ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094744

ADAPTIVE LARGE MARGIN FINE-TUNING FOR ROBUST SPEAKER VERIFICATION

Leying Zhang, Zhengyang Chen, Yanmin Qian

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Large margin fine-tuning (LMFT) is an effective strategy to improve the speaker verification system's performance and is widely used in speaker verification challenge systems. Because the large margin in the loss function could make the training task too difficult, people usually use longer training segments to alleviate this problem in LMFT. However, the LMFT model could have a duration mismatch with the real scenario verification, where the verification speech may be very short. In our experiments, we also find that LMFT fails in short duration and other verification scenarios. To solve this problem, we propose the duration-based and similarity-based adaptive large margin fine-tuning (ALMFT) strategy. To verify its effectiveness, we constructed fixed, variable length, and asymmetric verification trials based on VoxCeleb1. Experimental results demonstrate that ALMFT algorithms are very effective and robust, which not only achieve comparable improvement with LMFT in official VoxCeleb evaluation trials but also overcome performance degradation problems in short-duration and asymmetric scenarios respectively.

Index Terms— speaker verification, large margin fine-tuning, duration mismatch, asymmetric scenario

1. INTRODUCTION

In recent years, the development of deep learning made progress in the field of automatic speaker verification (ASV). The neural network based ASV model can be divided into three modules, the frame level speaker feature extraction [1, 2, 3], the pooling layer for statistics extraction [4, 5, 6], and the loss function for optimization [7, 8].

Based on the widely-used softmax function, researchers proposed angular softmax [9, 10] to optimize the speaker embedding in a hyper-sphere space. Further, the margin is added to minimize the within-class distance and maximize the between-class distance. The most popular margin-based loss is additive angular margin softmax (AAM) [11, 12]. In theory, increasing the margin within a reasonable range can make the speaker embedding more discriminative. However, the commonly used training segment for ASV is short, e.g. 2s, and a too-large margin could make the optimization task very challenging. Prior work proposed the large margin fine-tuning (LMFT) strategy [13], which is a secondary training stage for ASV systems and uses longer segments to fit the larger margin. Such strategy achieves great performance improvement on the VoxCeleb [14] dataset and VoxSRC 2020 competition [13].

Yanmin Qian is the corresponding author. This work was supported in part by China NSFC projects under Grants 62122050 and 62071288, in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102, and in part by Shanghai Science and Technology Committee under Grant No. 21511101100.

The commonly used VoxCeleb1 dataset [15] contains utterances of at least 3s, with an average length of 8s, which are the interview video audios collected from YouTube. However, in real verification scenarios, people may say very short phrases for verification and the LMFT model may encounter duration mismatch with such a scenario. Besides, our experiments also verify that the LMFT model fails in short-duration speaker verification.

In most studies, margin is a fixed value. But it can also vary dynamically. For example, Dyn-arcFace [16] implements dynamic margin according to distance between the target class and all other classes, ElasticFace [17] utilizes random margin drawn from a normal distribution in each training iteration, and AdaptiveFace [18] allows margin to vary by class. In addition, DAM-Softmax [19] designs a margin with a negatively correlated cosine similarity of the training sample, inspiring us to rethink the similarity calculated inside the classifier inferred from training samples.

Based on LMFT's drawbacks and inspired by the above works using dynamic margin, we propose the adaptive large margin fine-tuning (ALMFT) strategy from two perspectives, duration-based and similarity-based. These viewpoints reflect the training difficulty of the model for each training sample, and the penalty should be imposed adaptively corresponding to the training difficulty. We carefully design two algorithms to choose the optimal margin for training samples, which require little modification of the initial network.

In experiments, we observe that ALMFT achieves comparable improvement compared with LMFT on the official VoxCeleb1 evaluation set. Meanwhile, ALMFT addresses the performance degradation inherent in LMFT in short-time and asymmetric scenarios and even improves the baseline system in these scenarios.

2. ANALYSIS OF LARGE MARGIN FINE-TUNING

2.1. Angular softmax with margin

The commonly used softmax classification loss is presented as

$$\mathcal{L}_{\text{Softmax}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp^{\mathbf{w}_{y_i}^T\mathbf{x}_i+\mathbf{b}_{y_i}}}{\sum_{j=1}^{C}\exp^{\mathbf{w}_j^T\mathbf{x}_i+\mathbf{b}_j}} \tag{1}$$

where N is the number of training samples, C is the number of speakers in the training set, $\mathbf{x}_i$ is the speaker embedding of the i-th sample, y_i is the corresponding label index, $\mathbf{w}$ is the parameter of the last fully connected layer and $\mathbf{b}$ is the bias.

After normalizing the weights, zeroing the biases and introducing margins, we get the margin-based softmax loss function formulated in Eq.2, where s is a scaling factor, θ_j is the angle between $\mathbf{w}_j$ and $\mathbf{x}_i$, and the $\mathbf{w}_j^T\mathbf{x}+\mathbf{b}$ is rewritten as $s\cdot\cos(\theta_j)$ or $s\cdot f(\theta_j)$.

The angle function Eq.3 summarizes the forms of margin-based softmax loss. In the following sections, we only consider the AAM-softmax case, i.e., the case where $m_1 = 1$ and $m_3 = 0$ [8].

$$\mathcal{L}_{\text{Margin\_S}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp^{s\cdot f(\theta_{y_i})}}{\exp^{s\cdot f(\theta_{y_i})}+\sum_{j=1,j\neq y_i}^{C}\exp^{s\cdot\cos(\theta_j)}} \tag{2}$$

$$f(\theta_{y_i}) = \cos(m_1\theta_{y_i}+m_2)-m_3 \tag{3}$$

2.2. Large Margin fine-tuning

Margin plays a critical role in the ASV task, helping the network to extract more discriminative speaker embeddings [8]. However, simply enlarging the margin value dramatically increases training difficulty and causes performance degradation. Large margin fine-tuning (LMFT) strategy, proposed in VoxSRC 2020 [13] alleviates this problem. For a network already trained to converge, which is also considered as the baseline in our experiments, LMFT is a secondary training phase on top of this initial network to help create more robust speaker embeddings. To stabilize the system at the large margin setting, longer training utterances are applied to provide more speaker information. Meanwhile, the longer duration also matches the VoxCeleb1 testing environment with an average of 8s for utterances [20]. However, LMFT has some inherent drawbacks. Since longer training segment is used in the fine-tuning stage, LMFT is only effective in certain scenarios and is detrimental to duration-mismatched scenarios.

To further investigate this phenomenon, we compare fine-tuned systems with different margins and fine-tuning duration setups in Figure 1. In order to better show the pros and cons of each system, we calculate these systems' relative Equal Error Rate (EER) change compared with the baseline system. Besides, to simulate the evaluation scenarios with different test durations, we sample sub-segments with a specific duration from the original VoxCeleb evaluation set to construct new evaluation trials based on Vox1-E. From the results in Figure 1, we have three key observations.

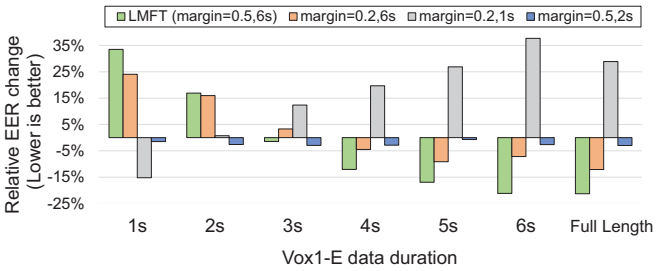

Fig. 1. Relative EER change of fine-tuned systems compared with the baseline system. We sample sub-segments with a specific duration from the official Vox1-E trial for evaluation.

Firstly, We observe that LMFT improves significantly over the pre-trained baseline model on the official VoxCeleb1 evaluation trial but has about 34% degradation on the short-time scenario such as 1s. This duration matching is beneficial for some specific testing data, yet this advantage could not transfer to other scenarios, such as short-time scenarios, and can even lead to performance degradation.

Secondly, we fine-tune the baseline system with utterances of lengths of 6s or 1s under a 0.2 margin, and we notice that training with fixed-length speeches only focuses on certain types of data and cannot get consistent improvement on more complex realistic scenarios. The system with 6s speeches (orange bars) improves significantly in the long-time scenario but causes a performance reduction in the short-time scenario. Similarly, fine-tuning with 1s data (grey bars) makes progress around 1s, but no improvement or even worse performance can be witnessed for test utterances over 3s. As a result, it is necessary to use dynamic and diverse lengths of speech to match more scenarios.

Thirdly, the margin is important, and choosing a reasonable margin also matters. By comparing the orange and green bars, we find that a higher margin helps the model extract stronger speaker embeddings. However, similar to previous studies [8, 12, 13], with 2s training utterances, the blue model is not capable to handle the high penalty by strengthening the margin to 0.5, and the improvement is not obvious. This inspires us to choose a larger margin with a reasonable judgment of the training difficulty it can receive.

In summary, considering the drawbacks of LMFT, we propose a method enabling the system to perceive audio inputs of different durations and to adjust the margin in accordance with data and its corresponding training difficulty during the fine-tuning phase.

3. ADAPTIVE LARGE MARGIN FINE-TUNING

The proposed adaptive large margin fine-tuning (ALMFT) method has two key points. On the one hand, to match more realistic test scenarios, we randomly select multiple lengths of speech for training to enrich the variety of training samples. On the other hand, to reasonably choose a larger margin, we can infer the acceptable training difficulty of the model from the data duration and cosine similarity. We let the margin be adjusted accordingly.

3.1. Duration-based adaptive margin

A larger margin helps to learn discriminative embedding while it increases the training difficulty. Meanwhile, longer training utterances contain more speaker information and are easier to classify, which can fit the larger margin and also adapt to the commonly used long-time evaluation scenario [20, 13]. Therefore, we introduce a dynamic function, allowing the margin to adapt to the training data duration. Here, we use the linear transformation to portray the incremental relation between margin and data duration shown in Eq.4. By setting the range of margin and the range of utterance duration, we can fit the line to get the values of A and B.

$$\text{Margin} = A \times \text{Duration} + B \tag{4}$$

3.2. Similarity-based adaptive margin

Psychological research proved that people should set challenging goals, but it is necessary to consider task complexity to avoid becoming too overwhelming [21]. Neural networks function like an imitation of the man brain [22], therefore, the margin in the loss calculator, serving as the reflection of challenges, should also satisfy the negative correlation with the training difficulty.

In the ASV system, the training difficulty can be reflected in two ways. On the one hand, by using longer speech, more information about the speaker is provided to reduce the difficulty of speaker classification. On the other hand, the arc-cosine function in AAM-softmax calculates the similarity between the current embedding and the target center, which also measures the classification difficulty [19]. A higher similarity indicates the capability of extracting more robust embedding, so we should strengthen the margin to increase the difficulty and vice versa.

To denote the positive correlation between margin and cosine similarity, we use the exponential function shown in Eq.5. α and β can be obtained by fitting this function with the margin range and statistically derived training difficulty. Due to the fast growth rate of the exponential function, the margin becomes extremely high when

cosine similarity is close to 1. To avoid this situation, we set a maximum margin γ to keep the margin in a reasonable range.

$$\text{Margin} = \min(\alpha \exp(\beta \times \text{Similarity}), \gamma) \quad (5)$$

3.3. Adaptive fitted parameter

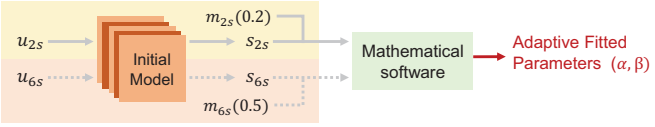

Fig. 2. Adaptive fitting parameter mechanism

Previous works have found the optimal margin for specific training duration, e.g. margin 0.2 for AAM loss with 2s training segment [23] and 0.5 with 6s training segment [13]. We utilize these setups as the benchmark to fit parameters in Eq.4 and Eq.5.

Based on the incremental relation between margin and duration in two optimal setups, we fit Eq.4 and obtain the values of A and B easily. However, to fit α and β in Eq.5, the data duration serves as a bridge to find the correspondence between margin and cosine similarity. As shown in Figure 2, we feed 2s or 6s segments into the initial model and calculate the average cosine similarity between embedding and target classification weight among all the segments. The obtained average similarity will correspond to the same margin as the input segment's duration. We apply mathematical software (such as Matlab) to fit the required function in parameters.

4. EXPERIMENTAL SETUP

4.1. Pre-trained baseline model

For all systems, We use the r-vector [2] as the backbone and train on the development set of the VoxCeleb2 dataset [14]. In the pre-trained baseline system, we apply data augmentation [24] and speed perturbation [25]. During the training process, we train the baseline system by randomly sampling 2s segments from utterances for 165 epochs. The AAM loss [11] is used for system optimization, where the scale ratio and margin equal 32 and 0.2 respectively [8]. The learning rate decreases exponentially from 0.1 to 0.00005. Cosine distance scoring is applied for all experiments.

4.2. Fine-tuning configuration

4.2.1. Large margin fine-tuning configuration

Because fine-tuning is the secondary stage of training, we choose the pre-trained baseline system as the initial model. We use the same configure in [13] to achieve LMFT system. We prolong the training utterances from 2s to 6s and increase the margin of the AAM-softmax from 0.2 to 0.5. We disable the speed perturbation and data augmentation [25]. The learning rate decreases exponentially from 0.0001 to 2.5e-05 during the 10 epochs.

4.2.2. Adaptive margin fine-tuning configuration

For the duration-based and similarity-based adaptive margin fine-tuning (D-ALMFT and S-ALMFT) strategies, the baseline system is used for initiation. Training utterances are randomly chosen from 1s to 6s. To maintain consistency with the parameters of LMFT, the margin range is also from 0.2 to 0.5. We do not utilize data augmentation or speed perturbation. The maximum margin γ in S-ALMFT is set to 0.7 to avoid unreasonable value. Other configurations are the same as those of LMFT described in Section 4.2.1.

4.3. Data preparation

VoxCeleb dataset is a large-scale audio-visual speaker recognition dataset extracted from videos in YouTube [15, 14]. In all experiments, VoxCeleb2-development set [14] is utilized for training. In addition to the official VoxCeleb1, we construct three other datasets based on VoxCeleb1 for evaluation.

(1) Official VoxCeleb1 [15] dataset is an audio-visual large-scale dataset, containing at least 3s and average 8s utterances.

(2) Fixed-length VoxCeleb1 dataset includes six subsets. Each contains fixed duration utterances from 1s to 6s respectively to analyze system performances in different speech duration scenarios.

(3) Variable-length VoxCeleb1 dataset is designed to verify the model's ability for extracting embedding that is robust to the speech duration. We let the duration of enroll and test utterances be chosen randomly from 1s to 6s

(4) Asymmetric dataset reflects realistic scenarios. We use full-length audios for enrollment but intercept 1s or 2s speeches for the test. Because enrollment is only once, and users tend to be more cooperative to record longer voices. However, convenience is often more important in real-world authentication, which requires the authentication system to respond within a very short time.

5. RESULTS AND ANALYSIS

5.1. System performance on the official VoxCeleb dataset

To study the effect of fine-tuning, we conduct experiments with LMFT and our proposed D-ALMFT and S-ALMFT strategies in Table 1. Cosine distance scoring is applied and equal error rate (EER) is used to evaluate the performance. We observe that both D-ALMFT and S-ALMFT strategies help the baseline system gain about 26.2%, 17.3% and 17.6% improvement in the official VoxCeleb1 O, E, and H trials, which is comparable with the LMFT system and is practical for ASV models in challenges.

5.2. System performance on the fixed-length dataset

VoxCeleb is an audio-visual large-scale dataset, but it is not an ideal evaluation dataset in real ASV scenarios for the following reasons. First, the average speech length of VoxCeleb is 8s, while it is difficult to encounter such long test data in real-life scenarios, such as smart wakeup and identity confirmation. Second, enroll and test speeches are often asymmetric in length. Test utterances are usually shorter due to environment and time limitations. To verify in more real-life scenarios, we regenerate three other datasets based on official VoxCeleb1, described in Section 4.3 .

In order to clearly observe the system performance under different durations, we construct the fixed-length dataset based on VoxCeleb1. The fixed-length dataset shows the system performance under different durations. Unfortunately, as shown in Table 1, the system with LMFT method causes 33% and 16% degradation on 1s and 2s scenarios respectively, although it succeeds in the official VoxCeleb1 dataset by taking advantage of data matching [20].

Similar to LMFT, ALMFT method boosts model performance as the data duration becomes longer. The system benefits from the duration matching and the enhanced intra-speaker compactness after giving a higher margin for longer utterances. Meanwhile, ALMFT avoids serious system degradation in short-time scenarios although with a slight performance decline at 1s. Due to the training data longer than 1s, it inevitably has an impact on the performance at 1s, but other test environments are not influenced.

Table 1. **System performance EER(%) comparison.** *LMFT: Traditional large margin fine-tuning; D-ALMFT: proposed Duration-based adaptive large margin fine-tuning; S-ALMFT: proposed Similarity-based adaptive large margin fine-tuning.*

Model	Official			Fixed-length						Variable-length			Asymmetric		
	O	E	H	1s	2s	3s	4s	5s	6s	O	E	H	O	E	H
Baseline	1.116	1.108	2.09	11.437	3.451	1.972	1.474	1.346	1.435	4.351	4.124	6.769	4.303	4.125	6.775
+LMFT	0.872	0.872	1.645	15.269	4.034	1.943	1.296	1.118	1.131	4.923	4.650	7.054	4.856	4.695	7.019
+D-ALMFT	0.824	0.917	1.723	12.415	3.489	1.816	1.287	1.155	1.176	4.306	4.075	6.524	4.142	3.967	6.340
+S-ALMFT	0.824	0.922	1.726	12.392	3.481	1.817	1.292	1.156	1.171	4.301	4.071	6.493	4.106	3.956	6.321

5.3. System performance on the variable-length dataset

It is not comprehensive to measure the performance of real-life ASV applications using fixed-length dataset. Therefore, we design the variable-length dataset to investigate the robustness of the model for data with different durations. Table 1 demonstrates the limitation of LMFT method in the variable-length scenario with a nearly 12% performance degradation. But both D-ALMFT and S-ALMFT methods overcome this problem and even make slight progress.

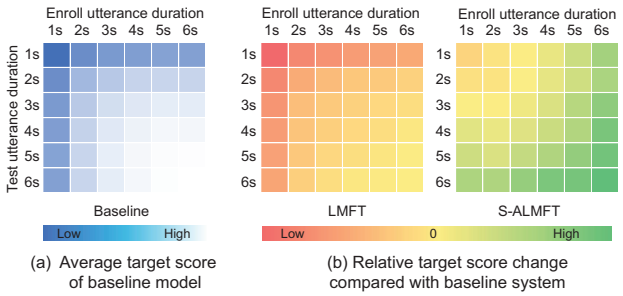

Fig. 3. Average score or relative average score change between embeddings extracted from utterances with the same speaker label and different durations.

If the ASV system is robust for the duration, the embedding extracted from short segments is similar to that extracted from long utterances. Figure 3 (a) compares the average score of speaker embedding extracted from utterances of different lengths based on target pairs (enroll and test utterances are from the same speaker) of variable-length dataset trial E. Darker color means less similarity and a weak ability to extract embedding robust to duration. The score decreases when either enroll utterance or test utterance duration declines, which can be reflected in the diagonal style color change. However, all scores are less than 0.61, indicating that there is still much room for improvement.

Figure 3 (b) presents the relative score change between the baseline system and LMFT or S-ALMFT system. The red upper left corner of the LMFT system represents the worse short-time embedding extraction ability. The green part of the S-ALMFT system shows an improvement in the variable-length dataset. This further indicates that our models address the inherent limitations of LMFT, and obtain performance gains in extracting embeddings robust to duration.

5.4. System performance on the asymmetric dataset

In real life, longer utterances are required in the enrollment process and speakers tend to speak shorter words to do verification for convenience. The information asymmetry provided by enroll and test utterances is challenging for ASV tasks as shown in Section 5.3. To simulate asymmetric scenario, we form the asymmetric dataset for evaluation described in Section 4.3.

Similar to the other datasets, Table 1 shows an unsatisfied performance of LMFT, with a 13.8% reduction in trial E. However, ALMFT methods achieve up to 4.5%, 4.1% and 6.7% performance improvement over baseline system on O,E and H trials respectively.

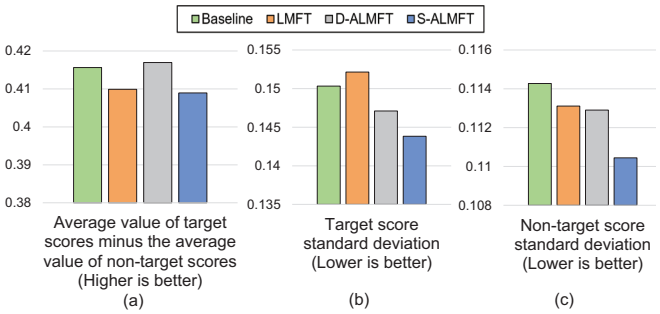

Fig. 4. Target and non-target speaker score distribution of systems on asymmetric dataset trial H.

Figure 4 compares the target pairs and non-target pairs score distribution and their corresponding standard deviation of different models. Figure 4 (a) represents the difference between the average score of target pairs and non-target pairs, which also indicates the inter-speaker distance. We notice that D-ALMFT increases this distance, while LMFT and S-ALMFT lead to a slight reduction. Figure 4 (b) presents the standard deviation of target pairs, indicating the intra-speaker compactness, which is enhanced by both D-ALMFT and S-ALMFT and is decreased by LMFT. Figure 4 (c) compares the standard deviation of non-target pairs and all three algorithms improve the ability to identify non-target speakers. Although S-ALMFT does not further enlarge inter-speaker separability, Table 1 shows that the system still outperforms the baseline system thanks to the reduced intra-speaker variance.

6. CONCLUSION

Focusing on the drawbacks of the conventional LMFT method, we emphasized the necessity of collaborative variation of training utterances and the training penalty. In this paper, we developed the adaptive LMFT methods that adjusted the margin according to utterance duration and similarity for each sample and its class center. In addition to the official VoxCeleb1, we constructed fixed-length, variable-length, and asymmetric datasets based on VoxCeleb1 to better simulate real-life scenarios. Finally, our newly proposed ALMFT gained comparable performance on official VoxCeleb1 compared with conventional LMFT, while without any performance degradation or even with slight improvements on other scenarios.

7. REFERENCES

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev 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.

[2] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[3] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[4] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[5] Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey, "Self-attentive speaker embeddings for text-independent speaker verification.," in *Interspeech*, 2018, vol. 2018, pp. 3573–3577.

[6] Leying Zhang, Zhengyang Chen, and Yanmin Qian, "Enroll-aware attentive statistics pooling for target speaker verification," *Proc. Interspeech 2022*, pp. 311–315, 2022.

[7] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Interspeech*, 2017, vol. 2017, pp. 999–1003.

[8] Yi Liu, Liang He, and Jia Liu, "Large margin softmax loss for speaker verification," *arXiv preprint arXiv:1904.03479*, 2019.

[9] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[10] Zili Huang, Shuai Wang, and Kai Yu, "Angular softmax for short-duration text-independent speaker verification.," in *Interspeech*, 2018, pp. 3623–3627.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos 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.

[12] Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1652–1656.

[13] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5814–5818.

[14] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[15] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[16] Jichao Jiao, Weilun Liu, Yaokai Mo, Jian Jiao, Zhongliang Deng, and Xinping Chen, "Dyn-arcface: dynamic additive angular margin loss for deep face recognition," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 25741–25756, 2021.

[17] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1578–1587.

[18] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z Li, "Adaptiveface: Adaptive margin and sampling for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11947–11956.

[19] Dao Zhou, Longbiao Wang, Kong Aik Lee, Yibo Wu, Meng Liu, Jianwu Dang, and Jianguo Wei, "Dynamic margin softmax loss for speaker verification.," in *INTERSPEECH*, 2020, pp. 3800–3804.

[20] Daniel Garcia-Romero, David Snyder, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, "x-vector dnn refinement with full-length recordings for speaker recognition.," in *Interspeech*, 2019, pp. 1493–1496.

[21] Edwin A Locke, "Toward a theory of task motivation and incentives," *Organizational behavior and human performance*, vol. 3, no. 2, pp. 157–189, 1968.

[22] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, pp. e00938, 2018.

[23] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.

[24] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[25] Miao Zhao, Yufeng Ma, Min Liu, and Minqiang Xu, "The speakin system for voxceleb speaker recognition challange 2021," *arXiv preprint arXiv:2109.01989*, 2021.