INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Training Speaker Enrollment Models by Network Optimization

Victoria Mingote, Antonio Miguel, Alfonso Ortega, Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{vmingote,amiguel,ortega,lleida}@unizar.es

Abstract

In this paper, we present a new approach for the enrollment process in a deep neural network (DNN) system which learns the speaker model by an optimization process. Most Speaker Verification (SV) systems extract representations for both the enrollment and test utterances called embeddings, and then, these systems usually apply a similarity metric or complex back-ends to carry out the verification process. Unlike previous works, we propose to take advantage of the knowledge acquired by a DNN to model the speakers from the training set since the last layer of the DNN can be seen as an embedding dictionary which represents train speakers. Thus, after the initial training phase, we introduce a new learnable vector for each enrollment speaker. Furthermore, to lead this training process, we employ a loss function more appropriate for verification, the approximated Detection Cost Function ($aDCF$) loss function. The new strategy to produce enrollment models for each target speaker was tested on the RSR-Part II database for text-dependent speaker verification, where the proposed approach outperforms the reference system based on directly averaging of the embeddings extracted from the enroll data using the network and the application of cosine similarity.

Index Terms: Speaker Verification, Enrollment Models, Embedding Dictionary, aDCF Loss

1. Introduction

Speaker Verification (SV) is a binary problem which consists of determining whether two different utterances belong to the same identity or different. These two utterances are widely known as enrollment utterance and test utterance. Mostly, current SV systems are trained to multi-class classification to obtain a representation for each of these utterances, which is called embedding or x-vector [1, 2]. However, this approach does not take into account the goal of the verification task to train discriminative embeddings. Therefore, after extracting the speaker embeddings, a back-end is applied to obtain the final verification scores. This back-end can be a simple cosine similarity [3] where the verification scores when each target speaker has more than one enrollment utterance are obtained by averaging all the enrollment embeddings, or an approach more sophisticated [4] to improve the discriminative ability which usually involves a more complex training process and high computational time. To alleviate these drawbacks, this paper presents a novel and straightforward approach to perform the verification process. This approach is based on training enrollment models for each speaker taking advantage of the information learned in the network from the training speaker set which allows the system to be more robust at the test stage. This is possible since the matrix from the last layer in DNN models can be interpreted as an embedding dictionary where the speaker identities of the training data are stored. As we will show, it is a more effective way to make the verification process since each enrollment model is trained to improve the discrimination ability and therefore, the system performance.

Ideally SV systems based on DNN should be trained to carry out directly the verification process, and also, all the parameters should be trained at the same time. For example, training an end-to-end system as a binary classification, so the system is able to determine between two utterances as a target or a non-target trial. This kind of systems have been trained successfully for text-dependent [5, 6] and text-independent [7] tasks, but it can be possible thanks to the availability of a large amount of training data. It also works in cases with strong pre-trained models as started point, e.g. in [8] where a DNN architecture is initialized to mimic a pre-trained i-vector and PLDA, and it is trained using a binary cross-entropy loss function as optimization metric . However, many SV systems in the state-of-the-art have proposed a similar framework which consists of the use of a model trained for multi-class classification with a Cross-Entropy (CE) loss function combined with an average pooling mechanism to produce embeddings. Once the embeddings are extracted, a back-end technique is employed to perform the verification process. Cosine similarity [3] and Probabilistic Linear Discriminant Analysis (PLDA) [9, 10] have been widely employed.

In pioneer DNN works with this framework, it was supposed, that successful classification models would be able to achieve great results in different test data. However, the test data can have different variability from train data, so it may not be always possible to generalize properly in unseen data. For that reason, recently, the CE loss function has been substituted by different variants of classification loss functions such as Angular loss (A-Softmax) [4] or Additive Angular loss (Arc-Softmax) [11] to increase the discrimination ability. On the other hand, different back-end approaches based on metric learning techniques are increasing as a relevant focus of research since these approaches allow to make the training process more appropriate to the evaluation procedure such as our previous work based on triplet neural networks [12, 13] combined with an approximation of the optimization of the AUC (aAUC) [14], contrastive loss [15], partial AUC loss (pAUC) [16], NeuralPLDA[17], or a binary DNN back-end [18]. However, these approaches are very sensitive to the training data selection to create the pairs or triplets, and this process also involves a slow convergence and a high computational cost.

Previously in [19], we addressed these issues proposing the aDCF loss function which is another alternative to CE loss function, and at the same time, it is a more suitable for SV task since this function is inspired by the Detection Cost Function (DCF) [20] which is one of the main verification metrics employed.

In this paper, instead of using a complex back-end, we propose an alternative to the verification process, which consists of training enrollment models. We can develop this approach easily thanks to the use of a learnable vector for each enrollment speaker that will be optimized using the speaker enrollment data

http://dx.doi.org/10.21437/Interspeech.2020-2325

and the network that has been trained and contains the information of the training speakers. Using this approach, we can consider the information stored in the last layer as competing speakers, therefore negative examples, and the enrollment data represent positive examples. This process has to be carried out for each enrollment speaker to produce a learned vector, which will be separated in terms of the detection metric from the training speakers. The process is extremely efficient since there is no need to select hard negatives, and only some learnable parameters are optimized while the rest of the network is frozen. To train the whole system, we optimize our aDCF loss function, which is more appropriate for the verification task. Furthermore, this function can be easily used in the new verification process to train the enrollment models since it is composed of multiple binary classifiers to produce the one-versus-all classification. Therefore, we can reproduce this expression as a binary classifier for the objective function of the enrollment training. Preliminary results outperform a cosine similarity metric and also show an improvement in the system calibration.

The rest of the paper is organized as follows. Section 2 provides a review of our loss function. In Section 3 we present the description of the new approach to train a enrollment model. The entire system description is detailed in Section 4, with the experiment protocol in Section 5. Results and analysis are given in Section 6. Conclusions are presented in Section 7.

2. aDCF Loss Function

In [19], we developed a loss function to replace the traditional CE loss, but keeping the philosophy of the multi-class training as Fig.1a) depicts. Furthermore, this loss function is a differentiable version of the Detection Cost Function (DCF) widely used in verification. Thus, with this approach, we minimize a weighted sum of the probability of misses ($\hat{P}_{miss}$) and the probability of false alarms ($\hat{P}_{fa}$) given by the costs that quantify the trade-offs between both types of errors. For m examples, we estimate $\hat{P}_{miss}$ by averaging the number of times the scores of target speakers N_{tar} are smaller than the decision threshold Ω. On the other hand, the $\hat{P}_{fa}$ is estimated by the average number of times the scores of non-targets N_{non} are greater than Ω. Therefore, we define the probability of false alarm and probability of miss by means of a sigmoid function of the difference between the score and the threshold to make a differentiable approximation of the binary counter, which enables the back-propagation of the gradients as follows,

$$\hat{P}_{fa}(\theta,\Omega) = \frac{\sum_{y_i \in y_{non}} \sigma(\alpha(s_\theta(x_i,y_i) - \Omega))}{N_{non}}, \quad (1)$$

$$\hat{P}_{miss}(\theta,\Omega) = \frac{\sum_{y_i \in y_{tar}} \sigma(\alpha(\Omega - s_\theta(x_i,y_i)))}{N_{tar}}, \quad (2)$$

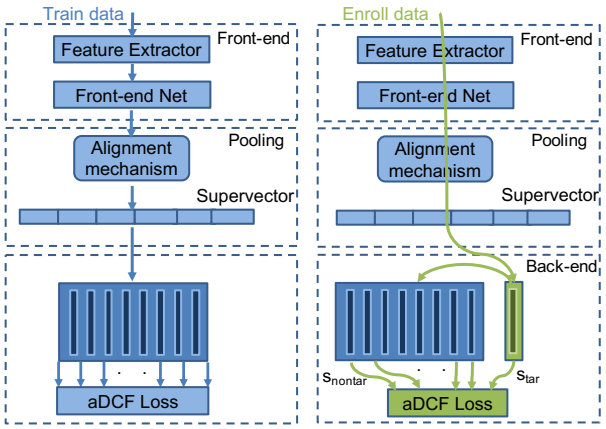
where $\sigma()$ is the sigmoid function, α is an adjustable parameter, and $s_\theta(x_i,y_i)$ is the score obtained from the last layer which is defined as a cosine distance layer as follows,

$$s_\theta(x_i,y_i) = \frac{x_i \cdot W_{y_i}^T}{\|x_i\| \cdot \|W_{y_i}^T\|}, \quad (3)$$

where x_i is the input sample, y_i is the class label, $\|x_i\|$ is the normalized input to the last layer, and $\|W_{y_i}^T\|$ is the normalized layer parameters of the speaker class y_i. Thus, using these expressions, we can now propose to minimize the following approximated loss function defined by,

$$aDCF(\theta,\Omega) = \gamma \cdot \hat{P}_{fa}(\theta,\Omega) + \beta \cdot \hat{P}_{miss}(\theta,\Omega), \quad (4)$$

where γ and β are tuneable parameters to provide more cost relevance to one of the terms over the other.



(a) Training phase (b) Enrollment phase

Figure 1: *(a) Left: Training phase, where the last layer can be seen as an embedding dictionary of the training speakers. (b) Right: Enrollment phase, where an enrollment model is trained for each target speaker.*

3. Training Enrollment Model

This paper proposes a novel approach to carry out the enrollment process in a SV system taking advantage of the information modelled during the training phase with the aDCF loss function to improve the system performance. Nowadays, some SV systems employ complex and strong back-end methods to perform the verification process. Most of them usually require a careful selection process of the input data which makes these methods very sensitive to this process.

To address the issue of data selection, we employ the matrix from the last layer of the architecture combined with the enrollment data to mimic the target/non-target process which is carried out in the verification task. In Fig.2, we interpret the matrix obtained from the last layer during the training process as an embedding dictionary, since each row learns as the training progresses a representation of the speaker information that is correctly classified when the embedding is multiplied by the corresponding row. Therefore, we can see each row weight as an embedding which represents a different speaker.

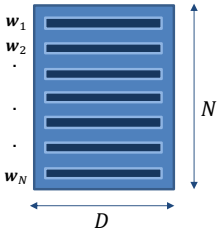


Figure 2: *Embedding dictionary from the last layer in the training phase, where each row represents one of the N train speakers and D is the dimension of the embedding.*

Once this matrix is well-trained in the training stage, we pass to the second phase, which is the enrollment phase represented in Fig.1b). In this phase, we add a new learnable vector w for each enrollment speaker that will also be evaluated similarly to (3) as we will see later. To initialize this layer, we have employed two different alternatives. First, we define random values for vector w, while in the other option, we initialize this vector using an averaging of the enrollment data of each speaker.

Moreover, during the training of the enrollment model, the same aDCF loss function that we have employed in the training phase is optimized. To optimize this function now, the score can be expressed as,

$$s_\theta(x_i) = \frac{x_i \cdot w^T}{\|x_i\| \cdot \|w^T\|}, \tag{5}$$

where $\|x_i\|$ is the normalized input to the enrollment model, and $\left\|w^T\right\|$ is the normalized layer parameters of the embedding obtained from the enrollment utterance. Using this expression, we obtain the scores of the enrollment utterances or targets s_{tar} and the scores of the embedding dictionary or non-targets s_{non}, which are directly used to optimize the aDCF loss. The optimization using this function is possible since it can be seen as a loss designed to make a one-versus-all multi-class classification with binary classifiers. Thus, we optimize the cost of classifying the enrollment utterances as the correct enrollment speaker avoiding the similarity to the stored models from the training set.

For the test phase, the test data is compared with each enrollment model trained during the enrollment phase, and we obtain directly the verification scores without the need of using another external metric. We also have to note that this procedure does not affect the efficiency or the computation cost at test time.

4. Supervector Neural Network System

In the following section, we briefly present the architecture of the system, which is depicted in Fig.1 for the front-end and pooling parts. The front-end in state-of-the-art SV systems is usually based on a DNN with a global average reduction mechanism to extract embeddings. However, this approach does not work efficiently for text-dependent tasks [21] since with the averaging the order of phonetic information in the utterance is dismissed. To address this problem, in previous works [14, 22, 23], we introduced an alignment method as a new layer into the Convolution Neural Network (CNN) architecture employed to replace the average pooling mechanism. This mechanism allows us to keep and encode the temporal structure of the phrase and the speaker information in a supervector. The alignment method employed in this work is a Gaussian Mixture Model (GMM) combined with a Maximum A Posteriori (MAP) adaptation [24].

5. Experimental Setup

5.1. Data

The experiments have been reported on the RSR2015 text-dependent speaker verification dataset [25]. This dataset comprises recordings from 157 males and 143 females. There are 9 sessions for each speaker pronouncing 30 different phrases. Moreover, this data is divided into three speaker subsets: background (bkg), development (dev) and evaluation (eval). In this work, we develop our experiments with Part II which is based on 30 short control commands which have strong overlap of lexical content, and we employ the bkg (97 speakers, 47 female/50 male) for training and dev data for calibration. The evaluation part is used for enrollment training and trial evaluation. This dataset has three evaluation conditions, but in this work, we have only evaluated the most challenging which is the Impostor-Correct case where the non-target speakers pronounce the same phrase as the target speakers. This condition is also the most employed in the text-dependent SV.

5.2. Experimental Description

To develop our experiments, we have employed a 20 dimensional Mel-Frequency Cepstral Coefficients (MFCC) stacked with their first and second derivatives as input to train the alignment mechanism and as input to the DNN. Furthermore, a 64 component GMM has been trained per phrase using the bkg partition. From these models, the alignment information is extracted to use in the alignment mechanism of our architecture.

In this work, a set of experiments was carried out to show the behaviour of the new approach proposed. The performance obtained using the architecture after the training phase to extract embeddings and applying a cosine similarity ($Cosine$) is compared to the results achieved with the training enrollment models approach ($EnrollModel$) proposed with two different alternatives for the layer initialization. The first alternative consists of a totally random initialization ($rand$), and the other alternative is initialized with an averaging of the enrollment embeddings (avg).

6. Results and Analysis

Table 1 presents Equal Error Rate (EER), NIST 2010 minimum and actual detection cost ($minDCF$ and $actDCF$) [26], and Cost of log-likelihood-ratio values ($Cllr$ and $minCllr$)[27].

Back-end			**Fem**	
Type	**Init**	**EER**	**min/actDCF**	**minCllr/Cllr**
Baseline (Cosine)	–	4.19	0.72/0.78	0.159/0.164
Enroll Model	rand	3.77	0.74/0.77	0.143/0.147
Enroll Model	avg	**3.52**	**0.69/0.72**	**0.132/0.135**
Improvement (%)		**15.99**	**4.17/7.69**	**16.98/17.68**

(a) Female results

Back-end			**Male**	
Type	**Init**	**EER**	**min/actDCF**	**minCllr/Cllr**
Baseline (Cosine)	–	5.80	0.91/1.02	0.218/0.228
Enroll Model	rand	5.42	0.89/0.92	0.204/**0.213**
Enroll Model	avg	**5.22**	**0.86/0.89**	**0.196**/0.228
Improvement %		**10.00**	**5.49/12.75**	**10.09/6.58**

(b) Male results

Back-end			**Fem+Male**	
Type	**Init**	**EER**	**min/actDCF**	**minCllr/Cllr**
Baseline (Cosine)	–	5.10	0.85/0.93	0.193/0.201
Enroll Model	rand	4.72	0.83/0.85	0.180/0.185
Enroll Model	avg	**4.46**	**0.79/0.82**	**0.170/0.174**
Improvement (%)		**12.55**	**7.06/11.83**	**11.92/13.43**

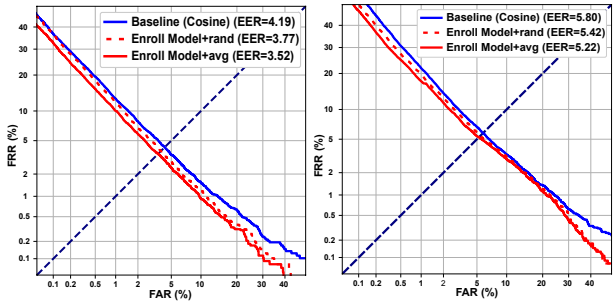
(c) Female+Male results

Table 1: *Experimental results on RSR2015 Part II [25] eval set, showing EER%,* Cllr, minCllr, actDCF *and* minDCF. *These results were obtained to compare the approach proposed with the two alternatives as initialization and the cosine baseline.*

We can observe that the proposed approach for the verification process with the two different initializations outperforms the baseline using a cosine similarity directly over the embeddings extracted from the architecture without applying any other back-end technique. Furthermore, we can see as a good initialization leads the enrollment training to better performance, but a random initialization also improves the baseline results. This fact reflects that training specific enrollment models for each enroll speaker helps to improve the discrimination ability, and therefore the text-dependent speaker verification process.

Moreover, whether we pay the attention in the difference between the values of the optimal DCF ($minDCF$) and Cllr ($minCllr$) with their correspondent actual value, we note that both alternatives for the training of the enrollment models have a minor difference between those values than using the cosine.

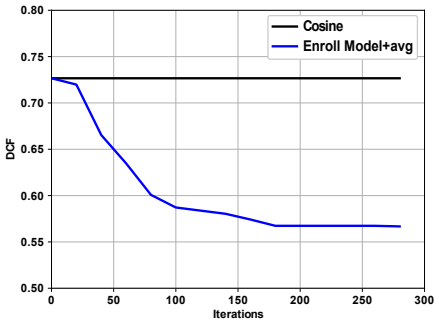
In addition, Fig. 3 represents the Detection Error Trade-off (DET) curves. These curves are grouped by gender to show better the results for each part of the Table 1. These representations clearly demonstrate that the training of the enrollment models with both initializations have a great performance in both subsets (female and male). In Fig.3a), we can see the DET curves of female results where the three results follow the same trend than in the other figures. However, note that the minDCF result in Table 1 for the $EnrollModel + rand$ in this data shows a slightly lower performance than $Cosine$ result, but in the correspondent DET curve, we observe that the overall performance including EER point are also better than $Cosine$ baseline.



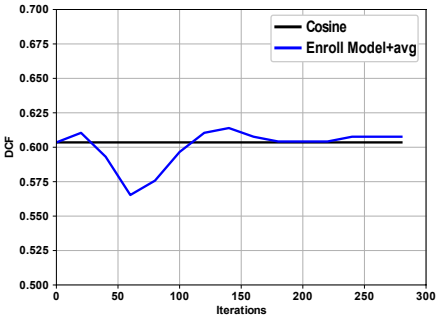
(a) Female DET curves (b) Male DET curves

Figure 3: *(a) DET curves for female results of the three back-ends. (b) DET curves for male results of the three back-ends.*

Finally, we have conducted an analysis of this new approach where we have made a brief study of the results obtained individually for each phrase. In Fig.4, we represent the evolution of the DCF metric for two different phrases during the training of the enrollment phase. Note that for each point in the curve the full trial list is evaluated and DCF computed for the selected phrase. In this representation, we observe two different behaviours which demonstrate that even though the global performance is better with the proposed approach, there is still room for improvement. We can find some phrases that the system works correctly with while some others do not behave as good as expected. For example, Fig.4a) shows one of the phrases which has a great performance and where we can see that the results with the proposed method improves considerably the final DCF result. While in Fig.4b), we observe that the use of the same training configuration for all the phrases may not be the best option. In this case, with less iterations we can find a better result, but the system does not converge to a better solution compared to the baseline.



(a) DCF Evolution Phrase 046



(b) DCF Evolution Phrase 054

Figure 4: *(a) DCF evolution in one of the phrases from the evaluation data which individually has a great performance during the training of the enrollment model. (b) DCF evolution in one of the phrases from the evaluation data which has one of the worst performance during the training of the enrollment model.*

7. Conclusions

In this paper, we have presented a novel approach to perform the verification process. This approach consists of the use of the embedding dictionary stored during the training phase in the matrix of the last DNN layer to train an enrollment model for each speaker. With this system, we mimic the test process where enrollment utterances are compared with test utterances to determine whether each pair of utterances is a target or non-target trial. Even though this is a preliminary study, the proposal has been able to improve the system performance, although, in the analysis part, we have checked that some limitations still exist. The results confirm that this technique is an interesting line of research, so we plan to work in different alternatives to initialize the weight vector combined with different Bayesian estimation approaches.

8. Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the project TIN2017-85854-C4-1-R, by the Government of Aragon (Reference Group T36_20R) and co-financed with Feder 2014-2020 “Building Europe from Aragon”, and by Nuance Communications, Inc. The Titan V used for this research was donated by the NVIDIA Corporation.

9. References

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

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[3] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Asian conference on computer vision*. Springer, 2010, pp. 709–720.

[4] Y. Li, F. Gao, Z. Ou, and J. Sun, "Angular Softmax Loss for End-to-end Speaker Verification," *arXiv preprint arXiv:1806.03464*, 2018.

[5] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, no. Section 3, pp. 5115–5119, 2016.

[6] K. Zhang, Z. Zhang, Z. Li, S. Member, Y. Qiao, and S. Member, "Joint Face Detection and Alignment using Multi - task Cascaded Convolutional Networks," *Spl*, no. 1, pp. 1–5, 2016.

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

[8] J. Rohdin, A. Silnova, M. Diez, O. Plchot, P. Matějka, L. Burget, and O. Glembek, "End-to-end DNN based text-independent speaker recognition for long and short utterances," *Computer Speech & Language*, vol. 59, pp. 22–35, 2020.

[9] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[10] N. Dehak, P. J. 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, 2011.

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

[12] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[13] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1633–1644, 2018.

[14] V. Mingote, A. Miguel, A. Ortega, and E. Lleida, "Optimization of the area under the roc curve using neural network supervectors for text-dependent speaker verification," *Computer Speech & Language*, vol. 63, p. 101078, 2020.

[15] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[16] Z. Bai, X. Zhang, and J. Chen, "Partial AUC Optimization Based Deep Speaker Embeddings with Class-Center Learning for Text-Independent Speaker Verification," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6819–6823.

[17] S. Ramoji, P. Krishnan, and S. Ganapathy, "NPLDA: A Deep Neural PLDA Model for Speaker Verification," *arXiv preprint arXiv:2002.03562*, 2020.

[18] J. Jung, H. Heo, J. Kim, H. Shim, and H. Yu, "RawNet: Advanced End-to-End Deep Neural Network Using Raw Waveforms for Text-Independent Speaker Verification," *Proc. Interspeech 2019*, pp. 1268–1272, 2019.

[19] V. Mingote, A. Miguel, D. Ribas, A. Ortega, and E. Lleida, "Optimization of False Acceptance/Rejection Rates and Decision Threshold for End-to-End Text-Dependent Speaker Verification Systems," *Proc. Interspeech 2019*, pp. 2903–2907, 2019.

[20] A. Martin and M. Przybocki, "The NIST 1999 speaker recognition evaluation—An overview," *Digital signal processing*, vol. 10, no. 1-3, pp. 1–18, 2000.

[21] E. Malykh, S. Novoselov, and O. Kudashev, "On residual CNN in text-dependent speaker verification task," in *International Conference on Speech and Computer*. Springer, 2017, pp. 593–601.

[22] V. Mingote, A. Miguel, A. Ortega, and E. Lleida, "Supervector Extraction for Encoding Speaker and Phrase Information with Neural Networks for Text-Dependent Speaker Verification," *Applied Sciences*, vol. 9, no. 16, p. 3295, 2019.

[23] V. Mingote, A. Miguel, D. Ribas, A. Ortega, and E. Lleida, "Knowledge distillation and random erasing data augmentation for text-dependent speaker verification," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6824–6828.

[24] D. A. Reynolds, R. C. Rose *et al.*, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[25] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2014.03.001

[26] "The NIST Year 2010 Speaker Recognition Evaluation Plan," 2010. [Online]. Available: https://www.nist.gov/sites/default/files/documents/itl/iad/mig/NIST_SRE10_evalplan-r6.pdf

[27] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.