Interspeech 2022
18-22 September 2022, Incheon, Korea

Adaptive Rectangle Loss for Speaker Verification

Ruida Li[1], *Shuo Fang*[1], *Chenguang Ma*[1], *Liang Li*[1]

[1]Ant Group

{ruida.lrd,fangshuo.f,chenguang.mcg,double.ll}@antgroup.com

Abstract

From the perspective of pair similarity optimization, speaker verification is expected to satisfy the criterion that each intra-class similarity is higher than the maximal inter-class similarity. However, we find that most softmax-based losses are suboptimal which encourages each sample to have a higher target similarity score only than its corresponding non-target similarity scores but not all the non-target ones. To this end, we propose a batch-wise maximum softmax loss, in which the non-target logits are replaced by the ones derived from the whole batch. To further emphasize the minority hard non-target pairs, an adaptive margin mechanism is introduced at the same time. The proposed loss is named Adaptive Rectangle loss due to its rectangle decision boundary. In addition, an annealing strategy is introduced to improve the stability of the training process and boost the convergence. Experimentally, we demonstrate the superiority of adaptive rectangle loss on speaker verification tasks. Results on VoxCeleb show that our proposed loss outperforms state-of-the-art by 10.11% in EER.

Index Terms: speaker verification, adapative rectangle loss

1. Introduction

Speaker verification(SV) determines whether a pair of speech segments belong to the same identity. A common method for SV tasks is modeling a feature extractor to map the speech segments into discriminative high-level features and then employing a metric function, typically the cosine, to measure the similarity of those features.

There are two main lines of research to model a feature extractor for speaker verification. The first one is to train a multi-class classifier that can separate different identities in the training set by classification losses(e.g. softmax cross-entropy loss[1, 2]). The other one leverages a metric loss function(e.g., triplet loss[3, 4]) to optimize similarity between samples. To boost performance for speaker verification, both softmax-loss based methods and metric-loss based methods seek to maximize intra-class similarity and minimize inter-class similarity.

However, both methods have some drawbacks for verification tasks. For the metric-loss based methods, the required pair/triplet mining procedure is time-consuming and performance-sensitive. For the softmax-loss based methods, the learned features are separable for the close-set classification problem but not discriminative enough for open-set verification tasks. Several softmax-based variants[5, 6, 7, 8, 9, 10, 11] have been proposed to enhance the discriminative power of learned deep features. Most variants introduce a margin penalty on the decision boundary to enforce the intra-class compactness and the inter-class discrepancy. Though the performance is boosted through the margin penalty, the mismatch between close-set training and open-set verification still intrinsically exists. During close-set training, softmax-based losses encourage each individual sample to have a higher target similarity score only than its corresponding non-target similarity scores. This manner only guarantees the separability of identities in the close-set. But for an open-set verification problem, it requires the minimal target similarity should be higher than the maximal non-target similarity. Viewing from the (s_n, s_p) space, where s_n,$s_p$is the non-target and target similarity respectively, the requirement $min(s_p) > max(s_n)$ yields a rectangle area as the green part shown in Figure1(c). However, most softmax-based losses present a larger convergent area than the rectangle. Specifically, given two typical softmax-based losses(AM-Softmax and Circle loss) as shown in Figure1 (a)(b), A and B are on the convergence boundary which determines whether a sample can be classified correctly. The samples satisfying $s_p^i > s_n^i$ in the solid and dashed green part can be well separated from the other classes during close-set training. However, from the viewpoint of verification, the samples in the dashed green part presenting like $(s_n^B > s_p^A)$ causes false acceptance or false rejection when compared with a fixed threshold and finally reduces the performance of speaker verification.

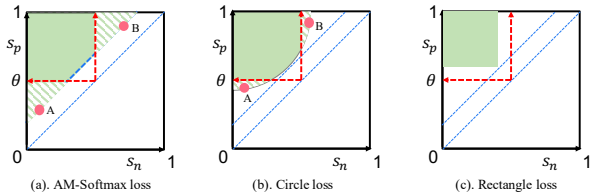

Figure 1: *Comparison of AM-softmax loss, Circle loss, and rectangle loss. s_p, s_n refers to the target and non-target similarity, respectively. θ is the decision threshold. The red border on the top left is the ideal convergent space for verification tasks, where minimal target similarity is higher than the maximal non-target similarity. Rectangle loss contributes to an optimal rectangle decision boundary.*

With these insights, to reach the optimal target, our intuition is that each target similarity score should be higher than all the non-target similarity scores. Motivated by this, we reformulate the softmax loss by replacing the non-target logits with all the ones derived from the whole batch. Then the original soft maximum function turns out to be batch-wise maximal to reach the optimal target for verification. Furthermore, considering the minority hard non-target pairs are easily dominated by the majority easy ones, we design an adaptive margin to mine the hard non-target pairs and emphasize their gradients. Together the proposed novel loss is named Adaptive Rectangle loss due to the rectangle decision boundary as shown in Figure 1(c). And an annealing strategy is introduced during training to improve the stability of the training process and boost the convergence. We summarize the contributions of this work as follows:

- We propose Rectangle loss: a batch-wise maximum loss for deep feature learning. By adopting all the non-

10.21437/Interspeech.2022-486

target similarity scores of the whole batch, Rectangle loss bridges the gap between close-set training and open-set verification.

- We design an adaptive margin mechanism to strengthen the gradient of hard non-target pairs. The mechanism can take full use of the hard pairs to enhance the discriminative power of learned deep features.
- We introduce an anneal training strategy to improve the stability of the training process and boost the convergence. Through a weighted average of the naive softmax loss and the proposed loss, a stable and smooth training procedure can be obtained.

The remainder of this paper is organized as follows. In section2, the Adaptive Rectangle loss is proposed. Section3 and Section4 introduce our experimental setup and results respectively, and we conclude our paper in Section5.

2. Adaptive Rectangle Loss

2.1. Rectangle Loss

The performance of verification models supervised by softmax-based losses suffers from the mismatch between the close-set training and open-set verification. Theoretically, for a speaker verification model, the convergence state is optimal when the minimal intra-class similarity is higher than the maximal inter-class similarity. Specifically, it requires each target similarity score should be higher than all the non-target similarity scores. However, most softmax-based losses are suboptimal which encourage each sample to have a higher target similarity score only than its corresponding non-target similarity scores.

Mathematically, the decision boundary of softmax-based losses can be uniformly expressed as $\alpha_n^i s_n^i - \beta_p^i s_p^i = m$, where s_n^i and s_p^idenote the inter-speaker similarity score and intra-speaker similarity score of i-th sample respectively, α_n^i and $\beta_p^i$are the weight factors and m is the margin. The decision boundary is a triangle or circle as shown in Figure1(a)(b)[10]. Actually, to achieve perfect accuracy, a rectangle decision boundary is needed as shown in Figure1(c). Specifically, given two typical softmax-based losses (AM-Softmax and Circle loss) as shown in Figure1 (a)(b), $s_n^B > s_p^A$ will result in the false acceptance or false rejection, though points above the boundary can be classified to their target classes.

Motivated by this insight, we propose a batch-wise maximum loss with a rectangle decision boundary to separate the target score s_p^i from all non-target scores s_n^j. The novel loss function named rectangle loss is defined as:

$$L_{rectangle} = \frac{1}{N}\sum_{i=1}^{N} log(1 + \frac{1}{N}\sum_{j=1}^{N}\sum_{k\neq y_j}^{C} e^{-\alpha*(s_p^i - s_n^{jk} - m)}) \quad (1)$$

where N is the mini-batch size and C is the number of speakers in the training set, α is the scale factor, $s_n^{jk} = cos(x_j, \omega_k)$ denotes non-target similarity score between the j-th sample x_j and the k-th class centre vector ω_k, s_p^i denotes the target similarity score of i-th sample.

As illustrated in Equation (1), the batch-wise summation symbol $\sum_{j=1}^{N}$is introduced to sum up all the non-target logits $cos(\theta_{j,k})$ traversing the whole batch. The original corresponding non-target logits are replaced by all the ones of the whole batch. More than just classifying samples into target speakers, rectangle loss encourages each sample to have a higher target similarity than all the non-target similarity scores derived from the whole batch. In other words, rectangle loss enforces the minimum intra-class similarity to be higher than the maximum inter-class similarity. Mathematically, the convergent condition of rectangle loss can be expressed as:

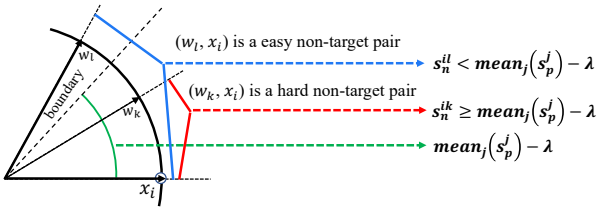

Figure 2: *The schematic diagram of easy and hard non-target pairs. Those that have higher similarity than marginal average target similarity are defined as hard non-target pairs.*

$$min(s_p) - m \geq max(s_n) \quad (2)$$

The convergence space can be expressed as follows, where μ denotes the $max(s_n)$:

$$\begin{cases} s_n \leq \mu \\ s_p \geq m + \mu \end{cases} \quad (3)$$

Geometrically, the formula (3) refers to a rectangle space. As shown in Figure1 (c), the (s_n, s_p) locates in a rectangle area and the minimum of s_p is higher than the maximum of s_n. The rectangle convergent space satisfies the optimal criterion that the minimum intra-class similarity is higher than the inter-class similarity even with a smaller margin. So we conclude that rectangle loss bridges the gap between the close-set training and open-set verification.

Considering that all the non-target pairs are optimized together, the gradient of the minority hard non-target pairs may be covered and further ignored. To address this issue, we design an adaptive margin mechanism to mine the hard non-target pairs and emphasize their less-optimized inter-class similarity score. It will be introduced in the following Section2.2.

2.2. Adaptive Margin mechanism

Pairs are constructed during training when comparing the deep feature x_i with the class centres represented by ω_k. Among the non-target pairs($x_i \notin$ class k), easy ones with small$s_n$are majority which results in the minority hard pairs with high $s_n$are covered and easily ignored. The minority hard pairs are important for boosting performance and should be paid more attention.

Inspired by this, we introduce an adaptive margin mechanism to mine the hard pairs and emphasize their gradients. First, the problem that who are hard non-target pairs should be mathematically addressed. As illustrated in Fig 2, we define the ones as hard non-target pairs which have higher similarity than the marginal average target similarity.

Mathematically, a binary function is defined as follows to identify whether a non-target pair is hard:

$$I(s_n^{ik}) = \begin{cases} 1, & s_n^{ik} - mean_j(s_p^j) + \lambda > 0 \\ 0, & s_n^{ik} - mean_j(s_p^j) + \lambda \leq 0 \end{cases} \quad (4)$$

where λ is a hyperparameter, s_p^jdenotes the target similarty score of the j-th sample x_j, s_n^{ik}refers to the non-target similarty score of the i-th sample x_i and the k-th class centre vector W_k.

$I(s_n^{ik})$ is a binary function to identify whether the i-th sample x_i and the class centre vector W_k is a hard pair.

By applying an inter-class marginal penalty, we design an adaptive margin mechanism to highlight the gradients of the hard non-target pairs. The mechanism defines the new non-target similarty as follows:

$$\widehat{s_n^{ik}} = s_n^{ik} + m * I(s_n^{ik}) - m/2 \tag{5}$$

where m denotes the margin. As illustrated in Equation (5), the adaptive margin mechanism decreases the easy non-target logits and increases the hard non-target logits. Since the gradient of the rectangle loss w.r.t. s_n^{ik} is positively correlated with logits s_n^{ik}, the reduction on easy logits and increase on hard logits have reinforced the gradients of the minority hard non-target pairs.

Finally, combining the rectangle loss and the adaptive margin mechanism, the adaptive rectangle loss is defined as:

$$L_{AR} = \frac{1}{N}\sum_{i=1}^{N} log(1 + \frac{1}{N}\sum_{j=1}^{N}\sum_{k\neq y_j}^{C} e^{-\alpha*(s_p^i - s_n^{jk} - m_{jk})}) \tag{6}$$

$$m_{jk} = m_1 + m_2 * I(s_n^{jk}) - m_2/2 \tag{7}$$

where m_1 and m_2 are the additive margin and the adaptive margin, the $I(s_n^{jk})$ is defined as Equation (4).

2.3. Gradient Analysis

From the viewpoint of gradient, we deeply look into the adaptive rectangle loss and analyze its advantages. The gradients of adaptive rectangle loss w.r.t. s_p^i and s_n^{ik} are derived as follows:

$$\left|\frac{\partial L_{rectangle}}{\partial s_p^i}\right| = \alpha * \frac{1}{1 + \frac{e^{\alpha*(s_p^i - m)}}{Z_0}} \tag{8}$$

and

$$\left|\frac{\partial L_{rectangle}}{\partial s_n^{ik}}\right| = \alpha * e^{\alpha * s_n^{ik}} * Z_1 \tag{9}$$

where Z_0 and Z_1 are defined as:

$$Z_0 = \frac{1}{N}\sum_{j=1}^{N}\sum_{k\neq y_i}^{C} e^{\alpha * s_n^{jk}} \tag{10}$$

$$Z_1 = \frac{1}{N} * \sum_{l=1}^{N} \frac{1}{e^{\alpha*(s_p^l - m)} + \frac{1}{N}\sum_{j=1}^{N}\sum_{k\neq y_i}^{C} e^{\alpha * s_n^{jk}}} \tag{11}$$

Because Z_0 and Z_1 are shared for all samples and have no relationship with index i, $\left|\frac{\partial L_{rectangle}}{\partial s_p^i}\right|$ has negative correlation only with s_p^i and is irrelevant with s_n^i. Likewise, $\left|\frac{\partial L_{rectangle}}{\partial s_n^{ik}}\right|$ has positive correlation only with s_n^i and is irrelevant with s_p^i. In conclusion, the gradient of s_p^ior s_n^ivaries only according to itself but not the common $(s_p^i - s_n^i)$to which the general softmax-based losses according.

Decoupling $s_p^i$and s_n^ihas advantages for verification tasks. We visualize the gradients of AM Softmax loss, Circle loss, and rectangle loss to make a further comparison. More specifically, as shown in Figure2 (b), suppose three samples A, B, C locating d in (s_n^A, s_p^A)=(0.2, 0.4), (s_n^B, s_p^B)=(0.4,0.45), (s_n^C, s_p^C)=(0.6,0.8), respectively. From the perspective of verification, A and C will lead to false acceptance or false rejection because $s_p^A < s_n^C$. Hence, a stronger gradient on s_p^A and s_n^C

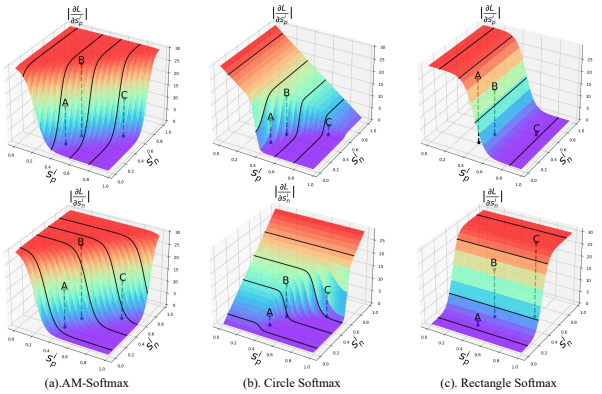

Figure 3: *The gradients of the loss functions. The gradient of rectangle loss w.r.t. s_n and s_p is decouplied with each other, resulting in the lower s_p or higher s_n always have a stronger gradient to be optimized.*

is needed to pull up $s_p^A$and push down the s_n^C. However, as shown in Figure3(a)(b), both AM-Softmax and Circle loss apply a higher gradient on s_n^B and s_p^B, because $(s_p^B - s_n^B)$is too small to separate B from other classes. But as illustrated in Figure3 (c) for rectangle loss, the lower s_p or higher $s_n$always have a stronger gradient to be optimized.

2.4. Annealing Strategy for Training

Adaptive rectangle loss contributes to an optimal rectangle decision boundary that well matches the target of verification tasks. However, adaptive rectangle loss also has more difficulty of network training due to the margin and sharp gradient. To address the problem, we use an annealing strategy to improve the stability of the training process and boost the convergence.

The final loss function is defined as a weighted average of Softmax and adaptive rectangle loss as:

$$L = \lambda * L_{AR} + (1 - \lambda) * L_{softmax} \tag{12}$$

$$\lambda = min(1, (max(iter - S_1, 0))/S_2) \tag{13}$$

where $iter$ is the training step, λ is the weights to balance between $L_{softmax}$ and L_{AR}. and S_1 and S_2 are two integers to control the annealing speed. As shown in Equation (12), at the beginning S_1 steps, the loss function is equivalent to $L_{softmax}$. In the next S_2 steps, loss function convert from $L_{softmax}$ to L_{AR} gradually,contributing to a smooth optimization process.

3. System configuration details

3.1. Dataset

The experiments are conducted on the publicly available datasets VoxCeleb1 and VoxCeleb2[12, 13]. We select the development set of VoxCeleb2 as the training set which includes 5994 speakers and over one million utterances. Besides, the training data is augmented using MUSAN[14] and RIR[15]. The test set is the VoxCeleb1. We employ three available test trials to evaluate the effectiveness of our method: VoxCeleb1-O-clean, VoxCeleb1-E-clean, VoxCeleb-H-clean. In our experiment, All evaluation performance is measured by Equal Error Rate(EER) and the minimum normalized detection cost(minDCF) with $P_{target} = 1e-2$.

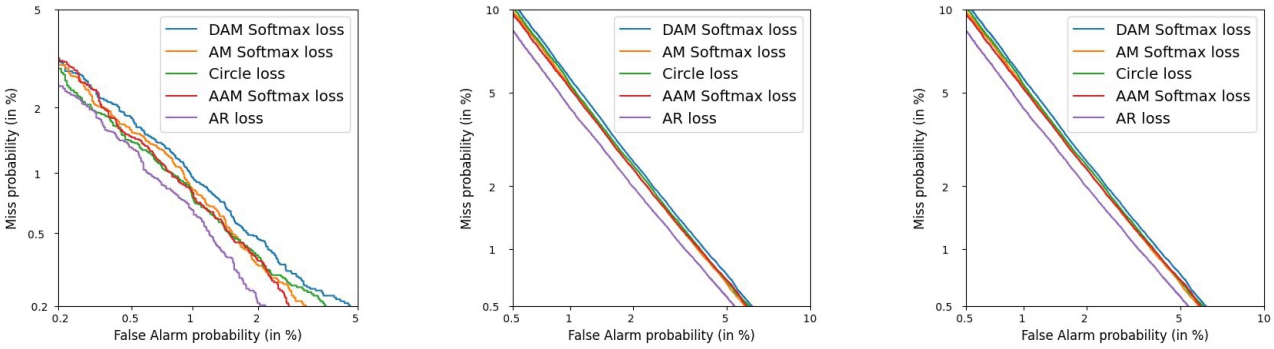

Figure 4: *Det curve on VoxCeleb1-O (a), VoxCeleb1-E (b), VoxCeleb1-H (c), respectively. The results show the adaptive rectangle loss outperforms the other losses from global sight*

3.2. Implementation Details

For data preprocessing, we follow the recent papers[16] to generate 80-dimensional log Melfilterbanks(Fbanks) from 25ms windows with10ms frameshift. The length of variable-length training samples varies from 2s to 4s, corresponding to 200 frames to 400 frames. And we exploit the cepstral mean normalization on the spectrogram. The feature extraction step is handled with Kaldi toolkit[17].

For the embedding network, we employ the widely used ECAPA-TDNN[16] architecture. ECAPA-TDNN consists of a 1-dimensional Squeeze-Excitation Res2Blocks, with a multi-layer feature aggregation to concatenate the information of different hierarchical levels, and a channel- and context-dependent statistics pooling layer is employed to extract statistic features, followed by a fully-connect layer to compute logits. In our experiment, we use 1024 channels in the convolutional frame layers; and set the nodes of the final fully-connected layer as 192, namely the final embedding is192-dimensional.

For the training process, the network is supervised by the proposed adaptive rectangle loss, and the annealing strategy is employed as described in Section2.4.The $m1$, $m2$ of Equation (7) is set to 0.15, 0.1, respectively. All models are trained with Adam optimizer[18] and the batch size is set to 64. The learning rate is started with 1e-3 and continuously decays by 0.95 epoch by epoch. We finish our experiment after 80 epochs.

4. Experimental Result

Results on the VoxCeleb dataset are the most widely used benchmark for speaker verification. In this paper, we conduct trials using DAM-Softmax, AM-Softmax, AAM-Softmax, Circle Loss, and the proposed adaptive rectangle loss to validate the effectiveness of our proposed adaptive rectangle loss.

Table 1 provides the performance of different loss functions on VoxCeleb1-O-clean, VoxCeleb1-E-clean, and VoxCeleb-H-clean. The results demonstrate the adaptive rectangle loss outperforms the series of margin-based losses. More specially, compared with the SOTA system, adaptive rectangle loss improve the system by 10.11%, 8.85%, 7.87% relative reduction in EER and 26.5%, 9.91%, 3.96% relative reduction in minDCF on the VoxCeleb1-O, VoxCeleb1-E, VoxCeleb-H respectively, denoting the superiority of adaptive rectangle loss.

In order to view the performance from a global sight, the DET curves of DAM-Softmax, AM-Softmax, AAM-Softmax, Circle loss, and Rectangle loss are drawn as shown in Fig. 4. For a fair comparison, the same ECAPA-TDNN architecture is adopted for all losses. The DET curves demonstrate the adaptive rectangle loss is superior to typical margin based losses on all three test sets.

Table 1: *The experimental result on voxceleb*

loss	EER(%)	minDCF
Voxceleb1-O		
DAM-Softmax[19]	0.95	0.1376
AM-Softmax[6]	0.91	0.1096
AAM-Softmax[7]	0.89	0.1040
Circle loss[10, 11]	0.89	0.1028
adaptive rectangle loss	**0.80**	**0.0755**
Voxceleb1-E		
DAM-Softmax[19]	1.23	0.1478
AM-Softmax[6]	1.18	0.1416
AAM-Softmax[7]	1.17	0.1392
Circle loss[10, 11]	1.13	0.1392
adaptive rectangle loss	**1.03**	**0.1254**
Voxceleb1-H		
DAM-Softmax[19]	2.26	0.2248
AM-Softmax[6]	2.17	0.2144
AAM-Softmax[7]	2.16	0.2233
Circle loss[10, 11]	2.20	0.2257
adaptive rectangle loss	**1.99**	**0.2059**

5. Conclusions

In this paper, we propose a batch-wise maximum loss called adaptive rectangle loss which bridges the gap between the close-set training and open-set verification. Moreover, we introduce an annealing strategy for a more smooth convergence process. The experiment result shows that our method achieves 10.11%, 8.85%, 7.87% relative reduction in EER compared with the current state-of-the-art systems. We wish that our explorations on learning deep discriminative speaker embeddings will benefit speaker verification tasks in both industry and academia.

6. References

[1] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[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] 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.

[4] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances." in *Interspeech*, 2017, pp. 1487–1491.

[5] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[6] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[8] Z. Huang, S. Wang, and K. Yu, "Angular softmax for short-duration text-independent speaker verification." in *Interspeech*, 2018, pp. 3623–3627.

[9] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," *arXiv preprint arXiv:1904.03479*, 2019.

[10] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6398–6407.

[11] R. Xiao, "Adaptive margin circle loss for speaker verification," *arXiv preprint arXiv:2106.08004*, 2021.

[12] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[13] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[14] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[15] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[16] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] D. Zhou, L. Wang, K. A. Lee, Y. Wu, M. Liu, J. Dang, and J. Wei, "Dynamic margin softmax loss for speaker verification." in *INTERSPEECH*, 2020, pp. 3800–3804.