INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Log-Likelihood-Ratio Cost Function as Objective Loss for Speaker Verification Systems

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

Many recent studies in Speaker Verification (SV) have been focused on the design of the most appropriate training loss function, which plays an important role to improve the recognition ability of the systems. However, the verification loss functions created often do not take into account the performance measures which are used for the final system evaluation. For this reason, this paper presents an alternative approach to optimize the parameters of a neural network using a loss function based on the log-likelihood-ratio cost function (CLLR). This function is an application-independent metric that measures the cost of soft detection decisions over all the operating points. Thus, prior or relevance cost parameters assumptions are not employed to obtain it. Moreover, this metric has a differentiable expression, so no approximation is needed to use it as the objective loss to train a neural network. CLLR function as optimization loss was tested on the RSR2015-Part II database for text-dependent speaker verification, providing competitive results without using score normalization and outperforming other similar loss functions as Cross-Entropy combined with Ring Loss, as well as our previous loss function based on an approximation of the Detection Cost Function (DCF).

Index Terms: Speaker Verification, Loss Functions, Metric Learning, aDCF, Cross-Entropy

1. Introduction

Biometric recognition systems are attracting a lot of research interest since they have become a crucial part of day-to-day life, e.g. face and speaker recognition systems. There are mainly two tasks that can be performed by these systems: identification and verification. In this study, we focus on speaker verification (SV) systems which are trained to provide a reliable binary decision of whether two speech samples belong to the same speaker or not. In order to make reliable decisions, SV systems must obtain proper log-likelihood ratios (LLRs) and compare them against a convenient threshold. The development of a method able to convert the outputs of the system into proper LLRs is usually known as calibration [1]. This process takes into account parameters such as the prior probability of a speech sample to belong to the legitimate speaker and the costs of make wrong decisions to adjust a few parameters to meet the requirements of the application. One of the main metrics used to check whether this step has been correctly done is the Detection Cost Function (DCF). Thus, in [2], we developed an approximation of this function (aDCF) to train the SV system directly with this final metric. However, aDCF has a drawback, since it is an application-dependent metric. Therefore, in this work, we propose the use of an alternative as objective loss to train deep neural networks which is based on another verification metrics, the log-likelihood-ratio cost function (CLLR) [1, 3]. The CLLR function is an application-independent evaluation measure where no assumptions of prior or cost parameters are needed.

State-of-the-art SV systems based on deep neural networks are usually trained without considering their main goal which is making a binary decision. However, in the last years, the design of new loss functions has been widely investigated to find the most suitable loss function to train deep learning systems. These efforts have been focused on two lines of study, on one side, the redesign of the identification loss function, and on the other side, the verification loss functions. The former is composed of Cross-Entropy (CE) loss with softmax output units [4, 5, 6] combined with a complementary loss such as Ring loss (RL) [7] or Center loss [8], and other studies have been focused on its variants such as Angular Softmax loss (A-Softmax) [9] or Additive Angular Margin loss (ArcFace) [10]. While the latter is based on metric learning approaches as triplet neural network [11, 12], contrastive loss [13], partial AUC loss (pAUC) [14] or NeuralPLDA [15].

In previous works [16, 2], we developed different loss functions to address the issue of training the system using one of the final evaluation metrics. The first one is aAUC loss function, which is an approximation of the area under the ROC curve combined with the triplet training philosophy. aAUC allows the optimization of the whole system performance during the training process and improves the final results. However, the use of a triplet strategy involves high computational cost and slows down the training process. While the second approach developed is aDCF loss function which is an approximation inspired by the Detection Cost Function (DCF). This function is based on the measure of the decision errors in verification systems. Unlike aAUC approach, aDCF loss function is trained following the philosophy of the existing multi-class loss function, so it is more efficient than the triplet training strategy. Despite its efficiency, this loss function has a drawback since it needs some prior and cost parameters assumptions to be used as objective loss function. Furthermore, we had to make an approximation of the real DCF metric to train a neural network.

In this paper, we propose a new objective loss function for training neural networks as an alternative to our previous aDCF loss function to substitute the classical identification losses. This function is based on the log-likelihood-ratio cost function (CLLR), which measures the overall quality of the scores for making soft decisions using the expected costs. Moreover, the CLLR function has a differentiable expression, so we do not have to approximate it as in the case of the DCF function. Therefore, training with the CLLR function as objective loss allows the end-to-end system to learn how to minimize the expected costs by obtaining good scores. Preliminary results outperform all the alternative loss function evaluated in text-dependent SV task: CE, CE combined with RL, A-Softmax, and aDCF.

The remainder of this paper is laid out as follows. Section

http://dx.doi.org/10.21437/Interspeech.2021-1085

2 provides a review of existing loss functions. In Section 3, we describe the proposal. The architecture employed to develop the system is explained in Section 4. Section 5 describes the experimental setup. Finally, Section 6 presents and discusses results and Section 7 concludes the paper.

2. Loss Functions

For the success of neural networks training, the chosen loss function has an important role since a suitable loss function can improve the discrimination ability of the SV systems. In this section, we describe some of the most extended state-of-the-art loss function and our previous alternative loss function.

2.1. Cross-Entropy Loss

Based on deep learning approaches, the main loss function has been traditionally the Cross-Entropy (CE) loss [4, 5, 6]. This function has been widely employed to solve the multi-class classification. CE loss can be defined as,

$$L_{CE} = -\frac{1}{m}\sum_{i}^{m} log \frac{\exp\left(W_{y_i}^T \cdot x_i + b_{y_i}\right)}{\sum_{j}^{N} \exp\left(W_j^T \cdot x_i + b_j\right)}, \qquad (1)$$

where x_i is the input sample with $i \in \{1, ..., m\}$ and m is the number of samples, y_i is the class label, W is the weight matrix, b indicates the bias value, W_{y_i} and W_j are the y_i and j column of W with $j \in \{1, ..., N\}$ and N is the total number of classes.

The CE loss improves the posterior probability of the training samples, which is not the best approach to achieve generalization in the representations learned.

2.2. Ring Loss

To improve the generalization in the representations, one alternative is to combine CE loss with a complementary loss such as Ring loss [7]. Using this loss, the system learns how to force the embedding norms to be close to the unit circle, which increases the generalization in the representation since all the vectors have similar module and different angles are used to represent the data. The Ring loss is formulated as,

$$L_R = \frac{\lambda}{2m}\sum_{i}^{m} (||x_i||_2 - R), \qquad (2)$$

where R is the target norm value, usually 1, λ is the loss weight, x_i is the input sample of the penultimate layer with $i \in \{1, ..., m\}$ and m is the number of samples.

2.3. Angular Softmax Loss

Many recent studies have pointed out the need to define new CE variants to address the lack of feature discrimination. Thus, another interesting approach to solve the generalization problems was to introduce a redesign of the CE loss. This loss function is known as Angular Softmax, or A-Softmax loss [9] which introduces an angular margin to learn angular discriminative embeddings.

$$L_{ANG} = -\tfrac{1}{m}\textstyle\sum_{i}^{m} log \frac{\exp\left(||x_i||\psi(\theta_{y_i,i})\right)}{\exp\left(||x_i||\psi(\theta_{y_i,i})\right) + \sum_{j \neq y_i}^{N} \exp\left(||x_i|| cos(\theta_{y_i,j})\right)}, \qquad (3)$$

where $\psi(\theta_{y_i,i})$ is the angle function which is a monotonic function defined as,

$$\psi(\theta_{y_i,i}) = (-1)^k cos(m\theta_{y_i,i}) - 2k, \qquad (4)$$

with $\theta_{y_i,i} \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}]$, $k \in [0, m-1]$, and m is an integer to control the angular margin.

2.4. aDCF Loss

Previous approaches have proved to be effective improving some recognition tasks. However, as we showed in [16] with aAUC, metrics related to the final verification measures increase the system performance. For this reason, we developed an approximation of the Detection Cost Function (*aDCF*) [2], which is inspired by DCF [17] and was chosen since it is one of the main metrics in the evaluation process for SV tasks. In addition, the implementation of this loss function allows us to keep the efficiency and speed of the multi-class training by interpreting the outputs of the final linear layer as scores of a verification task. With this approach, we seek to minimize the false alarm probability and the miss probability following the approximated loss function defined by,

$$aDCF(\theta, \Omega) = \gamma \cdot \hat{P}_{fa}(\theta, \Omega) + \beta \cdot \hat{P}_{miss}(\theta, \Omega), \qquad (5)$$

where γ and β are adjustable parameters to provide more cost relevance to one of the terms over the other, and $\hat{P}_{fa}$ and $\hat{P}_{miss}$ can be written as,

$$\hat{P}_{fa}(\theta, \Omega) = \frac{\sum_{y_i \in y_{non}} \sigma_\alpha(s_\theta(x_i, y_i) - \Omega)}{N_{non}}, \qquad (6)$$

$$\hat{P}_{miss}(\theta, \Omega) = \frac{\sum_{y_i \in y_{tar}} \sigma_\alpha(\Omega - s_\theta(x_i, y_i))}{N_{tar}}, \qquad (7)$$

where $s_\theta(x_i, y_i)$ is the score obtained from the last layer of the neural network, Ω is the decision threshold, N_{tar} is the number of target speakers, N_{non} are the non-target speakers, $\sigma_\alpha()$ is the sigmoid function and is defined as,

$$\sigma_\alpha(s) = \frac{1}{1 + exp(-\alpha \cdot s)}, \qquad (8)$$

where α is an adjustable parameter.

With aDCF, we developed a loss function related to the final measurement used in speaker verification. However, this loss function is designed to address the minimization of the decision errors as a function of a single threshold and the cost relevance parameters since it is an application-dependent metric. Thus, using aDCF, the system is trained to meet the requirements of a specific application.

3. CLLR Function

Motivated by the fact that aDCF is a good choice to train the verification systems, but it has the drawback of the fixed operating point, this paper presents a loss function for SV systems based on CLLR [1]. This loss function is a generalization of the previous aDCF since CLLR is formulated as an integral over all possible operating points of DCF. Besides, it is also related to another widely employed graph representation which is known as the Detection Error Trade-off (DET). DET curve is a representation of what happens whether the decision threshold is swept across its whole range, so CLLR can be viewed as a summary of the accuracy obtained over the whole DET curve. The integral of CLLR is defined as,

$$CLLR = \int_{\Omega} DCF(\Omega)\, d\Omega, \qquad (9)$$

where Ω are the overall spectrum of operating points to integrate over them.

This integral expression is not differentiable, so it can not be employed to train a neural network. However, in [1], an analytical closed-form expression was presented to solve this integral, so we do not need to make any approximation as we did with DCF. Using this expression, we can introduce it directly in the neural network as objective loss to optimize. CLLR defined with this differentiable solution measures a sum of the expected log costs of target examples ($Ctar$) and the expected log costs of non-target examples ($Cnon$). $Ctar$ is defined by the sum of the cost for each target example where whether the system assigns it correctly a high score for the target hypothesis, the cost will be low. While $Cnon$ is determined by the sum of the cost for each non-target example where this cost will be low whether the assigned score is low.

$$Ctar(\theta) = \sum_{y_i \in y_{tar}} \log(1 + exp(-s_\theta(x_i, y_i))), \qquad (10)$$

$$Cnon(\theta) = \sum_{y_i \in y_{non}} \log(1 + exp(s_\theta(x_i, y_i))), \qquad (11)$$

where $s_\theta(x_i, y_i)$ is the score obtained from the last layer of the neural network. Furthermore, in the implementation used in our system, we have introduced a τ parameter to divide the scores which is known as temperature scaling parameter [18].

Combining the previous expressions, we propose to minimize as objective loss the CLLR function defined as,

$$CLLR(\theta) = \frac{1}{2\log 2}\left(\frac{Ctar(\theta)}{N_{non}} + \frac{Cnon(\theta)}{N_{tar}}\right), \qquad (12)$$

where N_{tar} is the number of target speakers, and N_{non} are the non-target speakers. Using this expression, we can apply it directly to our system without any other assumption. As in our previous work [2], the optimization is performed iteratively by computing the CLLR for each minibatch and updating the model. Each minibacth is interpreted as a verification evaluation where the outputs of the last linear layer are interpreted as scores. Therefore, using the class labels, we define the target and nontarget scores.

In addition, note that CLLR function can be also interpreted as a measure of loss of information [3] since a CLLR value close to 0 represents that good scores have been obtained. Therefore, this means that these scores store a large amount of information which allows reducing the uncertainty about the speaker hypothesis to accept or reject one speaker. While whether a CLLR value close to 1 is obtained, there is a great loss of information, so the error rate is similar to the reference detector.

4. System Description

In this section, we briefly present the system architecture used in this work for text-dependent SV which is depicted in Fig.1 [19]. The backbone is composed of two Residual Network (RN) blocks with three layers each block. Furthermore, this architecture needs positional information [20] for the self-attention layers to provide good performance. Instead of using temporal positional information as many language modelling applications, we use the output of a phonetic classifier bottleneck [21, 22]. We concatenate this information before each RN block.

For the pooling part, two multi-head attention layers with two memory layers are alternated. Since we use a concatenation of multi-head self-attention layers, it is equivalent to the encoder part of a transformer, which can be seen analogously as an alignment method that allows assigning embeddings to several categories. This approach has been found useful for text-dependent tasks [23, 24]. Furthermore, the use of memory layers has also been proved helpful since these layers are able to store the knowledge obtained for the network during the training process. In addition, with the integration of the phoneme embeddings in the backbone part, the performance of the attention mechanism improves since the phoneme embeddings help to guide to the attention mask.

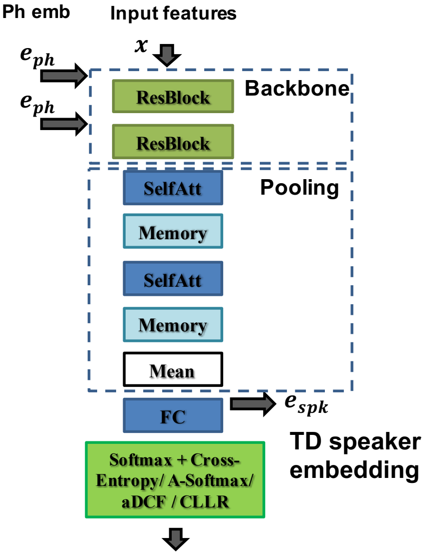

Figure 1: *Architecture for RN, SelfAttention and Memory layers network, composed of a backbone, a pooling and a embedding extraction.*

5. Experimental Setup

5.1. Data

For the experiments, we reported the results on the RSR2015 text-dependent speaker verification database [25]. It consists of speech samples from 157 males and 143 females. For each speaker, there are 9 sessions pronouncing 30 different phrases. The corpus is divided into three speaker subset: background (bkg), development (dev), and evaluation (eval). In this work, we develop our experiments with Part II, which is based on short control commands with a strong overlap of lexical content, and we employ only the bkg data for training. The eval data is used for enrollment and trial evaluation.

5.2. Experimental Description

To develop our experiments, 20 dimensions Mel-Frequency Cepstral Coefficients (MFCC) stacked with their first and second derivates are used as input to train the architecture. Moreover, we have extracted phonetic embeddings of 256 dimensions from a phonetic classifier network which are employed as positional information. Unlike our previous works, for these experiments, we have trained a single deep neural network architecture with all the phrases of the database instead of one model for each phrase. Motivated by the fact that we need to use more data to successfully train an architecture with a self-attention mechanism.

In this paper, a set of experiments was carried out to evaluate the new loss function proposed. We compare the system trained using some of the state-of-the-art loss functions with the proposed CLLR function. Once these systems are trained, we have evaluated them using only a cosine similarity without any score normalization technique or calibration step to show the effectiveness of the scores obtained by training with the different loss functions.

6. Results

Table 1 presents Equal Error Rate (EER), NIST 2008 (*DCF08* [26]) and 2010 minimum detection costs (*DCF10* [27]), and minimum cost of log-likelihood-ratio value ($CLLR$) [3] for the mentioned loss functions. Observing these results, we can conclude that the proposed CLLR function achieves the best results within the three metrics. Thus, we have checked that using this loss function to train the SV system, we have improved the quality of the scores for all the operating points since apart from the improvement of minCLLR, this system achieves the best results in the operating points where EER and DCF are evaluated. Additionally to the previous metrics, in the last two rows of the table, we present the relative improvement achieved comparing the system with CLLR with CE+RL and aDCF systems. These comparatives allow us to remark the fact that training with a loss function oriented to the goal task improves significantly the whole system performance.

Table 1: *Experimental results on RSR2015 Part II [25] eval set, showing EER%, NIST 2008 and 2010 min costs (DCF08, DCF10), and min CLLR. These results were obtained to compare the approach proposed with different loss functions.*

	Female			
Loss Function	**EER**	**DCF08**	**DCF10**	**CLLR**
CE	5.87	0.286	0.740	0.214
CE+RL	4.64	0.228	0.669	0.171
A-Softmax	4.99	0.251	0.703	0.189
aDCF	4.20	0.201	0.660	0.158
CLLR	**3.64**	**0.170**	**0.532**	**0.139**
CLLR vs CE+RL (%)	**21.55**	**25.43**	**20.47**	**18.71**
CLLR vs aDCF (%)	**13.33**	**15.42**	**19.39**	**12.02**

(a) Female results

	Male			
Loss Function	**EER**	**DCF08**	**DCF10**	**CLLR**
CE	6.37	0.304	0.803	0.236
CE+RL	4.92	0.244	0.712	0.184
A-Softmax	6.44	0.309	0.777	0.239
aDCF	4.90	0.232	0.668	0.182
CLLR	**4.07**	**0.200**	**0.588**	**0.157**
CLLR vs CE+RL (%)	**17.27**	**18.03**	**17.41**	**14.67**
CLLR vs aDCF (%)	**16.94**	**13.79**	**11.97**	**13.73**

(b) Male results

	Female+Male			
Loss Function	**EER**	**DCF08**	**DCF10**	**CLLR**
CE	6.22	0.302	0.784	0.229
CE+RL	4.79	0.237	0.706	0.179
A-Softmax	5.79	0.283	0.746	0.217
aDCF	4.64	0.226	0.677	0.175
CLLR	**3.96**	**0.189**	**0.567**	**0.151**
CLLR vs CE+RL (%)	**17.32**	**20.25**	**19.68**	**15.64**
CLLR vs aDCF (%)	**14.65**	**16.37**	**16.25**	**13.71**

(c) Female+Male results

Moreover, note that these performances have been obtained training a single model with all the phrases and without applying score normalization. Therefore, it shows that SV systems trained with one of the final verification metrics produce good scores to achieve promising results without the need for normalization.

In addition to the previous table, Fig.2 depicts DET curves which represent the decision errors sweeping the threshold over all the operating points. These curves show the results for female+male experiments. Note that these representations clearly demonstrate that DET curve obtained with CLLR system shows better results in all the operating points while the DET of aDCF system only outperforms the DET curve of CE+RL system in some points. This fact can be motivated by the assumption of the selected parameters to train with aDCF the system. Furthermore, in this work, we can observe the difficulties to adjust the A-Softmax loss training parameters to achieve good results.

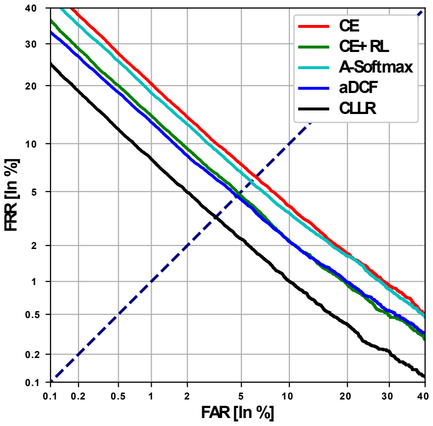

Figure 2: *DET curves for female+male results using the different loss functions evaluated.*

7. Conclusions

In this paper, we have presented a new loss function based on optimizing the quality of the scores over all the operating points. This CLLR function is an alternative to our aDCF loss function to replace the traditional identification loss functions as CE loss without the need of making prior or cost assumptions, and also we have not made any approximation of the real metric. Moreover, the use of CLLR allows the system to learn how to reduce the expected log costs of target and non-target examples. The evaluation was carried out in the text-dependent SV database RSR2015-part II. Results confirm that the SV systems trained with specific verification metrics are a good choice to improve the generalization of the learned representations. This is an interesting line of research where there is still some details to exploit the information that this application-independent loss function can give us.

8. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101007666, the Spanish Ministry of Economy and Competitiveness and the European Social Fund (TIN2017-85854-C4-1-R), by the Government of Aragon (Reference Group T36_20R), and by Nuance Communications, Inc.

9. References

[1] D. A. Van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker classification I*. Springer, 2007, pp. 330–353.

[2] 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.

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

[4] A. Krizhevsky, I. Sulskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information and Processing Systems (NIPS)*, pp. 1–9, 2012.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[6] Y. Srivastava, V. Murali, and S. R. Dubey, "A Performance Comparison of Loss Functions for Deep Face Recognition," *arXiv preprint arXiv:1901.05903*, 2019.

[7] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5089–5097.

[8] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] S. Ramoji, P. Krishnan, and S. Ganapathy, "NPLDA: A Deep Neural PLDA Model for Speaker Verification," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 202–209.

[16] 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.

[17] 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.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[19] V. Mingote, A. Miguel, A. Ortega, and E. Lleida, "Memory Layers with Multi-Head Attention Mechanisms for Text-Dependent Speaker Verification," *Proc. ICASSP 2021*.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[21] I. Viñals, D. Ribas, V. Mingote, J. Llombart, P. Gimeno, A. Miguel, A. Ortega, and E. Lleida, "Phonetically-Aware Embeddings, Wide Residual Networks with Time-Delay Neural Networks and Self Attention Models for the 2018 NIST Speaker Recognition Evaluation," *Proc. Interspeech 2019*, pp. 4310–4314, 2019.

[22] T. Zhou, Y. Zhao, J. Li, Y. Gong, and J. Wu, "CNN with phonetic attention for text-independent speaker verification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 718–725.

[23] 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.

[24] W. Cai, Z. Cai, X. Zhang, X. Wang, and M. Li, "A novel learnable dictionary encoding layer for end-to-end language identification," in *2018 ICASSP*, pp. 5189–5193.

[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.

[26] "The NIST Year 2008 Speaker Recognition Evaluation Plan," 2008. [Online]. Available: https://www.nist.gov/sites/default/files/documents/2017/09/26/sre08_evalplan_release4.pdf

[27] "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