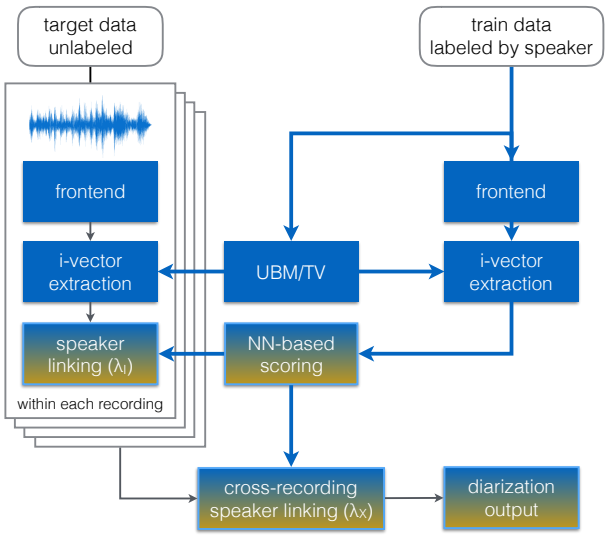

Figure 1: *Overview of the diarization framework.*

Subsequent sections are organized as follows: first, we describe the diarization framework and present the proposed neural network used for scoring. Then we describe the data used for the experiments and conclude with a discussion about the performances of the proposed system and the possible improvements.

2. Diarization Framework

Figure 1 presents the diarization framework, developed using the SIDEKIT toolkit [18]. First, each recording of the collection is independently processed. The frontend homogeneously segments the audio in terms of speakers, using Gaussian Divergence [19] and Bayesian Information Criterion (BIC) [20], so that an accurate *i-vector* can be extracted over each segment. All extracted *i-vectors* are then clustered, using a neural based similarity measure and a clustering method, with a clustering threshold λ_I. At the end of the diarization step, each within-recording cluster is represented by the average of its *i-vectors*.

Once all recordings have been processed, cross-recording linking between those averaged *i-vectors* is applied at the scale of the collection, using the same type of scoring and clustering, with a clustering threshold λ_X.

The 200-dimension *i-vector* representation used in the following is estimated over a GMM/UBM of 256 Gaussians with

http://dx.doi.org/10.21437/Interspeech.2017-270

diagonal covariance, computed on the $train$ corpus. The proposed similarity scoring is compared with cosine or PLDA of rank 100. Those dimensions were chosen after an exhaustive search. To cluster the *i-vectors*, two methods are compared: Complete-linkage Hierarchical Agglomerative Clustering (HAC) or Connected Components (CC) Clustering.

3. Triplet Ranking Framework

Contrary to the speaker diarization and linking framework of [21] where PLDA is used to compute similarities between *i-vectors*, we decide to replace PLDA by a Neural Network approach. The method is inspired by that of [15] and [16], where the triplet loss [17] is used to train a neural network embedding, which aims at separating faces or speakers. Instead of training a network on raw features (e.g. MFCCs), we propose to use the *i-vector* representation.

We propose to non-linearly project the *i-vectors* on a unit sphere that better separates speaker classes in terms of cosine similarity, using a simple feed-forward network f trained for that purpose. The similarity between two *i-vectors* (ϕ_1, ϕ_2) corresponds to the cosine similarity between the two embeddings $(f(\phi_1), f(\phi_2))$. To achieve better separability in the projection space, we adopt the triplet loss paradigm. We will call this variability compensation method for cosine scoring Triplet Ranking (TR) scoring.

From a training set of *i-vectors* representing different speakers, triplets (ϕ_a, ϕ_p, ϕ_n) are sampled so that ϕ_a (called *anchor*) and ϕ_p (called *positive*) represent the same speaker and ϕ_n (called *negative*) a different speaker. Triplet loss aims at better separating the speaker classes in the embedding space by maximizing the *anchor-positive* similarity, while minimizing the *anchor-negative* similarity. For the set of all possible N triplets $\mathcal{T} = (\phi_a^i, \phi_p^i, \phi_n^i)_{i\in[1..N]}$, the loss is defined as

$$L(\mathcal{T}) = \Sigma_i^N max(0, \Delta_i + \alpha) \quad (1)$$

$$\Delta_i = -\frac{f(\phi_a^i)f(\phi_p^i)^T}{\|f(\phi_a^i)\|\|f(\phi_p^i)\|} + \frac{f(\phi_a^i)f(\phi_n^i)^T}{\|f(\phi_a^i)\|\|f(\phi_n^i)\|} \quad (2)$$

α is a margin aiming at forcing a better separation between the classes. Ideally we want that for any triplet i, $\Delta_i + \alpha < 0$. To optimize training, it is faster to only select triplets that contribute to the loss. In the original paper [15], two different selection strategies are discussed: *hard-selection* corresponds to all triplets contributing to the loss (ie. $0 < \Delta_i + \alpha$), while *soft-selection* consists in excluding the hardest triplets (ie. keeping those in the margin, $0 < \Delta_i + \alpha < \alpha$).

The procedure is the following for each training epoch. For each speaker class which contains at least 3 *i-vectors*, we randomly select *i-vectors* pairs (ϕ_a^i, ϕ_p^i). Among the k-Nearest Neighbors (k-NN) of each *anchor* embedding $f(\phi_a^i)$, a *negative* ϕ_n^i is randomly picked so that $0 < \Delta_i + \alpha$ ($< \alpha$, depending on the selection strategy). All selected triplets are then used to update the network weights and gradients.

4. Experimental collections

The datasets used for our experiments are described in [22]. They consist in multiple recordings of various French radio and TV shows broadcast between 1998 and 2007, where speakers are identified by their first and last name. Speakers appearing in more than one recording of a dataset are called recurring (R.) speakers, as opposed to one-time (O.T.) speakers, who only speak in one recording. Only radio recordings were used to build the $train$ corpus, while both $target$ corpora contain TV recordings only. This was initially chosen to maximize the acoustic mismatch between the $train$ and $target$ data. Since we do not have any *development* data, a cross evaluation is performed between the two *target* collections.

Table 1: *Composition of target corpora.*

Corpus	LCP	BFM
Episodes	45	42
Labeled speech duration	10h08m	19h57m
One-Time speakers	127	345
Recurring speakers (2+ occurrences)	93	77
R. speakers (3+ occurrences)	48	35
Total speakers	220	422
O.T. speakers speech proportion	20.12%	44,84%
R. speakers (2+ occurrences) s.p.	79.88%	55,16%
R. speakers (3+ occurrences) s.p.	67.06%	45.94%
Average speaker time per episode	1m08s	1m58s

4.1. Train corpus

The $train$ corpus, used to train initial and contrastive systems, is composed of 317 audio files from ESTER [23] campaign corpora, taken from radio broadcasts, for a total of 190 hours of speech duration. The corpus contains 372 speakers that speak in at least three recordings, with a minimum speech time per recording of 10s.

4.2. Target corpora

We define two $target$ corpora built from REPERE [1] and ETAPE corpora [2]: LCP and BFM. LCP (resp. BFM), is the collection of all available recordings of the show *LCP Info* (resp. *BFM Story*), a French TV news broadcast (resp. talk-show). Those two shows have been selected because they both contain a decent number of episodes (more than 40), and there is a large amount of recurring speakers, who speak for more than 50% of the total speech duration of the collection. Numerical details about the two corpora are presented in table 1.

5. Experiments

Diarization and linking systems are evaluated with the Diarization Error Rate (DER). DER was introduced by the NIST as the fraction of speech time which is not attributed to the correct speaker, using the best match between references and hypothesis speaker labels. The scoring tool [24] evaluates within-recording and cross-recording speaker diarization. Cross-recording speaker diarization aims at labeling a recurring speaker in the same way, in every recording that composes a collection. For DER computation, a collar of 250ms is allowed, and the overlapping speech is not evaluated. In the following, we will call the cross-recording DER *X-DER*, as opposed to *I-DER*, for within-recording DER.

5.1. Neural Network Training

5.1.1. Implementation

Before evaluating the proposed TR scoring on the diarization and linking task, we first evaluated it on a speaker verification task, using a ground truth version *i-vectors* extracted using true labels of the $target$ datasets. The metrics used are the Equal Error Rate (EER) and minimum Detection Cost Function (minDCF), with a prior of 1%. This allows us to faster

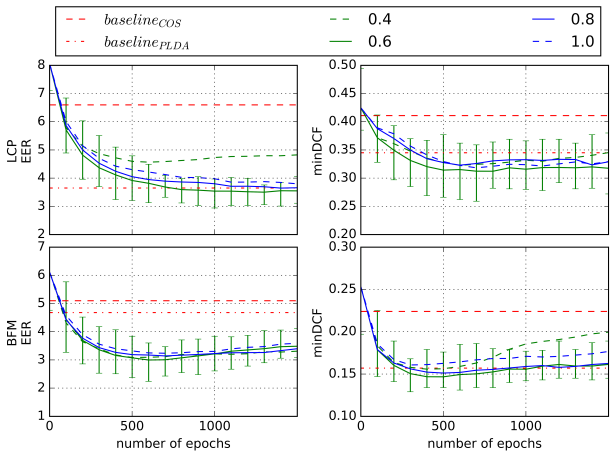

Figure 2: *Average EER and minDCF on both* $target$ *datasets, using different margins for training. Each condition is repeated 20 times.*

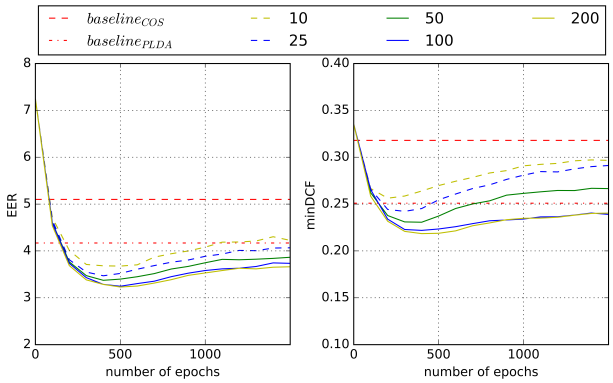

Figure 3: *Average EER and minDCF on the two* $target$ *datasets, using different number of nearest neighbors for training. Each condition is repeated 20 times and averaged.*

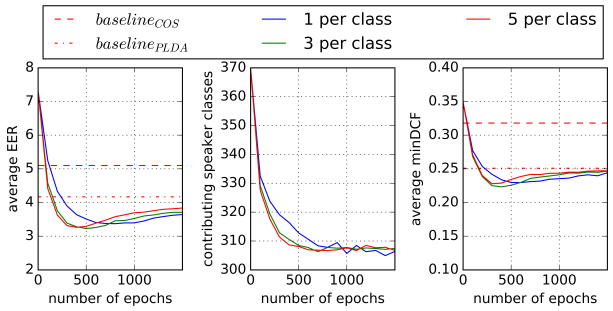

Figure 4: *Influence of the number of examples provided per class. Average EER, minDCF and number of speaker classes contributing to the loss, on the two* $target$ *datasets.*

explore different configurations. The neural network consists in a feed-forward layer of dimension 200 (same as the *i-vectors* dimension), followed by a *tanh* activation layer. The chosen optimizer is *Adadelta* [25] and the network is implemented with Keras [26]. In the next paragraphs, we explore some key aspects of the network configuration : choice of margin, number of k-NN and representativity of the classes. In all experiments, we decide to use the *soft* triplet selection strategy, as our preliminary explorations did not show significant differences with *hard selection*.

5.1.2. Choice of margin

Figure 2 presents the results on the ground truth *i-vectors* of both *target* datasets, as a function of the number of epochs. Different margins are tested, from 0.4 to 1.0 with a 0.1 step, using the same network initialization. For each margin, the experiment is repeated 20 times (ie. 20 different initializations). At each epoch, one triplet per speaker class is presented, the k-NN value of 100. For better readability, only one half of the curves is presented (0.2 step).

Results show that the proposed approach outperforms the cosine setup for all tested margins for both metrics. The optimal margin is of 0.6 and is presented with the (min, max) performances interval at each epoch, over all 20 experiments. When comparing with the PLDA setup, we see that the proposed scoring with a margin of 0.6 outperforms PLDA for both metrics, in average, while being very close in terms of EER for LCP and in terms of minDCF for BFM.

5.1.3. Number of nearest neighbors

In practice, using nearest neighbors for triplet selection is a way to accelerate training and to optimize triplet selection. A *negative i-vector* compliant to the margin constraint is more likely to be found in a certain neighborhood of an *anchor*. In figure 3, we explore the number of nearest neighbors and influence on the EER and minDCF. The k-NNs are updated every 50 epochs. The same evaluation protocol is used : contrastive evaluation of parameters using the same network initialization, repeated 20 times and averaged.

Results show that the higher the k-NN, the better the performances, but past 100, results do not improve that much, as the majority of good *negative* candidates must be located in the 100-NN area. We decide to keep the 100-NN configuration for the following experiments, as it is a good trade-off between performance and speed.

5.1.4. Representativity of the classes

In the literature [15, 16], triplet-loss neural networks are trained on raw input (acoustic features or images), and 40 examples per class are provided at each epoch. In our work, we train the network on *i-vectors*, and the number of examples per class is very limited (our $train$ corpus contains between 3 and 59 *i-vectors* per class, half of the classes containing 5 examples or less). Since we cannot provide much diversity in terms of speaker pairs (3 examples in a class means 6 possible *anchor-positive* pairs only), we explore the influence of the number of examples per class provided during training. Providing too many examples per class might lead to introduce some bias, as the same *anchor-positive* pairs would be selected at each epoch, even if diversity is enforced by the presence of *negative*. Another side effect of the low number of examples per class is that after some epochs, some classes stop contributing to the loss: there is no *negative* in the margin anymore. Results are presented in figure 4, where we compare performances using 1,3 and 5 examples per class, each experiment being repeated 20 times.

Results show that using more than one triplet per speaker class works better, as it provides some within speaker variability information. Using 3 triplets per class gives the best EER and minDCF and we will keep that configuration in the following

Table 2: *oracle diarization results, for both collections*

Corpus	LCP	BFM
oracle linking I-DER & X-DER	5.9	7.8
false speech	3.0	3.7
missed speech	0.7	1.0
speaker error	2.2	3.0

Table 3: *Baseline performances of contrastive and proposed systems, for HAC and CC clustering, in terms of I- and X-DER. λ_I & λ_X are the clustering thresholds, applied on the opposite of cosine similarity or PLDA likelihood ratio.*

scoring	clust.	λ_I	λ_X	LCP DER		BFM DER	
				I-.	X-.	I-.	X-.
cosine	HAC	-0.55	-0.5	8.5	19.5	13.6	23.8
	CC	-0.55	-0.65	8.5	22.6	12.4	19.3
$PLDA$	HAC	10	10	10.0	19.1	10.6	15.7
	CC	10	-20	8.7	21.2	9.9	13.6
TR_{avg}	CC	-0.55	-0.65	8.0	16.6	9.8	13.3
TR_{best}				**7.9**	**16.1**	**9.6**	**13.1**

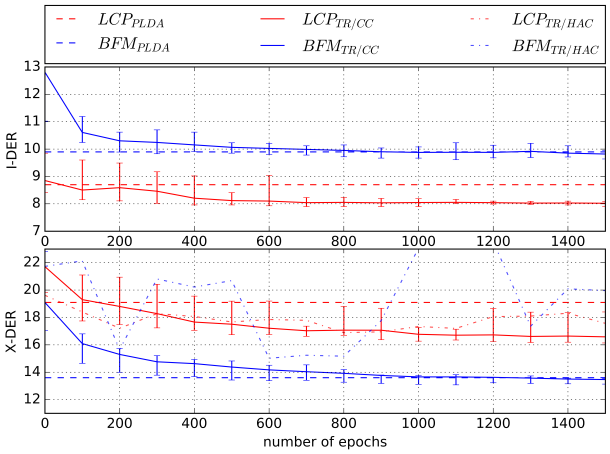

Figure 5: *Average X-DER on the two target datasets, using a margin of 0.6, soft selection strategy and 3 examples per class. Clustering configuration is ($\lambda_I = -0.55, \lambda_X = -0.65$).*

experiments. For all tested configurations, we see a reduction of around 17% of the contributing speaker classes during training.

5.2. Speaker Diarization and Linking

5.2.1. Oracle

Before running any diarization experiment, in table 2, we present the performances of an ideal speaker linking system. The speaker segmentation module (see section 2) aims at producing pure speaker segments to allow accurate *i-vector* extraction. An ideal linking would consist in labeling each segment according to its most speaking speaker. As seen in the table, around 60% of the DER is due to speech detection, while the last part consists in speaker error. This speaker error is due to the speaker segmentation module. Thus some *i-vectors* are extracted over segments containing more than one speaker.

5.2.2. Diarization Results

The Neural Network configuration being set, we now investigate the performances on the diarization and linking task, using I- and X-DER as an evaluation metric. We stop working with ground truth *i-vectors* and use the ones produced by speaker segmentation. Table 3 shows the *baseline* performances of the contrastive cosine and PLDA setups, for HAC and CC clustering. The clustering thresholds are common to both corpora.

In Figure 5, we present the averaged I- and X-DER over 20 speaker linking experiments, using CC clustering and 20 different TR scoring networks, for the two *target* corpora, as a function of the number of epochs. (min, max) intervals are also presented. Results indicate that for both corpora, after 1300 epochs, the proposed system (TR/CC) outperforms the PLDA setup ($PLDA$), in average, with an X-DER of 16.6% (resp. 13.3%) on LCP (resp. BFM), whereas PLDA achieved 19.1% (resp. 15.7%).

In [21], the authors used HAC for speaker linking, but our preliminary experiments with TR scoring using HAC gave inconclusive results. As shown in the lower part of Figure 5, HAC gives erratic variations in terms of DER depending on the number of epochs, for BFM. However, we noticed that using CC for clustering proved to be more adequate, with a smooth DER reduction throughout epochs. Indeed, the network tries to ensure the classes are separated by a fixed margin, while CC makes two elements belong to a same cluster if they are separated by less that a certain threshold. By design, both methods work well together.

It is worth noticing that optimal DERs are obtained for more than 1300 epochs, while in the previous speaker verification experiments, performances tended to slightly degrade in the same area, especially for BFM. This could be explained by the fact that the *i-vectors* used for speaker verification are different to those used for diarization, as the latter may be corrupted by speaker segmentation errors. For BFM, we noticed in section 5.1.2 that performances of TR scoring were close to PLDA in terms of minDCF, as opposed to LCP. The same conclusion applies for DER: minDCF simulates a situation where for a given *i-vector*, there are more possible impostors that correct candidates. This is also the case for speaker linking.

Finally, in the last rows of table 3, the averaged and best TR-based diarization results after 1500 epochs are presented, over the 20 linking experiments. We observe that the presented optimal clustering configuration is the same as the cosine-baseline one ($\lambda_I = -0.55, \lambda_X = -0.65$).

6. Conclusions

In this paper, we proposed an novel scoring method for *i-vectors*. A neural network projects *i-vectors* in a space that optimizes speaker separation in terms of cosine similarity. The triplet loss is used for training, with a margin close to the optimal speaker separation threshold in the initial *i-vector* space. Due to the small amount of *train* data, 3 triplets per speaker class are presented at each epoch and some classes stop contributing to the training after few iterations.

Cross-recording speaker diarization results on two distinct corpora show that the proposed method is competitive with state-of-the-art scoring methods for *i-vectors*, both in terms of within- and cross-recording DER. The triplet loss approach seems to be promising for future speaker diarization and linking architectures. Further work could consist in replacing the *i-vector* by a neural-based speaker embedding, or proposing a domain adaptation method for the neural scoring, as it was explored for PLDA.

7. References

[1] O. Galibert and J. Kahn, "The first official repere evaluation," in *Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, 2013.

[2] O. Galibert, J. Leixa, A. Gilles, K. Choukri, and G. Gravier, "The ETAPE Speech Processing Evaluation," in *Conference on Language Resources and Evaluation*, Reykyavik, Iceland, May 2014.

[3] H. Bourlard, M. Ferras, N. Pappas, A. Popescu-Belis, S. Renals, F. McInnes, P. Bell, and M. Guillemot, "Processing and Linking Audio Events in Large Multimedia Archives: The EU inEvent Project," in *Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France, August 2013.

[4] M. Ferràs and H. Bourlard, "Speaker Diarization and Linking of Large Corpora," in *Proceedings of IEEE Workshop on Spoken Language Technology*, Miami, Florida (USA), December 2012.

[5] H. Ghaemmaghami, D. Dean, and S. Sridha, "Speaker Attribution of Australian Broadcast News Data," in *Proceedings of Interspeech satellite workshop on Speech, Language and Audio in Multimedia (SLAM)*, Marseille, France, August 2013.

[6] D. A. Van Leeuwen, "Speaker linking in large data sets," Proc. Odyssey 2010.

[7] V.-A. Tran, V. B. Le, C. Barras, and L. Lamel, "Comparing Multi-Stage Approaches for Cross-Show Speaker Diarization," in *Proceedings of Interspeech*, Florence, Italy, August 2011.

[8] Q. Yang, Q. Jin, and T. Schultz, "Investigation of Cross-show Speaker Diarization," in *Proceedings of Interspeech*, Florence, Italy, August 2011.

[9] G. Dupuy, M. Rouvier, S. Meignier, and Y. Estève, "I-vectors and ILP Clustering Adapted to Cross-Show Speaker Diarization," in *Proceedings of Interspeech*, Portland, Oregon, USA, September 2012.

[10] M. Ferras, S. Madikeri, and H. Bourlard, "Speaker diarization and linking of meeting data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, no. 99, pp. 1–1, 2016.

[11] H. Ghaemmaghami, D. Dean, and S. Sridharan, "A cluster-voting approach for speaker diarization and linking of australian broadcast news recordings," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4829–4833.

[12] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[13] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Speaker Odyssey Workshop*, 2010.

[14] G. Bhattacharya, M. J. Alam, P. Kenny, and V. Gupta, "Modelling speaker and channel variability using deep neural networks for robust speaker verification," in *2016 IEEE Spoken Language Technology Workshop, SLT 2016, San Diego, CA, USA, December 13-16, 2016*, 2016, pp. 192–198.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[16] H. Bredin, "Tristounet: Triplet loss for speaker turn embedding," *arXiv preprint arXiv:1609.04301*, 2016.

[17] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1386–1393.

[18] A. Larcher, K. Aik Lee, and S. Meignier, "An extensible speaker identification sidekit in python," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[19] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Multi-stage speaker diarization of broadcast news," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 5, pp. 1505–1512, Feb. 2006.

[20] G. Schwarz *et al.*, "Estimating the Dimension of a Model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[21] G. L. Lan, D. Charlet, A. Larcher, and S. Meignier, "Iterative plda adaptation for speaker diarization," in *INTERSPEECH*, 2016.

[22] G. L. Lan, S. Meignier, D. Charlet, and P. Deléglise, "Speaker diarization with unsupervised training framework," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5560–5564.

[23] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Proceedings of Interspeech*, Brighton, Royaume Uni, Sept 2009.

[24] O. Galibert, "Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech." in *INTERSPEECH*, 2013, pp. 1131–1134.

[25] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.

[26] F. Chollet, "Keras," https://github.com/fchollet/keras, 2017.