IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Deep Learning for i-Vector Speaker and Language Recognition: A Ph.D. Thesis Overview

Omid Ghahabi

TALP Research Center, Department of Signal Theory and Communications
Universitat Politecnica de Catalunya - BarcelonaTech, Spain
omid.ghahabi@upc.edu

Abstract

Recent advances in Deep Learning (DL) technology have improved the quality of i-vectors but the DL techniques in use are computationally expensive and need speaker or/and phonetic labels for the background data, which are not easily accessible in practice. On the other hand, the lack of speaker-labeled background data makes a big performance gap, in speaker recognition, between two well-known cosine and PLDA i-vector scoring techniques. This thesis tries to solve the problems above by using the DL technology in different ways, without any need of speaker or phonetic labels. We have proposed an effective DL-based backend for i-vectors which fills 46% of this performance gap, in terms of minDCF, and 79% in combination with a PLDA system with automatically estimated labels. We have also developed an efficient alternative vector representation of speech by keeping the computational cost as low as possible and avoiding phonetic labels. The proposed vectors are referred to as GMM-RBM vectors. Experiments on the core test condition 5 of the NIST SRE 2010 show that comparable results with conventional i-vectors are achieved with a clearly lower computational load in the vector extraction process. Finally, for the LID application, we have proposed a DNN architecture to model effectively the i-vector space of languages in the car environment. It is shown that the proposed DNN architecture outperforms GMM-UBM and i-vector/LDA systems by 37% and 28%, respectively, for short signals 2-3 sec.

Index Terms: Deep Learning, Speaker Recognition, i-Vector, Deep Neural Network, Deep Belief Network, Restricted Boltzmann Machine

1. Introduction

The successful use of Deep Learning (DL) in a large variety of signal processing applications, particularly in speech processing (e.g., [1, 2, 3]), has inspired the community to make use of DL techniques in speaker and language recognition as well. A possible use of DL techniques in speaker recognition is to combine them with the state-of-the-art i-vector [4]. However, the main problem is that the use of DL increases highly the computational cost of the i-vector extraction process and phonetic and/or speaker labels are required for training, which are not always accessible (e.g., [5, 6, 7, 8, 9]).

Another possible use of DL is to represent a speech signal with a single low dimensional vector using a DL architecture, rather than the traditional i-vector algorithm. These vectors are often referred to as speaker embeddings (e.g., [10, 7, 11, 12, 13]). The need of speaker labels for training the network is one of the disadvantages of these techniques. Moreover, speaker embeddings extracted from hidden layer outputs are not so compatible with Probabilistic Linear Discriminant Analysis (PLDA) backend [14, 15] as the posterior distribution of hidden layer outputs are usually not truly Gaussian.

The first objective in this thesis is to make use of deep architectures for backend i-vector classification in order to fill the performance gap between the cosine (unlabeled-based) and PLDA (labeled-based) scoring baseline systems given unlabeled background data. The second one is to develop an efficient framework for vector representation of speech by keeping the computational cost as low as possible and avoiding speaker and phonetic labels. The last main objective is to make use of deep architectures for backend i-vector classification for Language Identification (LID) in intelligent vehicles. In this scenario, LID systems are evaluated using words or short sentences recorded in cars in four languages, English, Spanish, German, and Finnish.

The three main objectives are summarized in sections 2-4. Section 5 describes the experimental results. Section 6 lists the publications resulted from the Ph.D. thesis and section 7 concludes the paper.

2. Deep Learning Backend for i-Vector Speaker Verification

We have proposed the use of DL as a backend in which a two-class hybrid Deep Belief Network (DBN)-Deep Neural Network (DNN) is trained for each target speaker to increase the discrimination between target i-vector/s and the i-vectors of other speakers (non-targets/impostors) (Fig. 2). Proposed networks are initialized with speaker-specific parameters adapted from a global model, which is referred to as Universal Deep Belief Network (UDBN). Then the cross-entropy between the class labels and the outputs is minimized using the back-propagation algorithm.

DNNs usually need a large number of input samples to be trained efficiently. In speaker recognition, target speakers can be enrolled with only one sample (single session task) or multiple samples (multi-session task). In both cases, the number of target samples is very limited. A network trained with such limited data is highly probable to overfit. On the other hand, the

The Ph.D. thesis has been carried out under supervision of Prof. Javier Hernando in TALP Research Center, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya - BarcelonaTech, Spain. The thesis was supported in part by the Spanish projects TEC2010-21040-C02-01, PCIN-2013-06, TEC2015-69266-P, and TEC2012-38939-C03-0 and the European project PCIN-2013-067. The Author is now with EML European Media Laboratory GmbH, Heidelberg, Germany. The full thesis manuscript can be found Online in http://hdl.handle.net/2117/118780.

10.21437/IberSPEECH.2018-37

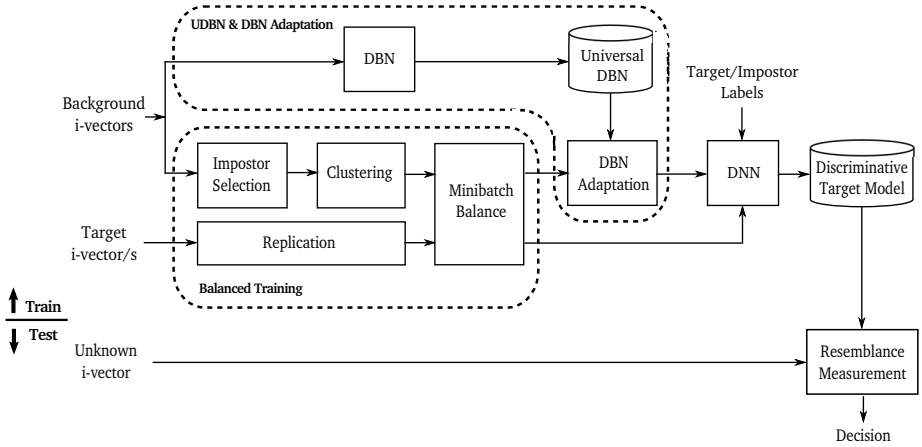

Figure 1: Block-diagram of the proposed DL-based backend on i-vectors for target speaker modeling.

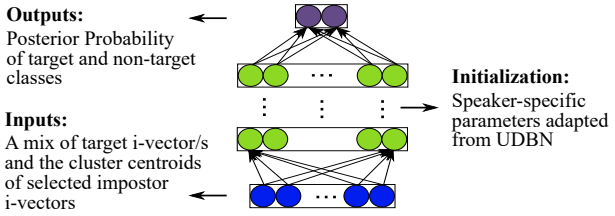

Figure 2: Proposed deep learning architecture for training of each speaker model.

number of target and impostor samples will be highly unbalanced, i.e., one or some few target samples against thousands of impostor samples. Learning from such unbalanced data will result in biased DNNs towards the majority class.

Fig. 1 shows the block diagram of the proposed approach. Two main contributions are proposed in this thesis to tackle the above problems. The balanced training block attempts to decrease the number of impostor samples and, on the contrary, to increase the number of target ones in a reasonable and effective way. The most informative impostor samples for target speakers are first selected by the proposed impostor selection algorithm. Afterwards, the selected impostors are clustered and the cluster centroids are considered as final impostor samples for each target speaker model. Impostor centroids and target samples are then divided equally into minibatches to provide balanced impostor and target data in each minibatch. On the other hand, the DBN adaptation block is proposed to compensate the lack of input data. As DBN training does not need any labeled data, the whole background i-vectors are used to build a global model, which is referred to as Universal DBN (UDBN). The parameters of the UDBN are then adapted to the balanced data obtained for each target speaker. At the end, given the target/impostor labels, the adapted DBN and the balanced data, a DNN is discriminatively trained for each target speaker. More details can be found in [16].

3. RBMs for Vector Representation of Speech

Recently, the advances in DL have improved the quality of i-vectors, but the DL techniques in use are computationally expensive and need phonetic labels for the background data. It has been proposed in this thesis an alternative vector-based representation for speakers in a less computationally expensive manner with no use of any phonetic or speaker labels.

RBMs are good potentials for this purpose because they have good representational powers and they are unsupervised and computationally low cost. It is assumed in this work that the inputs of RBM, i.e., visible units, are GMM supervectors and the outputs, i.e., hidden units, are the low dimensional vectors we are looking for. The RBM is trained given the background GMM supervectors and will be referred to as URBM. The role of the URBM is to learn the total session and speaker variability among the background supervectors. Different types of units and activation functions can be used for training the URBM but we have proposed a variant of ReLU, which will be referred to as Variable ReLU (VReLU), for this application. It will be shown in section 5 that the proposed VReLU does not suffer from the problems with sigmoid and ReLU and works the best. After training the URBM, the visible-hidden connection weight matrix is used to transform unseen GMM supervectors to lower dimensional vectors which will be referred to as GMM-RBM vectors in this work.

In fact, the proposed VReLU is defined as follows and is compared with ReLU function in Fig. 3,

$$f(x) = \begin{cases} x & x > \tau \\ 0 & x \leq \tau \end{cases} \quad , \quad \tau \in N(0,1) \tag{1}$$

Given the GMM supervectors and the URBM parameters, the GMM-RBM vectors are extracted as follows,

$$\boldsymbol{\omega}_r = \boldsymbol{W}\boldsymbol{\Sigma}_{ubm}^{-1/2}\mathcal{N}^{-1}(\boldsymbol{u})\tilde{\mathcal{F}}(\boldsymbol{u}) \tag{2}$$

where $\boldsymbol{\Sigma}_{ubm}$ is the diagonal covariance matrix of the UBM, $\boldsymbol{W}$ is the connection weights from URBM, and $\mathcal{N}(\boldsymbol{u})$ and $\tilde{\mathcal{F}}(\boldsymbol{u})$ are zeroth and centralized first order Baum-Welch statistics, respectively.

Like in case of i-vectors, resulting GMM-RBM vectors are mean normalized and whitened using the mean vector and the whitening matrix obtained on the background data.

The comparison of equation 2 with that of i-vector in equation 3 implies clearly that GMM-RBM vector extraction needs much less computational load. More details can be found in [16].

$$\boldsymbol{\omega} = \left(\boldsymbol{I} + \boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\mathcal{N}(\boldsymbol{u})\boldsymbol{T}\right)^{-1}\boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\tilde{\mathcal{F}}(\boldsymbol{u}) \tag{3}$$

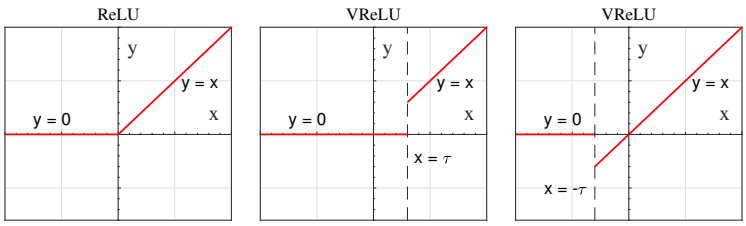

Figure 3: Comparison of ReLU and proposed VReLU. τ is randomly selected from a normal distribution with zero mean and unit variance per each hidden unit and per each input sample. (b) and (c) are two examples of VReLU with positive and negative τ.

4. Deep Learning Backend for i-Vector Language Identification

Figure 4 shows the architecture of DNNs we have proposed in this work. The inputs are i-vectors and the outputs are the language class posteriors. The softmax and sigmoid are used as the activation functions of the internal and the output layers, respectively. In order to Gaussianize the output posterior distributions, we have proposed to compute the output scores in Log Posterior Ratio (LPR) forms as in [16].

As the response time of the LID system is important in the car, the computational complexity of the classifier should also be taken into account. Therefore, we have proposed to choose the size of the first hidden layer as the lowest power of 2 greater than the input layer size. From the second hidden layer towards the output, the size of each layer will be half of the previous layer. For example, the configuration of a 3-hidden-layer DNN will be as 400-512-256-128-4, where 400 is the size of the input i-vectors and 4 is the number of language classes. It will be shown in section 5 that, in this way, we can decrease the computational complexity to a great extent while keeping the classification accuracy.

Two forms of i-vectors are considered as inputs to DNNs, raw i-vectors and session-compensated i-vectors. LDA and WCCN are two commonly used techniques for session variability compensation among i-vectors. Although LDA performs better than WCCN for the LID application when cosine scoring is used, we will use only WCCN session-compensated i-vectors as the inputs to DNNs. This is because the number of the language classes is very few in this application and, therefore, the maximum number of meaningful eigenvectors will be also few (number of classes minus one). We implemented different DNN architectures with LDA-projected i-vectors as inputs but no gain was observed. The use of raw i-vectors is advantageous as no language-labeled background data is required. More details can be found in [16].

5. Experimental Results

This section summarizes the main results obtained on the experiments for each main contribution presented in sections 2-4.

The full database provided in the National Institute of Standard and Technology (NIST) 2014 speaker recognition i-vector challenge [17] is used for the experiments in section 2. Rather than speech signals, i-vectors are given directly by NIST in this challenge to train, test, and develop the speaker recognition systems. This enables system comparison more readily with consistency in the front-end and in the amount and type of the background data [17]. Three sets of 600-dimensional i-vectors are provided: development, train, and test consisting

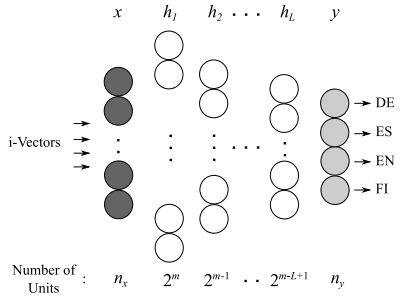

Figure 4: Proposed DNN architecture used for i-vector language identification (L denotes the number of hidden layers and $m = \lceil \log_2^{n_x} \rceil$).

Table 1: Performance comparison of the proposed DNN system with other baseline systems on NIST 2014 i-vector challenge.

Unlabeled Background Data	**Progress Set**		**Evaluation Set**	
	EER (%)	minDCF	EER (%)	minDCF
[1] cosine	4.78	0.386	4.46	0.378
[2] PLDA (Estimated Labels)	3.85	0.300	3.46	0.284
[3] Proposed DNN-1L	5.13	0.327	4.61	0.320
[4] Proposed DNN-3L	4.55	0.305	4.11	0.300
Fusion [2] & [4]	**2.99**	**0.260**	**2.70**	**0.243**
Labeled Background Data				
[5] PLDA (Actual Labels)	2.23	0.226	2.01	0.207
Fusion [2] & [5]	2.04	0.220	1.85	0.204
Fusion [4] & [5]	2.13	0.221	2.00	0.196
Fusion [2] & [4] & [5]	**1.88**	**0.204**	**1.74**	**0.190**

of 36,572, 6530, and 9634 i-vectors, respectively. The number of target speaker models is 1306 and for each of them five i-vectors are available. Each target model will be scored against all the test i-vectors and, therefore, the total number of trials will be 12,582,004. Three baseline systems are considered in this work for evaluation: cosine, PLDA with actual labels, and PLDA with estimated labels. The size of hidden layers is set to 400. Table 1 compares the performance of the proposed DNN systems with other baseline systems in terms of minDCF and EER. The interesting point is that the combination of the DNN-3L and PLDA with estimated labels in the score level improves the results to a great extent. The resulting relative improvement compared to cosine baseline system is 36% in terms of minDCF on the evaluation set. This improvement with no use of background labels is considerable compared to 45% relative improvement which can be obtained by PLDA with actual labels.

Table 2: Performance comparison of proposed GMM-RBM vectors and conventional i-vectors on the **evaluation** set core test condition-common 5 of NIST SRE 2010. GMM-RBM vectors and i-vectors are of a same size of 400.

	cosine		PLDA	
	EER (%)	minDCF	EER (%)	minDCF
[1] i-Vector	6.270	0.05450	4.096	0.04993
[2] GMM-RBM Vector (Trained with ReLU)	6.638	0.06228	4.517	0.05085
[3] GMM-RBM Vector (Trained with VReLU)	6.497	0.06099	3.907	0.05184
Fusion [1] & [3]	**5.791**	**0.05238**	**3.814**	**0.04673**

Table 3: Comparison of LID systems for short signals recorded in car. Performance values are reported based on LER (%).

Duration of Test Signals (in sec) **Number of Samples**	$t < 2$ 2,472	$2 \leqslant t < 3$ 2,355	$t \geqslant 3$ 5,591	All 10,418
[1] GMM-UBM	**9.98**	4.56	4.70	6.02
[2] i-Vector + Cosine	17.28	6.58	5.00	8.09
[3] i-Vector + WCCN + Cosine	14.50	5.03	3.42	6.31
[4] i-Vector + LDA + Cosine	12.41	3.96	2.32	5.03
[5] i-Vector + WCCN + DNN	12.06	3.30	**2.30**	4.60
[6] i-Vector + DNN	11.01	**2.87**	2.58	**4.54**
Fusion [6] & [4]	11.63	3.41	**1.95**	4.48
Fusion [6] & [1]	10.20	3.04	2.49	4.41
Fusion [6] & [4] & [1]	11.12	3.37	**1.96**	**4.39**

For the experiments in Section 3, the NIST 2010 SRE [18], core test-common condition 5, is used for evaluation. Table 2 compares the performance of GMM-RBM vectors, which are obtained with URBMs trained with ReLU and VReLU, with traditional i-vectors on the evaluation set. The use of proposed VReLU shows better performance than the use of ReLU in both cosine and PLDA scoring. At the end, the best results are achieved with score fusion of i-vectors and GMM-RBM vectors which shows about 7-7.5% and 4-6.5% relative improvements in terms of EER and minDCF, respectively, compared to i-vectors. For score fusion, BOSARIS toolkit [19] is used.

For the experiments of Section 4, the database has been recorded within the scope of the EU project SpeechDat-Car (LE4-8334) [20]. Table 3 summarizes the results for all the techniques in four categories based on the test signal durations: less than 2 sec, between 2 and 3 sec, more than 3 sec, and all durations. The first two categories are more interesting because the decision should be made fast in this application. Both i-vector+DNN systems show superior performance compared to i-vector + LDA baseline system. The frame-based GMM-UBM baseline system works better than other systems only for test signals shorter than 2 sec. However, the accuracy is still high in comparison to other categories.

6. Publications

1. O. Ghahabi and J. Hernando, Restricted Boltzmann machines for vector representation of speech in speaker recognition, *Computer Speech & Language*, vol. 47, pp. 16-29, 2018.
2. O. Ghahabi and J. Hernando, Deep Learning Backend for Single and Multisession i-Vector Speaker Recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 807-817, Apr. 2017, **(Awarded the best RTTH doctorate student article in 2017)**.
3. O. Ghahabi, A. Bonafonte, J. Hernando, and A. Moreno, Deep neural networks for i-vector language identification of short utterances in cars, *in Proc. INTERSPEECH*, 2016, pp. 367-371.
4. O. Ghahabi and J. Hernando, Restricted boltzmann machine supervectors for speaker recognition, *in Proc. ICASSP*, 2015, pp. 4804-4808.
5. O. Ghahabi and J. Hernando, Deep belief networks for i-vector based speaker recognition, *in Proc. ICASSP*, May 2014, pp. 1700-1704, **(Awarded the Qualcomm travel grant)**.
6. O. Ghahabi and J. Hernando, i-vector modeling with deep belief networks for multi-session speaker recognition, *in Proc. Odyssey*, 2014, pp. 305-310.
7. O. Ghahabi and J. Hernando, Global impostor selection for DBNs in multi-session i-vector speaker recognition, *in Advances in Speech and Language Technologies for Iberian Languages, ser. Lecture Notes in Artificial Intelligence*. Springer, Nov. 2014.
8. P. Safari, O. Ghahabi, and J. Hernando, From features to speaker vectors by means of restricted boltzmann machine adaptation, *in Proc. Odyssey*, 2016, pp. 366-371.
9. P. Safari, O. Ghahabi, and J. Hernando, Speaker recognition by means of restricted boltzmann machine adaptation, *in Proc. URSI*, 2016, pp. 1-4.
10. P. Safari, O. Ghahabi, and J. Hernando, Feature classification by means of deep belief networks for speaker recognition, *in Proc. EUSIPCO*, 2015, pp. 2162-2166.
11. G. Raboshchuk, C. Nadeu, O. Ghahabi, S. Solvez, B. M. Mahamud, A. Veciana, and S. Hervas, On the acoustic environment of a neonatal intensive care unit: Initial description, and detection of equipment alarms, *in Proc. INTERSPEECH*, 2014, pp. 2543-2547, **(Awarded the ISCA travel grant)**.

7. Conclusions

The main contributions of this thesis have been presented in three main works. In the first one, a hybrid architecture based on DBN and DNN has been proposed to discriminatively model each target speaker for i-vector speaker verification. It was shown that the proposed hybrid system fills approximately 46% of the performance gap between the cosine and the oracle PLDA scoring systems in terms of minDCF. In the second work, a new vector representation of speech has been presented for text-independent speaker recognition. Gaussian Mixture Model (GMM) supervectors have been transformed by a Universal RBM (URBM) to lower dimensional vectors, referred to as GMM-RBM vectors. The experimental results show that the performance of GMM-RBM vectors is comparable with that of traditional i-vectors but with much less computational load. In the third work, a DNN architecture has been proposed for i-vector LID of short utterances recorded in cars. It has been shown that for test signals with duration 2-3 sec the proposed DNN architecture outperforms GMM-UBM and i-vector/LDA baseline systems by 37% and 28%, respectively.

8. References

[1] A. Mohamed, D. Yu, and L. Deng, "Investigation of full-sequence training of deep belief networks for speech recognition," in *Proc. Interspeech*, 2010, pp. 2846–2849.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[3] A. Senior, H. Sak, and I. Shafran, "Context Dependent Phone Models For LSTM RNN Acoustic Modelling," in *Proc. ICASSP*, 2015, pp. 4585–4589.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[5] Y. Lei, N. Scheffer, L. Ferre, and M. Mclaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. ICASSP*, 2014, pp. 1714–1718.

[6] F. Richardson, D. Reynolds, and N. Dehak, "Deep Neural Network Approaches to Speaker and Language Recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[7] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, Oct. 2015.

[8] T. Pekhovsky, S. Novoselov, A. Sholokhov, and O. Kudashev, "On autoencoders in the i-vector space for speaker recognition," pp. 217–224, 2016.

[9] J. Villalba, N. Brümmer, and N. Dehak, "Tied variational autoencoder backends for i-vector speaker recognition," in *Proc. Interspeech*, 2017, pp. 1004–1008.

[10] E. Variani, X. Lei, E. McDermott, I. Lopez Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, 2014, pp. 4052–4056.

[11] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" *Proc. Interspeech*, pp. 1497–1501, 2017.

[12] G. Bhattacharya, J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," *Proc. Interspeech*, pp. 1517–1521, 2017.

[13] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Proc. Interspeech*, pp. 999–1003, 2017.

[14] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.

[15] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey)*, 2010.

[16] O. Ghahabi, "Deep learning for i-vector speaker and language recognition," PhD dissertation, Universitat Politecnica de Catalunya, 2018, [Online]. Available: http://hdl.handle.net/2117/118780.

[17] NIST. (2014) The NIST speaker recognition i-vector machine learning challenge. [Online]. Available: http://nist.gov/itl/iad/mig/upload/sre-ivectorchallenge_2013-11-18_r0.pdf.

[18] NIST, "The NIST year 2010 speaker recognition evaluation plan," 2010, [Online]. Available: https://www.nist.gov/itl/iad/mig/speaker_recognition_evaluation_2010.

[19] N. Brummer and E. Villiers, "BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," 2011, [Online]. Available: https://sites.google.com/site/bosaristoolkit/.

[20] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "Speechdat-car. a large speech database for automotive environments," in *Proc. LREC*, 2000.