Adaptation strategy and clustering from scratch for new domains in speaker recognition

Pierre-Michel Bousquet, Mickael Rouvier

Adaptation strategy and clustering from scratch for new domains in speaker recognition

Pierre-Michel Bousquet, Mickael Rouvier

University of Avignon - LIA, France

{pierre-michel.bousquet, mickael.rouvier}@univ-avignon.fr

Abstract

This paper investigates the domain adaptation back-end methods introduced over the past years for speaker recognition, when the mismatch between training and test data induces a severe degradation of performance. The conducted analyses lead to suggest some ways, experimentally validated, for the task of collecting in-domain data and making the most of the first data at hand. The proposed strategy helps to quickly increase accuracy of the detection, without omitting to take into account the practical difficulties of the task of data collecting in real-life situations, and without the cost and delay for forming the expected large and speaker-labeled in-domain dataset. Moreover, a new approach of artificial speaker labeling by clustering is proposed, that dispenses of collecting a preliminary annotated in-domain dataset, with a similar gain of efficiency.

Index Terms: Speaker recognition, domain adaptation, clustering

1. Introduction

Our era, marked by inflation of data and computing power, has favored the introduction of neural networks in machine learning techniques and, therefore, in speaker recognition. Unfortunately, the scope of the huge databases used for training the network does not necessarily span the wide variety of settings found in real life (for speaker utterances: channel, device, duration, language, type and level of noise, reverberation, etc.). Much progress has been made in recent years to adapt existing models to new domains (without waiting for large amounts of matching data), boosted by recent evaluation campaigns as NIST SRE 2016-18-19 [1, 2] which focused on severe shifts due to language mismatch. These advances are due to the DNN-based representation of utterances (x-vector embedding), data augmentation and also to refinement of the back-end process.

The time has come to thoroughly analyze the scenario of a new domain, not included or weakly in the training databases, on which detecting the speaker is desired. The goal being to fast provide a system that achieves recognition performance comparable to one that is provided all knowledge of the domain mismatch. This paper investigates this task of adaptation from scratch, in terms of data collection strategy (what is the quantity -but also quality- of data to gather for making the models more accurate ?) and of back-end methods (what is the impact of the various adaptation techniques, supervised or unsupervised, given the available resources ?).

To go further, the use of clustering for domain adaptation is reviewed: this technique allows supervised approaches despite the lack of annotated data, by identifying the resulting classes of a clustering to speaker labels. This led us to propose a new approach, which dispenses of pre-labeled data without disquieting loss of accuracy.

The outline of the paper is as follows: section 2 describes the domain adaptation methods, section 3 presents our contributions, section 4 reports the results of our strategy analysis and of the proposed clustering, then we conclude in section 5.

2. Domain adaptation

2.1. Methods

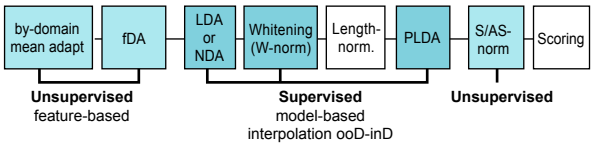

Figure 1: *Steps of the embedding-based speaker recognition back-end process, with the types of domain adaptation that can be applied.*

Figure 1 shows the most usual stages of a back-end process for speaker recognition systems based on embeddings (x-vector, i-vector...) and the different adaptation techniques that can be included. The mean and total covariance of the out-of-domain and in-domain training datasets are used for unsupervised adaptation, when no speaker label is available for development. The methods are feature-based (i.e. based on the embeddings), transforming the vectors to reduce the shift between target and out-of-domain distributions. Mean-adaptation-by-domain subtracts to each embedding the mean parameter of its domain, estimated on a dataset. Correlation alignment (CORAL [3, 4]) attempts to move the out-of-domain distribution towards the target ones, leading to transform out-of-domain data into *pseudo in-domain* data. Introduced in [5], feature-Distribution Adaptor (fDA) is a variant of this method that takes into account the sparsity of the in-domain dataset. Let us note that the unsupervised adaptation can also be model-based, modifying the matricial parameters of the PLDA model: methods such as CORAL+ [6] apply specific correlation alignments to the out-of-domain within- and between-speaker covariance matrices. Also, score normalizations (S-norm or AS-norm) can be considered as unsupervised adaptation methods playing on scores instead of features, as they use a non-labeled in-domain subset for impostor cohort.

When speaker labels of an in-domain sample are available, supervised adaptation can be carried out. The most usual method [7] is a kind of MAP approach, that boils down to a linear interpolation between in-domain and out-of-domain PLDA parameters, as follows:

$$\begin{aligned} \mathbf{B} &= \alpha \mathbf{B}_{in} + (1-\alpha)\,\mathbf{B}_{out} \\ \mathbf{W} &= \alpha \mathbf{W}_{in} + (1-\alpha)\,\mathbf{W}_{out} \end{aligned} \quad (1)$$

where $\alpha \in [0,1]$ and $\mathbf{B}_{...}$, $\mathbf{W}_{...}$ respectively denote the between- and within- classe covariance matrices of in-domain (*in*) and out-of-domain (*out*) training datasets. As shown in Figure 1, we generalize this interpolation to all the possible stages of the system (LDA, whitening). With this tactic, performance improvements are observed on all our experiments.

2.2. Clustering

The challenge of speaker recognition with domain mismatch is greatly simplified when an in-domain speaker-annotated corpus is available. The resulting supervised methods better fit models by building on specific speaker and nuisance variabilities. In [8], it is proposed to carry out an unsupervised clustering of in-domain unlabeled data and to identify the resulting clusters with speaker-classes. The agglomerative hierarchical clustering (AHC) used in [8] takes into account all the variabilities, but assuming that the potential splitting in clusters is mainly due to the speaker-latent variable has proved to be relevant in [8]. The authors show that the exact and artificial speaker-partitions are similar (at least for their experimental setup, NIST and Switchboard domains). The trick of the method is to make use of a speaker-based clustering similarity, the log-likelihood ratio, for better focusing on the speaker partitioning. Once AHC has been carried out, the number of clusters must be determined. In [8], several methods are proposed but a new strategy was adopted during NIST SRE18 and SRE19 campaigns [9]: given a number of clusters, a supervised system (adapted by using the dataset to cluster with its hypothesized labels) computes the scores of a pre-existing in-domain development trial dataset. The optimized number of clusters is the one that provides the best accuracy on this test sample, in terms of error rate and/or cost function. This approach, that takes advantage of additional information about the target domain, turns out to be more performing than those proposed in [8], at least for language mismatch.

Table 1: *In-domain datasets used at each step of the state-of-the-art clustering.*

Inputs: •X dataset to cluster •X_{dev} pre-existing in-domain labeled development set	
step 1. AHC	
fDA	X_{dev}
LDA,WCCN,PLDA	-
S-norm cohort	X_{dev}
scored	$X \rightarrow$ *AHC*
step 2. For each #class k	
fDA	X
LDA,WCCN,PLDA	X
S-norm cohort	X
scored	$X_{dev} \rightarrow$ best perf $k_{optimal}$

Table 1 describes the steps of this method. Let us note that, in step 1 (computation of similarities for AHC), the in-domain labeled development set denoted by X_{dev} is also the learning sample of feature-based adaptation (fDA) and the impostor cohort for score-normalization. This allows to deliver similarities that are less shifted by the out-of-domain information. During step 2, the dataset X to cluster is injected at each stage of adaptation (unsupervised and supervised) as in-domain learning sample. The labeled development dataset is used to make up a trial dataset, eventually constrained to match the desired probability of the target (i.e. *same speaker*).

The requirement of pre-existing labeled data could raise concern in a scenario of adaptation from scratch, where the first data that are gathered have to fast involve a significant gain of accuracy. This issue is discussed in Section 3.

3. Adaptation strategy and clustering from scratch

Analysis of the adaptation strategy focuses on the gain of adapted systems as a function of the available data and methods. Three parameters are selected for the cross-referenced analysis: the total number of in-domain speakers, the size of speaker samples and the adaptation technique (unsupervised, supervised, both techniques). The experiments designed for this analysis are described in 4.1.2.

Clustering unlabeled in-domain data in order to take advantage of a supervised adaptation requires making available a speaker-labeled in-domain development dataset, used for finding out the optimal number of clusters. Considering the most usual scenario in real-world situation, the first data that we have quick and large access to are unlabeled. Moreover, the pre-existing labeled sample has to be informative enough, in terms of variability, to provide a robust trial dataset. We study here the ability of clustering without any in-domain speaker label, while keeping the EER-DCF criteria described above. This will allow to fast form an as accurate as possible speaker recognition system in mismatch condition.

The first step of clustering, as described in previous Table 1, can be easily carried out without pre-existing in-domain labeled set. As shown in Algorithm 1, the dataset X to cluster is used as in-domain learning sample for unsupervised adaptation. In this way, the shift induced in similarities by out-of-domain data is reduced.

The key point of the proposed approach is in step 2. How to determine the optimal number of clusters without available in-domain trials ? On the one hand, AHC is an unsupervised method, that neither relies on *a priori* nor requires annotated data. A popular method to determine the optimal number of clusters is referred to as "the elbow criterion" [10]: while aggregating clusters, the part of variance explained by the clustering in the total variance will decrease (or, equivalently, the within-cluster variance will increase). Beyond a satisfying number of cluster, this variance will suddenly fall apart. This will induce an elbow into the curve of this variance as a function of the number of clusters. This means that going on to aggregate data beyond this or these thresholds will lead to merge heterogeneous data. Let us note that, as this elbow cannot always or easily be unambiguously identified, other criteria have been introduced in the clustering field [10].

On the other hand, the clustering similarity is the LLR-score, focused on the speaker variable. This leads us to consider the common error measurement in the field (EER) as equivalent to the percentage of variance unexplained by the clustering, which is used for determining the elbow location. It is obvious that computing this error measure with speaker-labels identified with the clustering classes (that is, from keys provided by the cluster-matching) yields almost zero error in the extreme case "1 speaker = 1 class", as only scores of a vector with itself are target (*same speaker*). This case also corresponds to the maximal percentage of variance explained by the clustering, thus to a null within-cluster variance. EER can be seen as an equivalent, for our issue, to the within-cluster variance in the general case of clustering. Therefore, it is to be hoped that the curve of

EER against the number of clusters would have the same general shape, decreasing to zero with one or more intermediate elbows. Moreover, beyond the exact number of speakers, going on splitting the speaker samples would cause some observations of the same speaker (on opposite sides of the class-boundaries) to be stated as *different speakers*. The DCFs with the usual operating points of speaker recognition are very sensitive to false alarms and, thus, would degrade. The behaviour of DCF curves beyond the exact number of speakers deserves to be observed.

Algorithm 1 Clustering from scratch

Input: in-domain dataset $X = \{x_i\}_{i=1}^{n}$ to cluster

Compute PLDA model using X as in-domain training dataset for fDA, LDA, WCCN and PLDA.
Compute the $n \times n$ score matrix **S** of crossed pairs of X:
$\mathbf{S}_{i,j} = LLR(x_i, x_j)$
Compute AHC using **S** as similarity matrix.

for $q = 1$ to n **do**
Use AHC to cluster X in q classes
Set a $n \times n$ key table **K** to:
if class(x_i) = class(x_j)
$\mathbf{K}_{i,j}$ = '*same speaker*'
else
$\mathbf{K}_{i,j}$ = '*diff. speaker*'
Use score matrix **S** and key table **K** to compute EER/DCF.

Output: the class-labels corresponding to the number of classes q that satisfies the elbow criterion (these labels are identified to speaker labels).

The outcome of this study is described in Algorithm 1: the trial set for determining the optimal number of classes is formed from the dataset X to cluster itself. It can be comprised of all crossed pairs, eventually constrained to match the desired probability of the target.

4. Experiments and results

4.1. Experimental setup

4.1.1. System components

For acoustic features MFCC are extracted by using Kaldi toolkit [11] with 23 cepstral coefficients and log-energy, a cepstral mean normalization being applied with a window size of 3 seconds. Voice Activity Detection removes silence and low energy speech segments. The simple energy-based VAD uses the C0 component of the acoustic feature. Training and extracting x-vector is done by using Kaldi toolkit [11]. The training corpus encompasses Switchboard and NIST-SRE' 04, 05, 06, 08. The setting of x-vector network follows the recipe in Kaldi [1], in which we have modified the following settings. First, in order to increase the diversity of the acoustic conditions in the training set, a 5-fold augmentation strategy is used, that adds four corrupted copies of the original recordings to the training list. The data augmentation consists of adding noise, music, and mixed speech (babble) drawn from the MUSAN database [12] and adding reverberation by using simulated room impulse responses (RIR). Second, an attentive statistics pooling layer is implemented, calculating weighted means and standard deviations over frame-level features scaled by an attention model. This enables x-vectors to only focus on important frames.

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2

4.1.2. Experimental protocol

Table 2: *Datasets used for experiments.*

	#spks	#segs
data for development and clustering		
SRE18 CMN2 dev. enroll.-test	25	1741
SRE18 CMN2 eval. enroll.-test	188	13451
SRE19 CTS eval. enroll.-test:		
female (50%)	62	4890
male (50%)	35	2449
Total	**310**	**22531**
data for test		
SRE19 CTS eval. enroll.-test:		
female (50%)	63	4967
male (50%)	35	2639
Total	**98**	**7606**
Trial dataset for test	target	non-target
female	20 000	1 980 000
male	20 000	1 980 000
Total	**40 000**	**3 960 000**
	(1%)	(99%)

The domain of our experiments is the Tunisian-Arabic language, which is the core domain of the NIST-speaker recognition evaluations 2018 (referred to as CMN [13]) and 2019 (CTS). This language is far enough from those of the speaker recognition training databases for inducing severe mismatches. The in-domain corpus for development and test is described in Table 2. The development dataset merges the enrollment-test segments delivered for NIST-SRE18 development/test and half of the enrollment-test segments delivered for NIST-SRE19 test. The other half is set aside for making up a trial dataset of test (the 50% split takes gender into account to avoid eventual bias in the results). It contains 4M trial pairs, randomly and uniformly picked up with the constraints of being equalized by gender and of target prior equal to 1%.

For analyzing the adaptation strategy, two parameters, the number of speakers and the number of segments per speaker, are varied in order to sweep different total amount of segments and, also, given a fixed amount to assess the impact of speaker-class variability (are development datasets comprised of small samples of many speakers more efficient than those comprised of many segments from few speakers ?)

Each time, a subset is picked up from the 310 speakers-size development dataset and employed for adapted models. The second development dataset is fixed, and only intended for testing. Three alternatives are considered and experimented:

- a system applying unsupervised adaptation only (i.e. feature-based and score-normalization). Even if the clustering methods described above could add useful information about the in-domain distribution, this system will be used for comparison to systems using no-adaptation as well as supervised adaptation.
- a system applying supervised adaptation only (by interpolation of parameters during LDA, whitening and PLDA stages).
- a system applying the full pipeline (unsupervised and supervised). Here, the goal is to assess the usefulness of

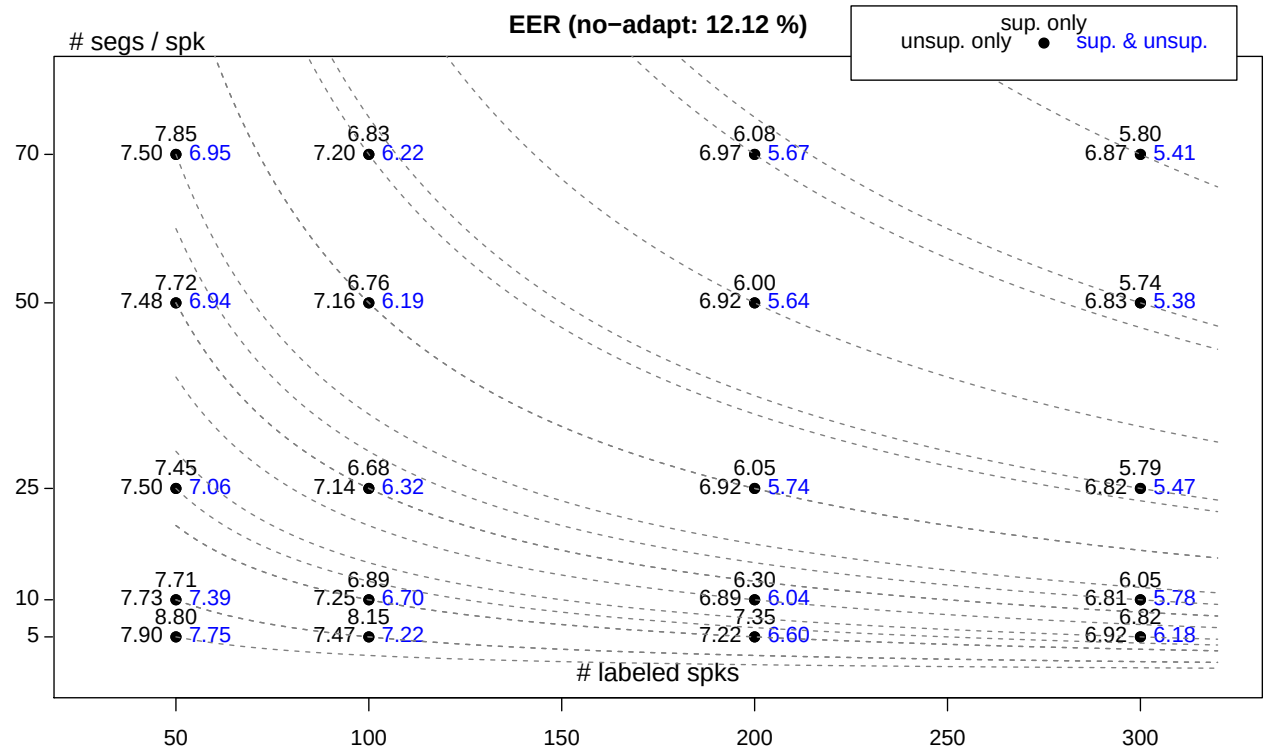

Figure 2: *Adaptation strategy: performance (EER) of systems with adaptation for three cases: unsupervised only, supervised only, both techniques. A curve corresponds to a given total amount of segments.*

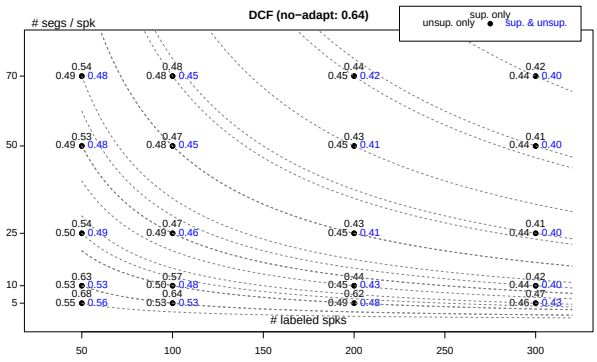

Figure 3: *Performance (DCF) of systems with adaptation for three cases: unsupervised only, supervised only, both techniques. A curve corresponds to a given total amount of segments.*

unsupervised techniques while speaker labels are available.

The optimal interpolation parameter α of equation 1 could be distinct from an experiment to another. Each time, we report the best result obtained by sweeping this parameter. This allows to assess the optimal accuracy of each sub-system.

For experimenting the clustering from scratch, several datasets are tested. First, the datasets to cluster are subsets of the previous development dataset, comprised of the same amount of speakers (250) with a fixed number of segments per speaker (successively 5, 10, 25, 50 or all available) or with a variable number of segments per speaker, randomly and uniformly selected between 5 and 25 for each speaker. The trial set for scoring is the whole set of pairs from X embeddings. To compare clustering from scratch with state-of-the-art, a development sample is generated (the dataset X_{dev} of Table 1), comprised of all the segments of the 60 remaining speakers in the development set. Lastly, efficiency of each method is assessed on the test trial dataset described in table 2. Second, in order to evaluate performance on a known evaluation, an experiment is carried out with the enrollment-test vectors of SRE18 evaluation to cluster, the enrollment-test vectors of SRE18 development as dataset X_{dev} of Table 1, Algorithm 1 and the SRE19 evaluation trial dataset for test. For all the experiments, the interpolation parameter α of equation 1 is set to 0.6, which is the chosen value in [9]

4.2. Adaptation strategy

Figure 2 shows performance in terms of equal error rate (EER) of unsupervised and supervised adapted systems depending on the number of speakers (50, 100, 200 or 300) and segments per speaker (5, 10, 25, 50 or 70) of the in-domain development dataset. The case 70 corresponds to all the segments available for these speakers (70 is the mean). The same graph for decisional cost function DCF is displayed in Figure 3. The DCF is the mean value of cost functions with target probabilities set to 0.01 and 0.05 and the cost of misses and false alarms set to 1, as proposed in NIST-SRE18 and SRE19 evaluations.

The power of such adaptation methods, mostly based on alignment of elliptic distribution parameters (mean and covariance), is outstanding even with very limited amounts of in-

domain data. The major proportion of improvement is induced by unsupervised techniques: from EER: 12.12% (resp. DCF: 0.64) to [6.92,7.90] (resp. [0.46,0.55]).

Making speaker-label available allows the use of supervised techniques, that make systems more accurate but does not exclude unsupervised techniques (feature-based and S/AS-norm). These approaches well combine in terms of accuracy, as can be observed at a glance at Figure 2 (blue results).

Each dashed curve corresponds to a given total amount of segments. By sweeping the curves, it can be observed that performance improve with the number of speakers. For example, fixing 5000 segments, 100 speakers with 50 segments yields EER: 6.19%, DCF: 0.45 whereas 200 speakers with 25 segments yields EER: 5.74%; DCF: 0.41. This demonstrates that gathering data from a few speakers, even with many utterances per speaker, limits the gain of adapted systems and, in a scenario from scratch, the delay to take advantage of domain adaptation.

This observation can be explained: the $r \times r$ matrix of between-speaker covariance (with r usually between 100 and 200 after LDA) is estimated from the speaker factors of the training dataset. With less than 300 speakers in this training sample, each of one providing one speaker factor (the expected average of the speaker vectors), hoping for an accurate estimate is unrealistic.

4.3. Clustering from scratch

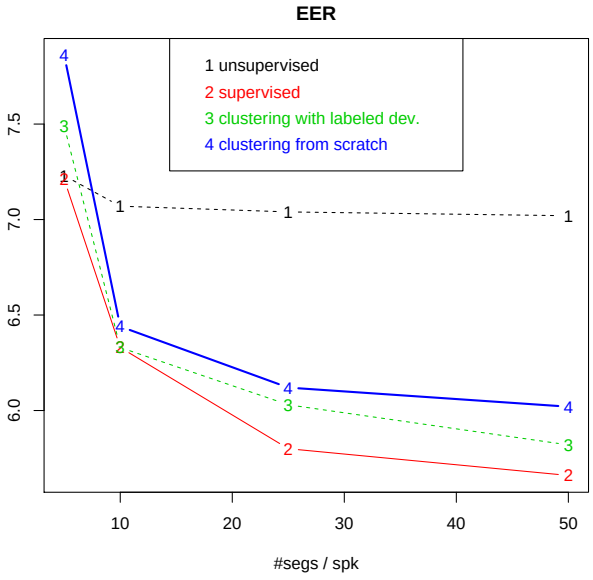

Figure 4: *Performance in terms of EER of the first four systems reported in Table 3 with the different approaches.*

The previous analysis assessed the impact of supervised adaptation. It recalls the usefulness of artificial labeling. In real-life conditions, clustering must be carried out without the need of annotated development set.

Results of our experiments are reported in Table 3 and displayed, for the first four cases, in Figure 4. It is worth noting that, for clustering from scratch, results that are reported are those with a number of clusters optimized by Algorithm 1 while, for state-of-the-art clustering, results are the best ones obtained by sweeping the number of cluster (which are not necessarily those of the optimized number of clusters determined by step 2 of Table 1). By this way, the power and relevance of the proposed clustering can be better assessed.

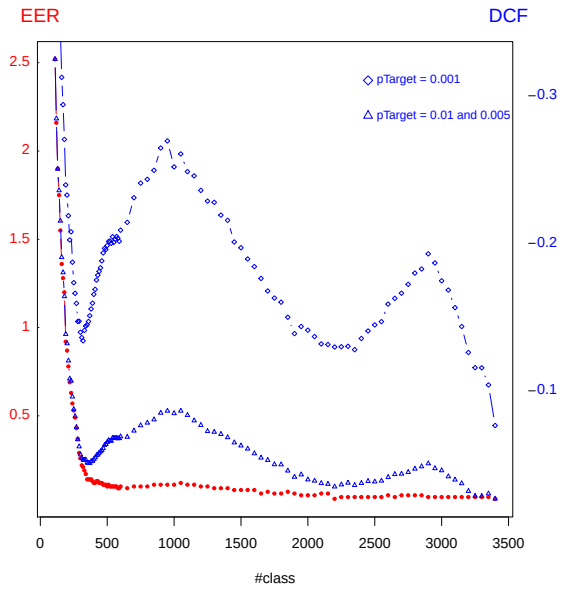

Figure 5: *Error measurements as criteria for determining the optimal number of clusters. The reported results correspond to the loop of Algorithm 1.*

As can be seen, the gap between unsupervised and supervised approaches (curves 1 et 2 of Figure 4) is greatly reduced by the artificial labelings provided by clustering methods. Above all, clustering from scratch yields similar results to the usual clustering, slightly worse in terms of EER, equal in terms of DCF. For the last experiment (last row of Table 3) with SRE18 evaluation enrollment-test vectors to cluster, 82.9% of the gap in terms of EER between unsupervised and supervised approaches is removed by the state-of-the-art clustering, 68.5% by clustering from scratch. Let us note that, for the first system (first row of Table 3) with only 5 segments per speaker, performance of clustering from scratch is disappointing but results of supervised adaptation are similar to those of unsupervised approach.

The effectiveness of this method may seem surprising. Figures 5 allows to better observe its behaviour: the Figure details the loop of Algorithm 1 for the fifth system of Table 3. EER and DCF for this development step are reported as a function of the AHC number of classes. As was presumed above, the slope of the EER curve suddenly slows down around the neighborhood by excess of the exact number of speakers (here 250). Moreover, the values of DCFs (with two operating points, in order to better analyze relevance of the method) reach local minima before converging to 0, the first one in the same neighborhood. The fact that the model was partially constrained by the initial training set (mismatched but wide in terms of information) could help to explain this local behaviour. Figure 6 details performance of the clustering from scratch on our test trials, as above for the fifth system of Table 3. The three vertical lines indicate the exact number of speakers (here 250), the number of clusters corresponding to the minimal EER rate and the one determined by the optimization-step 2 described in Figure 5. As shown, the loss of gain, compared to a system supervised by

Table 3: *Comparison of the proposed clustering from scratch to all the alternatives, supervised or unsupervised, on datasets of various sizes. For readability, results on the first four datasets are displayed in Figure 4.*

	unsupervised		**supervised**					
			with exact spk-labels		by clustering with a pre-existing labeled dev. set (best #class)		by clustering from scratch (optimized #class)	
#segs (#spks × #segs/spk)	EER	DCF	EER	DCF	EER	DCF	EER	DCF
1250 (250×5)	7.23	0.517	7.21	0.517	7.49	0.559	7.86	0.564
2500 (250×10)	7.07	0.488	6.33	0.468	6.33	0.476	6.44	0.474
6250 (250×25)	7.04	0.482	5.80	0.447	6.03	0.456	6.12	0.460
12500 (250×50)	7.02	0.483	5.66	0.452	5.82	0.459	6.02	0.459
3750 (250×[5 to 25])	7.05	0.487	6.00	0.456	6.17	0.474	6.27	0.465
To cluster: enroll.-test sre18 eval.								
13451 (188×[41 to 112])	5.10	0.401	3.99	0.371	4.18	0.378	4.34	0.376

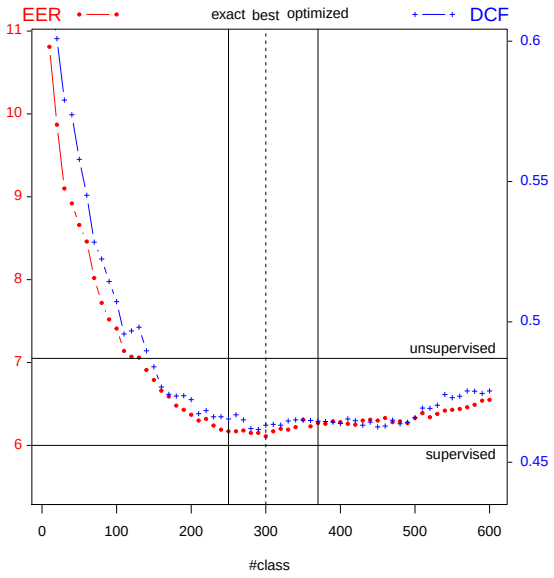

Figure 6: *Performance of the adapted system using clustering from scratch as a function of the number of clusters, compared to unsupervised and supervised (with the exact speaker labels) adaptation.*

exact labels, is very slight, demonstrating that the optimization criteria turn out to be relevant. Accordingly, two criteria can be proposed to determine the optimal number of classes: the elbow of the EER curve and the first local minimum of a DCF curve.

Note how the method overestimates the number of speakers (Figure 5) but manages to attain interesting recognition performance (Figure 6), and how the DCF curve reaches a second local minimum (Figure 5). We hope to better explain these phenomena in future work.

5. Conclusion

The challenge of speaker recognition on specific domains and, more generally, on data from various conditions has been partially overcome in recent years by the use of DNN-based embeddings, data augmentation and specific transformations into the back-end process. This paper analyzes the tasks of collecting in-domain data and making the most of the first data at hand. It is shown that the major proportion of improvement is due to unsupervised domain adaptation approaches. Even if supervised approaches well combine to achieve best accuracy, future work about adaptation in back-end process should focus on the supervised approaches which may, hopefully, enhanced. Also, it is observed that a small sample of in-domain data can significantly reduce the gap of performance, but when favoring the amount of speakers rather than of segments per speaker.

For taking advantage of supervised adaptation methods, the expensive and time-consuming process of human-assisted data annotation can be avoid by clustering but, for mismatch of language, this method requires pre-labeled in-domain data. In this paper, a new approach of artificial speaker-labeling is introduced, doing without pre-existing in-domain labeled data while achieving equivalent performance.

More broadly, the development of committees of systems, each of them fitted to a specific domain by using a small in-domain sample (with a pre-classifier computing *a posteriori* domain probability of embeddings), may be considered and, hence, suggests some fruitful avenue for future work.

6. Acknowledgements

This research was supported by the ANR agency (Agence Nationale de la Recherche), RoboVox project (ANR-18-CE33-0014).

7. References

[1] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 NIST speaker recognition evaluation," in *International Conference on Speech Communication and Technology*, 2017, pp. 1353–1357.

[2] S. O. Sadjadi, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2018 NIST speaker recognition evaluation," in *International Conference on Speech Communication and Technology*, 2019, pp. 1483–1487.

[3] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," *CoRR*, vol. abs/1511.05547, 2015. [Online]. Available: http://arxiv.org/abs/1511.05547

[4] J. Alam, G. Bhattacharya, and P. Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adap-

tation," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2018.

[5] P.-M. Bousquet and M. Rouvier, "On Robustness of Unsupervised Domain Adaptation for Speaker Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2958–2962. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1524

[6] K. A. Lee, Q. Wang, and T. Koshinaka, "The CORAL+ algorithm for unsupervised domain adaptation of PLDA," *CoRR*, vol. abs/1812.10260, 2018. [Online]. Available: http://arxiv.org/abs/1812.10260

[7] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2014, pp. 4047–4051.

[8] S. Shum, D. Reynolds, D. Garcia-Romero, , and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2014.

[9] J. Villalba and al., "The JHU-MIT System Description for NIST SRE18," in *NIST Speaker Recognition Evaluation Workshop*, 2018.

[10] T. S. Madhulatha, "An overview on clustering methods," *CoRR*, vol. abs/1205.1117, 2012. [Online]. Available: http://arxiv.org/abs/1205.1117

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[12] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[13] K. Jones, S. Strassel, K. Walker, D. Graff, and J. Wright, "Call My Net corpus: A multilingual corpus for evaluation of speaker recognition technology," in *International Conference on Speech Communication and Technology*, 2017.