A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

Curricular SincNet: Towards Robust Deep Speaker Recognition by Emphasizing Hard Samples in Latent Space

Labib Chowdhury[1], Mustafa Kamal[2], Najia Hasan[3], Nabeel Mohammed[4]

Abstract: Deep learning models have become an increasingly preferred option for biometric recognition systems, such as speaker recognition. SincNet, a deep neural network architecture, gained popularity in speaker recognition tasks due to its parameterized sinc functions that allow it to work directly on the speech signal. The original SincNet architecture uses the softmax loss, which may not be the most suitable choice for recognition-based tasks. Such loss functions do not impose inter-class margins nor differentiate between easy and hard training samples. Curriculum learning, particularly those leveraging angular margin-based losses, has proven very successful in other biometric applications such as face recognition. The advantage of such a curriculum learning-based techniques is that it will impose inter-class margins as well as taking to account easy and hard samples. In this paper, we propose Curricular SincNet (CL-SincNet), an improved SincNet model where we use a curricular loss function to train the SincNet architecture. The proposed model is evaluated on multiple datasets using intra-dataset and inter-dataset evaluation protocols. In both settings, the model performs competitively with other previously published work. In the case of inter-dataset testing, it achieves the best overall results with a reduction of 4% error rate compare to SincNet and other published work.

Keywords: Biometric Authentication, Speaker Recognition, Angular Margin Loss, Curriculum Learning.

1 Introduction

Speaker Recognition(SR) is widely adopted in real-life scenarios as it has brought remarkable changes in security systems, authentication programs, automated identifications, and forensics. SR is divided into two subtasks: - Speaker Verification (SV) and Speaker Identification (SI). SV involves the comparison of two speech signals and determining whether they belong to the same person. It is simply a validation task where the system is required to indicate whether a speech signal given matches the subject who is being considered. Unlike SV, SI is not a validation task but instead can be considered as a search problem, where given a voice sample of a person, the system attempts to identify the speaker from a list of previously registered speakers.

[1] Department of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh, labib.chowdhury@northsouth.edu
[2] Department of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh, mustafa.kamal@northsouth.edu
[3] Department of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh, najia.tasnim@northsouth.edu
[4] Department of Electrical & Computer Engineering, North South University, Dhaka, Bangladesh, nabeel.mohammed@northsouth.edu

Before the emergence of deep learning in this field, the popular method included the i-vector method [De11], where the features were extracted from MFCC coefficients and Filter-bank Features [Va14], [RD15], [Sn17]. These features are then used in a variety of classifiers, including Probabilistic Linear Discriminant Analysis (PLDA) [PE07] and heavy-tailed PLDA [Ma11]. Numerous recent SR tasks have been based on the popular SincNet [RB18] architecture and as can be appreciated. SR is a very challenging task due to audio signals having a high dimension. Unlike other methods, SincNet can work directly on audio signals because it leverages the parameterized sinc function, which extracts features from audio signals. The deeper layers of the network later process these features.

Biometric systems such as SR and Facial Recognition (FR) can be considered as open-set problems, where the number of classes is not fixed [CM20]. The original SincNet model was trained using the softmax loss [RB18]. Following studies have incorporated various angular margin-based loss functions with SincNet, to achieve better results in both inter-dataset and intra-dataset testing [NZ19], [CM20]. While existing models achieve excellent performance on standard datasets [NZ19], [CM20], the study performed in [CM20] demonstrated that these results do not carry over when performing the inter-dataset evaluation, raising a question about the generalizability of these models. To address this, this study proposes the use of a curriculum learning based loss function and incorporates it with the SincNet architecture. Previously curriculum learning based loss function [Hu20] obtained outstanding performance on biometric tasks such as FR. Influenced by such findings, in this study, we propose Curricular SincNet (CL-SincNet), where we use SincNet architecture as a feature extractor and incorporate curricular loss with it. The contributions of our paper are as follows:

- To the best of our knowledge, we are the first one to introduce curriculum learning applied in the angular space to the speaker recognition task.
- We conducted extensive experiments on two popular speaker recognition datasets, TIMIT and LibriSpeech, and achieve competitive performance on both. In the case of LibriSpeech, we do better than previously published studies. In fact, our approach reduces the frame error rate by 17% in intra-dataset testing.
- Most significantly, we find our proposed approach achieves a lower Classification Error Rate (CER), compared to previously published models in inter-dataset testing. In fact, our proposed approach reduces the CER by 4% when trained on LibriSpeech and tested on TIMIT, thus indicating the better generalizability of our approach.

2 Background Study

This section includes a brief discussion of some background of the softmax loss function and the later part of the discussion includes the SincNet architecture.

Softmax loss is usually defined as the pipeline combination of the last fully connected layer, softmax function, and cross-entropy loss [Wa18]. Softmax loss can be formulated

as:

$$L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{W_k^T f_i + b_k}}{\sum_{c=1}^{C} e^{W_c^T f_i + b_c}} \tag{1}$$

Here, f_i denotes the feature vector from last fully connected layer, W_k represents the kth row of weight matrix W and b_k, b_c are the bias scalar value of respective index value k and c. C is the total number of classes and the number of training samples in a mini-batch is N. By setting bias, $b_k, b_c = 0$ and ensuring W_k^T and f_i are unit norm, equation 1 can be rewritten equation 2

$$L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s\cdot\cos\theta_k}}{e^{s\cdot\cos\theta_k} + \sum_{c=1,c\neq k}^{C} e^{s\cdot\cos\theta_c}} \tag{2}$$

Here s is rescaling parameter and θ_k is angle between weight vector W_k and feature vector f_i. Softmax loss in equations 1and 2 result in a decision boundary between two classes without having any margin being imposed [Wa18]. However for open-set problems, particularly in biometric recognition areas, margin-based loss functions in particular angular margin-based loss functions have obtained superior and encouraging results [De19], [Hu20].

To this end, authors of [De19] proposed arcface loss function that mitigates the issue with softmax loss by imposing a margin in angular space, thus creating more robust and larger decision boundaries between classes. The formulation of [De19] is as follows:

$$L_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s\cos(\theta_{k,i}+m)}}{e^{s\cos(\theta_{k,i}+m)} + \sum_{c=1,c\neq k}^{C} e^{s\cos(\theta_c,i)}} \tag{3}$$

Where authors added an additional margin with the angle between the target weight vector and the feature vector and then rescale the feature by multiplying with s. Although this loss function is verified to obtain good performance [De19] it does not consider each sample's difficulties into consideration [Hu20].

The authors of [Hu20] proposed a new loss function where they leverage curriculum learning and introduce a modulation coefficient in the negative cosine similarity. Authors defined positive cosine similarity as $\cos(\theta_k + m)$, which is same as [De19] but they changed the representation of negative cosine similarity from $\cos\theta_j$ to $N(t, \cos\theta_j)$. The loss is defines as follows:

$$L_{CurricularLoss} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(\cos(\theta_{k,i}+m)}}{e^{s(\cos(\theta_{k,i}+m)} + \sum_{c=1,c\neq k}^{C} e^{sN(t,\cos\theta_c)}} \tag{4}$$

The modulation coefficient function $N(t, \cos\theta_c)$ is defined as in [Hu20]

$$N(t, \cos\theta_j) = \begin{cases} \cos\theta_j, & \cos(\theta_k + m) > \cos\theta_j \\ \cos\theta_j(t + \cos\theta_j), & \cos(\theta_k + m) < \cos\theta_j \end{cases} \tag{5}$$

According to equation 5 a sample is considered to be easy if the angle between the embedding vector and the target weight vector plus the margin is still smaller than the angle between the embedding vector and the weight vector of non-ground truth classes. At the beginning of the training, the hyper-parameter t should be closed to zero so that the model can emphasize the easy samples first, gradually t will increase and the model will focus on the hard example. Since t will increase, the hard sample will be emphasized with larger weights in the later part of the training. the value of t is adaptive in the loss function, as the training goes the estimate of t is formulated as:

$$t^{(k)} = \alpha r^{(k)} + (1-\alpha)t^{(k-1)} \tag{6}$$

Here, $r^{(k)}$ is the average of positive cosine similarity of k-th batch, defined as $r^{(k)} = \sum_i \cos\theta_{y_i}$, α is a momentum parameter, the author from [Hu20] defined $\alpha = 0.99$.

3 Proposed model Curricular SincNet

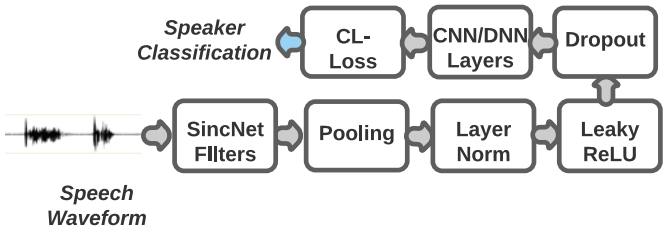

Fig. 1: Portrayal of our proposed architecture

Recent work such as [RB18], [NZ19], [CM20] improved significantly in SI and SV tasks and reports competitive results. Recently [NZ19], [CM20] used the SincNet architecture and incorporated margin-based losses. The initial convolution operation is performed by using parameterized sinc function to extract low-level features from the raw audio waveform and the only learnable parameters of the convolutional filter are the high and low cutoff frequencies. The convolution operation is expressed in equation 7

$$y[n] = x[n] * g[n,\theta] \tag{7}$$

Where $x[n]$ represents a chunk of audio/speaker's signal, $y[n]$ is the filtered output. g is a predefined function and it is defined as equation 8, where a_1 & a_2 represents the low and high cutoff frequencies

$$g[n,a_1,a_2] = 2a_2\frac{\sin(2\pi a_2 n)}{2\pi a_2 n} - 2a_1\frac{\sin(2\pi a_1 n)}{2\pi a_1 n} \tag{8}$$

It has been verified that introducing margin-based losses as a last layer of SincNet, helps to increase the distance between classes and decrease the intra-class distance in the embedding space [CM20], [NZ19]. Although the previously mentioned work showed competitive results, they did not consider each sample's difficulties during the training time.

To address this, in this study we propose Curricular SincNet aka CL-SincNet. We use the same feature extractor as others which is SincNet and we use the curricular loss function as outlined in equation 4 to optimize the network. The motivation is that this approach can allow the model to learn features by allowing the network to learn easy samples first and the harder samples later in the training loop. The graphical representation of our CL-SincNet is shown in figure 1.

4 Experimental Setup

This section describes the details of datasets we use to train and evaluate our model, training and testing procedures, and the metrics we use to measure the model's performance.

4.1 Datasets

We consider TIMIT [Ga93] and LibriSpeech [Pa15] for training and evaluation our model. These two datasets are very popular in SR tasks. TIMIT dataset has 462 classes, or unique speakers, and each class has 8 samples. LibriSpeech has a total 2484 number of classes/unique speakers with a total number of 21933 samples. We used 12-15 seconds of audio for training and 2-6 seconds for testing.

4.2 Training & Testing Procedure

For the training procedure, we use similar settings as [RB18] except for the last layer. We discarded this layer, and instead used the output of the previous layer. We normalized both the feature vector and the row vector of weights by L2 normalization and calculated the cosine similarity of the easy sample and hard sample with corresponding labels. For the two hyperparameters in equation 4 we used the same value as used in [Hu20], which is $m = 0.5$ and $s = 64$. To train the model we use a mini-batch of 128 and the learning rate is set to 10^{-2}. To evaluate our model, we use the same two protocols as [CM20] i.e - Intra dataset test and Inter dataset test. An intra-dataset test is performed to evaluate the Speaker Verification performance and an inter-dataset test is performed to evaluate the Speaker Identification performance. All codes are available at github's project repository[3].

For SV we use Frame Error Rate (FER) and Classification Error Rate (CER) in percentage to demonstrate the performance of our proposed model. These are widely used metrics to measure the performance in SR-based task [RB18], [NZ19], [CM20]. FER is calculated over a window of 200 ms while CER is calculated by averaging the posterior probabilities computed at each frame of the sample and voting for the speaker with the highest average probability. We also use CER(%) in inter-dataset testing for SI task.

The motivation of using the aformentioned metrics is, to demonstrate a fair comparison with recently published works [CM20], [NZ19], [RB18].

[3] `https://github.com/jongli747/Curricular-SincNet`

5 Results

In this section, we discuss our results in two parts. Initially, we speak about speaker verification tasks in the intra-dataset protocol, and then we will talk about speaker identification in an inter-dataset protocol.

5.1 Speaker Verification

	FER(%)		CER(%)	
Configuration	**TIMIT**	**LibriSpeech**	**TIMIT**	**LibriSpeech**
SincNet [RB18]	47.38	45.23	1.08	3.2
AM-SincNet [NZ19]	28.09	44.73	0.36	6.1
AF-SincNet [CM20]	**26.90**	44.65	**0.28**	5.7
Ensemble-SincNet [CM20]	35.98	45.97	0.79	7.2
ALL-SincNet [CM20]	36.08	45.92	0.72	6.4
CL-SincNet (Ours)	37.36	**27.63**	1.08	**0.64**

Tab. 1: Comparison of FER(%) and CER(%) evaluation for both TIMIT and LibriSpeech

As we mentioned earlier in section 4.2 we used FER and CER in percentage as evaluation metrics for SV task(intra- dataset test protocol). Table 1 presents the FER and CER obtained by our model on the TIMIT and LibriSpeech datasets. The performance reported in the previously published models is also shown for comparison. From table 1 we can see that on the LibriSpeech dataset, our proposed model outperforms previously published methods with a significant reduction of FER and CER. In FER, we can see that our proposed model not only outperforms the previously published model, but we have also achieved at least 17% less error rate on the LibriSpeech dataset. Moreover, in terms of CER on the LibriSpeech dataset, our proposed approach has achieved the lowest error rate of 0.64%, reducing the CER by 2.5% in the speaker verification task. Although our model does not show better performance than [NZ19], [CM20] on TIMIT, it is worth mentioning that, TIMIT is a comparatively smaller dataset than LibriSpeech.

5.2 Speaker Identification (Inter-Dataset Evaluation)

For the SI task, we usually compare $x : n$, where x is the given speaker's sample and n is a set of registered lists of speakers. Usually, Cosine Similarity or Euclidean Distance is used for evaluation, this study considered Cosine Similarity.

We have adopted the protocol of inter-dataset testing from [CM20] for the evaluation of the SI task, where the model is trained on one dataset and tested using a different independently collected dataset. This is a good indicator of the generalizability of a model. Table 2 presents the CER obtained by our model on the TIMIT and LibriSpeech datasets. For the sake of simplicity, we refer to the TIMIT trained LibriSpeech tested model as protocol-1 and LibriSpeech trained TIMIT tested model as protocol-2. The first two columns represent the result of protocol-1, and the last two columns represent the results of protocol-2.

Protocol-1		**Protocol-2**	
Configuration	**CER (%)**	**Configuration**	**CER (%)**
SincNet[RB18]	10.09	SincNet[RB18]	10.94
AM-SincNet[NZ19]	9.39	AM-SincNet[NZ19]	13.10
AF-SincNet[CM20]	9.14	AF-SincNet[CM20]	10.83
Ensemble-SincNet[CM20]	8.10	Ensemble-SincNet[CM20]	12.87
ALL-SincNet[CM20]	7.15	ALL-SincNet[CM20]	10.72
CL-SincNet(Ours)	**6.39**	CL-SincNet(Ours)	**6.06**

Tab. 2: Comparison of interdataset evaluation for both TIMIT and LibriSpeech

It is worth mentioning that no samples or classes are overlapped between the two datasets. Our proposed CL-SincNet outperforms the previously published model in both settings. At protocol-1, our proposed model has achieved 0.8% less error rate than compared to previously published work. Most significantly in protocol-2, our proposed CL-SincNet has achieved a 6.06% error rate which is a reduction of more than 4% error rate than compared to previously published works. As we have mentioned earlier, the TIMIT dataset is small than the LibriSpeech dataset, which may be a reason why an improvement in TIMIT is less significant. Tables 1, 2 indicate that our proposed CL-SincNet has the capacity to generalized better than other published models with significant performance improvements.

6 Conclusions

This study has proposed Curricular SincNet (CL-SincNet), where we leverage an angular margin-based curriculum learning loss function on the SincNet architecture for the speaker recognition task. The proposed CL-SincNet has manifested superior results compared to previously published studies [RB18], [NZ19], [CM20]. Our proposed approach reduces the frame error rate by 17% on the LibriSpeech dataset for speaker verification tasks in intra-dataset test protocol and reduces the 4% classification error rate in inter-dataset testing for speaker identification tasks. The results indicate that introducing such a curriculum learning-based loss function on SincNet architecture can have positive outcomes for open-set biometric recognition systems.

References

[CM20] Chowdhury, Labib; Zunair, Hasib; Mohammed, Nabeel: Robust Deep Speaker Recognition: Learning Latent Representation with Joint Angular Margin Loss. Applied Sciences, 10:7522, 2020.

[De11] Dehak, N.; Kenny, P. J.; Dehak, R.; Dumouchel, P.; Ouellet, P.: Front-End Factor Analysis for Speaker Verification. Trans. Audio, Speech and Lang. Proc., 19(4):788?798, May 2011.

[De19] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699, 2019.

[Ga93] Garofolo, John S.; Lamel, Lori; Fisher, William M.; Fiscus, Jonathan G.; Pallett, David S.; Dahlgren, Nancy L.: DARPA TIMIT:: acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1. 1993.

[Hu20] Huang, Yuge; Wang, Yuhan; Tai, Ying; Liu, Xiaoming; Shen, Pengcheng; Li, Shaoxin; Jilin Li, Feiyue Huang: CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. pp. 1–8, 2020.

[Ma11] Matejka, P.; Glembek, O.; Castaldo, F.; Alam, M. J.; Plchot, O.; Kenny, P.; Burget, L.; Cernocky, J.: Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4828–4831, 2011.

[NZ19] Nunes, João Antônio Chagas; Macêdo, David; Zanchettin, Cleber: Additive margin sincnet for speaker recognition. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–5, 2019.

[Pa15] Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S.: Librispeech: An ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210, 2015.

[PE07] Prince, S. J. D.; Elder, J. H.: Probabilistic Linear Discriminant Analysis for Inferences About Identity. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8, 2007.

[RB18] Ravanelli, Mirco; Bengio, Yoshua: Speaker recognition from raw waveform with sincnet. In: 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 1021–1028, 2018.

[RD15] Reynolds, Fred Richardson Douglas A.; Dehak, Najim: A Unified Deep Neural Network for Speaker and Language Recognition. CoRR, abs/1504.00923, 2015.

[Sn17] Snyder, David; Garcia-Romero, Daniel; Povey, Daniel; Khudanpur, Sanjeev: Deep Neural Network Embeddings for Text-Independent Speaker Verification. In: INTERSPEECH. 2017.

[Va14] Variani, E.; Lei, X.; McDermott, E.; Moreno, I. L.; Gonzalez-Dominguez, J.: Deep neural networks for small footprint text-dependent speaker verification. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4052–4056, 2014.

[Wa18] Wang, Feng; Cheng, Jian; Liu, Weiyang; Liu, Haijun: Additive margin softmax for face verification. IEEE Signal Processing Letters, 25(7):926–930, 2018.