INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Improving training datasets for resource-constrained speaker recognition neural networks

Pierre-Michel Bousquet, Mickael Rouvier

LIA - Avignon University

`firstname.lastname@univ-avignon.fr`

Abstract

Tackling the increase in complexity, which is now the main factor of improvement in deep learning, this paper proposes a new algorithm of data selection for training speaker recognition systems. The method starts from an initial training dataset, then the algorithm scans new data to determine the most useful speakers for completing the initial ones. The resulting training dataset improves the model, in terms of accuracy and ability to generalize, while maintaining the learning complexity reasonable. This algorithm is unsupervised, as it does not need any metadata on the new utterances and, therefore, compatible with self-supervised learning. By selecting only 30% of the speakers from a new database, the proposed algorithm is able to achieve very similar performances to the system with all speakers added.

Index Terms: subset selection, speaker recognition, embeddings, deep learning, x-vectors

1. Introduction

The success of Deep Neural Network (DNN) discriminative approaches in Speaker Recognition (SR), as in many areas of machine learning, relies on the large amount of observations used to train the model. These discriminative models do not tend to saturate as quickly as their probabilistic counterparts: experimentation has shown that their accuracy can be further improved by expanding the training databases, with no limits detected at this time. The availability of an increasing number of open datasets, the continuous inflation of information gathered by companies, have enriched these databases. In addition, the introduction of the self-supervised learning [1, 2], which allows the use of large unlabeled datasets, has definitely made it possible to build giga-databases. Consequently, their processing for modeling makes the learning phase much more complex. Moreover, data augmentation, which makes systems more robust, multiplies the size of the dataset. Even if this increasing computation burden comes with technological advances in terms of calculation speed, data storage and access, it raises some issues: on the one hand, collecting unlimited amount of data and using them for DNN training is restricted to gigantic corporations. On the other hand, as remarked in [3], ”deep neural network training spends most of the computation on examples that are properly handled, and could be ignored.”

The challenge of training systems with a limited amount of data or a selection of the most efficient data has been addressed in several studies and from various views. Some studies have revealed the relativity of training data in terms of modeling information [4, 5]. Importance sampling schemes have been proposed [3, 5, 6, 7, 8, 9, 10, 11], which emphasize the most informative examples during the iterative learning process, in order to improve and accelerate the training of neural network [12]. Training a system for SR with limited amount of data has been analyzed [13, 14], focusing on the number of speakers, sessions per speaker, duration of the utterance, etc...

The task of subset selection attempts to find a small subset of most informative items from a ground set. It has found numerous applications (see for example references 1-20 of [6]). Finding a well-optimized subset is an NP-hard problem [15], which requires heuristics or meta-heuristics to find near-optimal solutions. Genetic algorithms have been used, as an undersampling technique [16] or to optimize parameters and select features [17]. To be applied to SR, a subset selection algorithm must fit the current x-vector representation produced by deep learning.

The problem addressed here is described in the following. To the best of our knowledge, this problem has not been tackled in SR, nor in related fields. First, a DNN model was learned using an initial training dataset. Next, a new utterance dataset, comprised of a large number of speakers unknown to the system, is collected. Their identities can be available or estimated by pseudo-labels. Then, the initial model is used to extract the x-vectors of the new utterances. The goal of this study is to select a small sample of this new dataset, by using an unsupervised method (no other information provided about these utterances: language, gender, channel, background noise, sound environment, other metadata), then to use this selected sample as additional corpus for training a new DNN. The algorithm should be done in a way that significantly improves the accuracy and ability to generalize of the system, while keeping a reasonable learning complexity. The proposed method will select speakers, not single utterances one by one, to avoid gathering outliers or degraded segments.

The paper is organized as follows: Section 2 describes the framework of our approach and the resulting algorithm, Section 3 details our experimental setup and discusses the results we obtained on various evaluation corpora, then we conclude in Section 4.

2. Subset selection method

The subset selection method presented here is based on output distributions and clusterings. We detail them below, then introduces our proposed selection metric and algorithm.

2.1. Framework of the method

Once the learning phase is achieved, it is possible to extract, from both training and test utterances, the conditional distributions over labels $i \in \{1...N\}$ given the x-vector x through a softmax function $p(i|x) = e^{sz_i} / \sum_{j=1}^{N} e^{sz_j}$ where z_i are the logits and s is a temperature scaling. The penalty margin [18],

10.21437/Interspeech.2023-210

typically used during the learning phase for fastening convergence and limiting overfitting, is ignored in the rest of the paper.

Our method of selection also relies on clustering. An agglomerative hierarchical clustering (AHC) is carried out on the output distributions of the initial training dataset. The distance between two training speakers can be the symmetric Kullback-Leibler divergence (often referred to as Jeffreys divergence) between the speaker samples. Denoting by χ_1, χ_2 the samples of two initial training speakers s_1, s_2 and by $p(x) = [p(i|x)]_{i=1,N}$ the output distribution presented above, this distance is equal to:

$$\mathbf{J}(s_1, s_2) = \frac{1}{|\chi_1|\,|\chi_2|} \sum_{x_1 \in \chi_1} \sum_{x_2 \in \chi_2} \left[D_{KL}(p(x_1)\,||p(x_2)) + D_{KL}(p(x_2)\,||p(x_1)) \right] \tag{1}$$

As these output distributions are fitted by the model during the learning phase to discriminate the speaker labels, clustering should be inefficient on it: indeed, the output distributions of the initial speaker utterances are, ideally, one-hot-target vectors equal to 1 for the target label 0 otherwise, leading to infinite entropy between the speakers.

AHC allows to provide clusterings in any number of classes. The good results we observed with these clusterings, in terms of inter-class variability, confirmed their ability to capture the latent variabilities contained in the training data. These latent variables are more or less independent of the speaker identity and interfere with it during modeling. Therefore, they tend to limit the discrimination between the speakers used for training, maintaining some proximity between speaker-groups. This property of AHC on training data is an opportunity to identify underlying speech attributes that are underrepresented in the initial training dataset.

2.2. Metric used for selection

It can be remarked that a great entropy of an output distribution (too smooth, too spread between the speaker labels) can reveal a difficulty of the model to classify a new speaker, thus its originality with regard to the initial training speakers. But analyses we conducted have shown that the current metrics of the information theory are too poor to detect the most useful data of a new dataset.

To find out the most complementary samples inside an available new speaker dataset, the measure proposed here relies on the ratio of probability of an observation x (an x-vector) a posteriori of a class (in one of the previous clusterings, carried out on the initial training dataset) and a priori. Defining the output distribution of the sample χ of a new speaker s as the average vector $p(i|s) = \left[\frac{1}{|\chi|}\sum_{x\in\chi} p(i|x)\right]_{i=1,N}$ this ratio can be written:

$$l\left(s, C_k^{(K)}\right) = \frac{P\left(s|C_k^{(K)}\right)}{P(s)} = \frac{P\left(C_k^{(K)}|s\right)}{P\left(C_k^{(K)}\right)} = \frac{\sum_{i\in C_k^{(K)}} p(i|s)}{f_k^{(K)}} \tag{2}$$

where $C_k^{(K)}$ is the k^{th} class for the clustering of the initial training dataset in K classes, $f_k^{(K)}$ its frequencies $\left|C_k^{(K)}\right|/N$ and the notation $i \in C_k^{(K)}$ means that the i^{th} training speaker belongs to $C_k^{(K)}$. The prior $P\left(C_k^{(K)}\right)$ is identified to $f_k^{(K)}$ (same probability of the speakers inside a class) and the probability of a label a posteriori of a speaker s is identified to the output distribution $p(i|s)$. In what follows, we refer to this ratio $l\left(s, C_k^{(K)}\right)$ as the *lift*, by analogy with the theory of association rules [19].

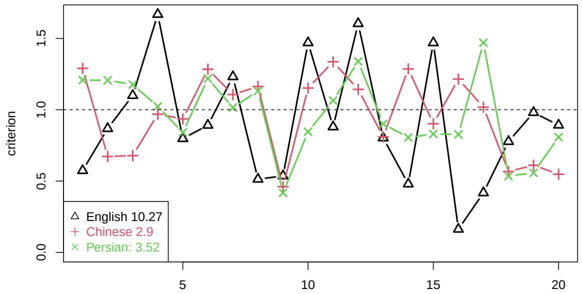

Figure 1: *Three examples of series of lifts (equation 2) from the clustering of the initial training dataset in* 20 *classes, for three speakers: English, Chinese and Persian. The value in the legend is the criterion of equation 3.*

2.3. Originality criterion

A value of lift greater than 1 indicates that the knowledge of a class increases the probability of the observation. Conversely, a low value reflects some independence between the observation and the class. The series of lifts of a new speaker along classes 1 to K of a K-size clustering can provide useful information about the originality of this speaker. When some lifts are higher than one while the majority are low, it means that this speaker sample contains some values of latent variables that are well represented in the initial training dataset, hence can be considered as *in-domain*. Conversely, a too smooth series of lifts indicates that this speaker does not match these variabilities.

As the series of lifts is not a probability distribution, entropy cannot be computed on it. Moreover, the originality criterion must take into account the specificity of the lift ratio. The criterion proposed here for selecting new speakers is:

$$L(s) = \frac{1}{K_M - 1} \sum_{K=2}^{K_M} \frac{\max_{k\in[1,K]} l\left(s, C_k^{(K)}\right)}{\min_{k\in[1,K]} l\left(s, C_k^{(K)}\right)} \tag{3}$$

The selected speakers are those who minimize this criterion [1]. Figure 1 shows an example of series of lifts, for three speakers, English, Chinese and Persian, unknown to the model, with the clustering in $K = 20$ classes computed on the initial training dataset of Voxceleb2 (details in Section 3). The values of the criterion $L(s)$ are reported in the legend. It can be seen that the series of lift of the English speaker, which is ”in-domain” since the majority of VoxCeleb2 speakers are English-native, is more varied than those of the other two, leading to a high value of the criterion ($= 10.27$). The value of the criterion for the Persian speaker is slightly higher than for the Chinese speaker, due to a peak for the 17^{th} class. Even if this graph only depicts a single example, it confirms the link between this clustering and the latent variable of language contained in the

[1] Note that other criteria that we tested (based on variance, etc...) proved to be less efficient.

utterances, and recalls that this variability is preserved in the x-vectors. This allows the method to be tested on other underlying speech attributes.

3. Experiments

Table 1: *Number of speakers per family of language in the* 5994*-size Voxceleb2 training dataset.*

language	#spk		language	#spk	
Arabic	60	1%	Misc	416	6.9%
East Asia	198	3.3%	North-Europa	926	15.4%
English	2875	48%	South-Europa	694	11.6%
Indian	472	7.9%	Spanish	353	5.9%

Table 2: *The new dataset candidate to complete the initial Vox-Celeb2 training dataset: numbers of utterances, speakers and percentage of each corpus among the* 4245 *new speakers.*

corpus	**language**	**#utt**	**#spk**	**% spk**
LibriSpeech	English	281241	2338	55,1%
MLS	French	15173	142	3,3%
	German	23155	176	4,1%
	Italian	11309	65	1,5%
	Polish	1254	11	0,3%
	Portuguese	2073	42	1%
	Spanish	12945	86	2%
DeepMine	Persian	85764	588	13,9%
Cn-celeb1	Chinese	107951	797	18,8%
Total		540865	4245	100%

3.1. Experimental setup

The x-vector extractor used in this paper is a variant based on ResNet-101. The extractor was trained on the development part of the Voxceleb 2 dataset [20], cut into 4-second chunks and augmented with noise, as described in [21] and available as a part of the Kaldi-recipe. It contains about 1M segments (+ 4M augmented) of 5994 speakers. As input, we used 60-dimensional filter-banks. The speaker embeddings are 256-dimensional and the loss is the angular additive margin with temperature scaling equal to 30 and margin equal to 0.2. The ResNet architecture have four residual blocks, each with feature maps of sizes 128, 128, 256, and 256, and block multipliers of 3, 12, 15, and 3. The activation function used in the ResNet is SiLU. We use stochastic gradient descent with momentum equal to 0.9, a weight decay equal to 2.10^{-4} and initial learning rate equal to 0.2. The implementation is based on PyTorch. The loss function minimized during the learning phase is the cross-entropy completed by label-smoothing [22, 23]. For scoring, we use the cosine metric.

The method of selection is applied on a dataset comprised of 4245 speakers gathered from four free datasets :

- the English part of Librispeech [24], a dataset derived from read audiobooks from LibriVox,
- a sub-corpus of the Multilingual LibriSpeech (MLS) [25], in which we select all the languages except English,
- the DeepMine [26, 27] dataset of Persian speakers used for training during SDSV challenge 2020 Task 2 [28],
- Cn-celeb1 speaker recognition training dataset [29], collected from Chinese celebrities.

Details of the dataset are reported in Table 2. For comparison, Table 1 reports the families of languages of the initial training dataset VoxCeleb2.

The relevance of the subset selection algorithm is tested on the following evaluation sets. VoxCeleb1-E and H (cleaned versions) [30, 31]. VoxCeleb1-E Noisy dataset [2], derived from VoxCeleb1-E cleaned by augmenting it with the Microsoft Scalable Noisy Speech Dataset (MS-SNSD) [3] [32]. SdSV challenge Task 2 Partition 2 [28] in which all the speakers are Persian native but the second partition consists of cross-language trials where the enrollment utterances are in Persian while the test utterances are in English. TED-x Spanish [4], derived from the public-available created from TED talks in Spanish, consisting of segments having a duration range between 3 and 10 seconds in which we randomly selected 2M pairs (1.6M non-targets pairs and 0.4M targets pairs), all with the same gender. DiPCo [33], a far-field speaker verification corpus issued from DiPCo corpus. The results are reported in terms of Equal Error-Rate (EER) and normalized minimal detection cost (DCF) with the probability of a target trial set to 0.01 and the cost of miss-detection and false alarm set to 1. For computing the criterion of equation 3, the clusterings are swept up to $K_M = 100$. As the systems reported for evaluation rely on distinct training datasets, the number of iterations and learning rates of the network are tailored to their size.

3.2. Results

Table 3 shows the results of four systems used for testing the selection method: the first system, also known as *Sys 1*, is trained on the initial dataset of VoxCeleb2 only (5994 speakers). Then, 1200 of the 4245 speakers of the new dataset are added to the training set: randomly chosen (row 2 of the table, *Sys 2*) or by using the selection method (row 3, *Sys 3*).

For VoxCeleb evaluations (VoxCeleb1-E, 1-H Cleaned and 1-E Noisy), the results obtained are not surprising, as the accuracy is increased by inflating the training set. But the randomly picked up subset in system 2 contains 668 English speakers versus 269 only in system 3 using the method. These results show that the unsupervised approach detects the most informative data and can boost performance on evaluations that are considered fully in-domain. For VoxCeleb1-E noisy (where the noise augmented data of the new utterances improve performance), the method is also slightly more accurate than random. For the Cn-celeb1 Chinese evaluation, the method yields a much more accurate system compared to the baseline. In terms of EER as well as DCF, performance are very competitive, even compared to systems found in the literature that are trained using larger Chinese datasets. For SdSV Task 2 partition2, as explained above, this partition combines two difficulties : (1) a mismatch of language between enrollment and test and (2) utterances of Persian but also non-native English speakers (two specificities that are underrepresented in the training data). The gain of adding 1200 new speakers is spectacular. But random and method selections yield similar performance, highlighting the difficulty to deal with mismatch between enrollment and test without asymmetric modeling. For the Spanish evaluation Ted-

[2] http://dipco-sre.univ-avignon.fr/downloads/voxceleb1.tar.gz

[3] https://github.com/microsoft/MS-SNSD

[4] http://dipco-sre.univ-avignon.fr/downloads/tedx_spanish.tar.gz

Table 3: *Evaluation of the subset selection method: starting from the system of row 1, stated as baseline, comparison of the benefits of the method over random selection (row 3 vs 2). The last system (row 4), with the overall dataset, allows to estimate the impact of the method when the training size grows. This impact can be assessed more easily in the last row, with the ratios of relative EER and DCF.*

Training dataset				**Evaluations**													
				VoxCeleb						Cn-celeb1		SdSV		Ted-X		DiPCo	
	VoxCeleb2	New	**Total**	1-E		1-H		1-E noisy				Task 2 Part.2		Spanish		Far-Field	
	#spk	#spk	**#spk**	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF
sys1	5994	-	**5994**	1.00	0.105	1.70	0.164	3.84	0.294	11.70	0.513	3.75	0.304	1.94	0.154	5.63	0.337
sys2	5994	1200 (random)	**7194**	0.92	0.100	1.60	0.151	3.44	0.280	8.77	0.441	2.60	0.231	1.66	0.144	5.81	0.340
sys3	5994	1200 **(method)**	**7194**	0.90	0.095	1.55	0.156	3.37	0.268	6.87	0.380	2.51	0.251	1.43	0.105	5.53	0.342
sys4	5994	4245 (all)	**10239**	0.85	0.093	1.48	0.144	3.21	0.27	6.57	0.363	1.66	0.131	1.39	0.090	5.85	0.320
ratios of relative EER and DCF: $\frac{val_{sys1} - val_{sys3}}{val_{sys1} - val_{sys4}}$				67 %	83 %	68 %	40 %	75 %	108 %	94 %	89 %	59 %	31 %	93 %	77 %	(-)	-29 %

X, there are 20 Spanish speakers randomly picked up versus only 13 selected with the method. However, system 3 achieves a very significant gain of performance, confirming the ability of the method to expand the scope of the model. For the far-field corpus DiPCo, system 2 degrades performance, whereas system 3 yields a significant gain of performance (in terms of EER), despite the lack of matching data in the training corpus.

System 4 is trained by using all the speakers: initial and new. Comparing this system to system 3 shows that the method contributes significantly to reducing error rates as the training size grows. This can be observed by comparing the ratios of size $1200/4245 \approx 28\%$ and relative EER or DCF $\frac{val_{sys1} - val_{sys3}}{val_{sys1} - val_{sys4}}$ which are reported in the last row of Table 3. For DipCo, the performance in terms of EER, compared to the baseline, is better with the method but worse with the overall dataset. This reminds that inflating the training dataset does not necessarily make the system more robust. Overall, the system using the method is much more robust than the original system and much less complex than the one using all the data.

3.3. Analysis

The method aims to detect and retain all attributes that are underrepresented in the original data, using an unsupervised approach. We propose to better assess this ability in terms of linguistic variability. Figure 2 shows the number of selected speakers per corpus by increasing order of the criterion, from the most relevant on the left to the first 1200 on the right. The method automatically selects Chinese speakers primarily, which are sparsely represented in the initial VoxCeleb2 training dataset. 85% of the available speakers of Cn-celeb1 are selected, whereas Cn-celeb1 represents only 19% of the whole dataset. Similar numbers of speakers of LibriSpeech and SdSV are selected, but only 12% of the available data from the former versus 34% from the latter, whereas their frequencies in the whole dataset are 55% and 14% respectively. Persian and Chinese languages are poorly represented in VoxCeleb2, but the figure shows that modeling Persian by analogy with other languages contained in VoxCeleb2 is easier than for Chinese. Selected speakers from other corpora remind that language is only one cause of variability in the speech signal.

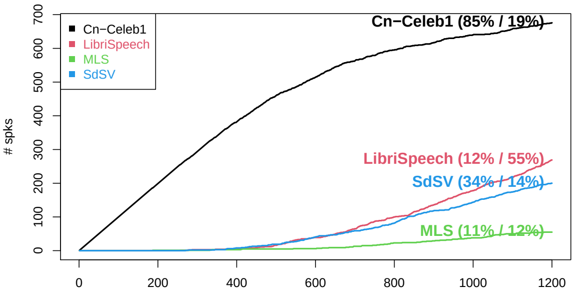

Figure 2: *Number of selected speakers per corpus, in increasing order of size of the selected subset. The percentages indicate the proportions of speakers inside the selected subset (e.g.* 85% *of Cn-celeb1 corpus are in the* 1200 *selected) then inside the whole dataset (e.g.* 19% *of the speakers belong to Cn-celeb1).*

4. Conclusion

By the time of massive collection of supervised data and of self-supervised learning, the subset selection of training data is likely to be a necessity for all organizations wishing to improve the accuracy of their systems within the limits of their computing power. Overcoming the issue of sparse training data has favored the generalization of discriminative neural network approaches, which perform well when fed with sufficiently large and various training data. A new algorithm of subset selection for speaker recognition model training is proposed here, which takes into account the specificity of the current x-vector produced by deep neural network and keeps the learning complexity reasonable. Its benefits, in terms of accuracy and generalization capabilities, are tested and demonstrated on various evaluation sets.

The selection method is unsupervised (no metadata required on the utterances to select) and, therefore, more general than domain adaptation focused on a specific variability. The advantage of an unsupervised method stems from the following observation: some underrepresented attributes of the speech signal can be fitted relatively well by the model, by combining what is available to it (e.g., a language from a family of languages, a sound environment from similar environments). Therefore, only an unsupervised approach can determine the most representative samples to efficiently complement the model.

Future work will extend this study, to test the ability of the method to detect all underlying speech signal attributes that may be underrepresented in initial training material. The application of the method in other fields than speaker recognition will also be considered.

5. Acknowledgements

This work was supported by the VoicePersonae project financed by the French National Research Agency (ANR) under contract ANR-18-JSTS-0001 and it was performed using HPC resources from GENCI-IDRIS (Grant 2022-AD011013257R1)

6. References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 2021.

[3] A. Katharopoulos and F. Fleuret, "Not all samples are created equal: Deep learning with importance sampling," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. International Conference on Machine Learning*, Montreal, Quebec, 2009.

[5] P. Xu, A. Gunawardana, and S. Khudanpur, "Efficient subsampling for training complex language models," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1128–1136.

[6] E. Elhamifar and M. C. De Paolis Kaluza, "Subset selection and summarization in sequential data," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] G. Alain, A. Lamb, C. Sankar, A. Courville, and Y. Bengio, "Variance reduction in SGD by distributed importance sampling," in *ICLR*, 2016.

[8] I. Loshchilov and F. Hutter, "Online batch selection for faster training of neural networks," in *ICLR*, 2016.

[9] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 118–126.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering." in *CVPR*. IEEE Computer Society, 2015, pp. 815–823.

[11] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *ICLR*, 2016.

[12] S. Garcia, J. Derrac, J. R. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 417–435, 2012.

[13] N. Vaessen and D. A. van Leeuwen, "Training speaker recognition systems with limited data," in *Interspeech*, 2022.

[14] M. K. Nammous, K. Saeed, and P. Kobojek, "Using a small amount of text-independent speech data for a bilstm large-scale speaker identification approach," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 3, 2022.

[15] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, pp. 1157–1182, 2003.

[16] J. Ha and J. Lee, "A new under-sampling method using genetic algorithm for imbalanced data classification," *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, vol. 3, pp. 1157–1182, 2003.

[17] M. R. G. Raman, N. Somu, K. Kannan, R. Liscano, and V. S. S. Sriram, "An efficient intrusion detection system based on hypergraph - genetic algorithm for parameter optimization and feature selection in support vector machine," *Knowl. Based Syst.*, vol. 134, pp. 1–12, 2017.

[18] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[19] S. Tuffery, *Data Mining and Statistics for Decision Making*, J. W. . Sons, Ed. Chichester, GB, 2011.

[20] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech*, Sep 2018.

[21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*. IEEE, 2018, pp. 5329–5333.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[23] P.-M. Bousquet and M. Rouvier, "Jeffreys divergence-based regularization of neural network output distribution applied to speaker recognition," in *ICASSP*, 2023.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[25] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech*, 2020.

[26] H. Zeinali, H. Sameti, and T. Stafylakis, "DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in Persian and English." in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2018, pp. 386–392.

[27] H. Zeinali, L. Burget, and J. Cernocky, "A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database," in *Proc. ASRU 2019 The 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.

[28] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "SdSV Challenge 2020: Large-Scale Evaluation of Short-Duration Speaker Verification," in *Interspeech*, 2020, pp. 731–735.

[29] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *ICASSP*, 2020.

[30] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Interspeech*, Aug 2017.

[31] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP*, 2019, pp. 5791–5795.

[32] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," in *Interspeech*, 2019.

[33] M. Rouvier and M. Mohammadamini, "Far-field speaker recognition benchmark derived from the DiPCo corpus," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, 2022, pp. 1955–1959.