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

Angular Margin Centroid Loss for Text-independent Speaker Recognition

Yuheng Wei[1], *Junzhao Du*[1], *Hui Liu*[1]

[1]School of Computer Science and Technology, Xidian University, China

yhwei1993@gmail.com, dujz@xidian.edu.cn, liuhui@xidian.edu.cn

Abstract

Speaker recognition for unseen speakers out of the training dataset relies on the discrimination of speaker embedding. Recent studies use the angular softmax losses with angular margin penalties to enhance the intra-class compactness of speaker embedding, which achieve obvious performance improvement. However, the classification layer encounters the problem of dimension explosion in these losses with the growth of training speakers. In this paper, like the prototype network loss in the few-short learning and the generalized end-to-end loss, we optimize the cosine distances between speaker embeddings and their corresponding centroids rather than the weight vectors in the classification layer. For the intra-class compactness, we impose the additive angular margin to shorten the cosine distance between speaker embeddings belonging to the same speaker. Meanwhile, we also explicitly improve the inter-class separability by enlarging the cosine distance between different speaker centroids. Experiments show that our loss achieves comparable performance with the stat-of-the-art angular margin softmax loss in both verification and identification tasks and markedly reduces the training iterations.

Index Terms: speaker recognition, intra-class compactness, inter-class separability, angular margin, embedding

1. Introduction

Speaker recognition (SR) [1] includes two types of tasks, verification (SV) and identification (SID). Given utterances from an unknown speaker, SV determines whether the speaker matches his/her claimed identity, while SID classifies the speaker as one of enrollment speakers. SR is also classified into text-dependent and text-independent methods, depending on whether the spoken contents are constrained into a specific lexical set.

A typical SR framework consists of two major components: a frond-end modeling speakers and a back-end discriminating them based on a similarity measure. The classical SR paradigm is the combination of i-vector [2] and probabilistic linear discriminant analysis(PLDA) [3], which has dominated the SR field over the past decade. With the rise of deep learning, deep neural network (DNN) as a powerful tool is used for SR and achieves incredible performance. The main usage of DNN is to extract speaker representations (speaker embeddings) from raw utterances, which replace i-vectors to model speakers in form of compact vectors. Deep speaker embedding gradually outperforms i-vector especially when dealing with short utterances, benefiting from the large scale of training data, the exquisite structures of DNN and the well-designed loss functions.

The loss functions used to learn speaker embedding in the DNN training can be roughly branched into two categories: classification and end-to-end. Softmax Loss is the most typical classification loss which trains DNN to predict the speaker labels for training samples. In this case, speaker embeddings such as D-vector [4] and X-vector [5] are derived from the activations of the last hidden layer before the classification layer, which are abundant in speaker characteristics. SR tasks usually face a open-set circumstance in which testing speakers may not appeared in the training dataset. This means the speaker embeddings should be discriminative enough for unseen speakers, but the softmax loss only makes speaker embeddings roughly separable. To address this issue, the end-to-end losses based on deep metric learning are proposed to directly optimize the speaker embedding and the similarity metric at the same time, such as Contrastive Loss [6] and Triplet Loss [7][8]. The popular triplet loss optimizes speaker embeddings so that the distance between the same speaker is smaller than the distance between the different speakers over a threshold. However, the triplet loss has two drawbacks: (1) The number of triplets explosively grows with the training dataset scale. (2) It relies on a complicated triplet mining strategy to search effective triplets.

Recently, several softmax variants are proposed to boost the discriminative power of face representations in the face recognition field, including SphereFace [9], CosFace [10] and ArcFace [11], which are also introduced into the SR field [12][13][14][15]. These methods project the representations as well as the class weight vectors onto a hypersphere by L2-normalization, so that the classification result depends on the angle between the representation and the class weight vector. Furthermore, a margin penalty is imposed into the target logit before softmax normalization to concentrate the speaker embeddings belonging to the same identity together, which encourages the intra-class compactness. However, one major drawback of above softmax variants is that the size of the weight matrix in the classification layer linearly increases with the number of training speakers.

Generalized end-to-end (GE2E) loss [16] is an enhanced constrative loss, which directly optimizes the cosine distance between speaker embeddings and their centroids without a complicated sample selection procedure like the triplet loss or a classification layer like the softmax loss. The alike method is also studied in [17], which adopts the Prototype Network Loss [18] proposed in the few-shot learning field [19] to optimize the Euclidean distance between embeddings and centroids. However, these losses are still not discriminative enough for unseen speakers. In this paper, we introduce an additive angular margin penalty between the speaker embeddings and their centroids to shrink the intra-class angular distance, so that the speaker embeddings belonging to the same speaker tightly gather around their centroid. As analyzed in RegularFace [20], the explicit consideration of the inter-class discrepancy is crucial to the face representation discrimination. Inspired by this, We also push the embedding centroids far away from each other by increasing the cosine distances between centroids. Our contributions are summarized as follows:

1. We propose Angular Margin Centroid (AM-Centroid) Loss to directly optimize the cosine distances between speaker embeddings and their corresponding centroids with an additive angular margin, which learns more dis-

http://dx.doi.org/10.21437/Interspeech.2020-2538

criminative speaker embeddings by explicitly enhancing the intra-class compactness and the inter-class discrepancy simultaneously.

2. We compared the AM-Centroid Loss with the start-of-the-art losses widely-used in the SR filed, such as the additive angular margin softmax loss, the triplet loss and the GE2E loss. Experiment results demonstrate the effectiveness of the proposed loss.

This paper is arranged as follows. In Section 2, we review the angular softmax loss and the GE2E loss ,which are the start-of-the-art losses used in SR methods. And then we illustrate our AM-Centroid loss. In Section 3 and Section 4, we give the experiment details and results. In Section 5, a conclusion of this paper is given.

2. Proposed Approach

Our AM-Centroid loss is inspired by the angular softmax loss with the additive angular margin penalty and the GE2E loss. We firstly revisit them and then introduce the AM-Centroid loss.

2.1. Additive Angular Margin Softmax Loss

Let $x_i \in \mathbb{R}^d$ refer to the d-dimensional speaker embedding belonging to the y_i-th speaker class. W_j denotes the j-th column of the classification matrix $W \in \mathbb{R}^{d\times n}$. The number of speaker classes is n and the mini-batch size is K. The standard softmax loss is defined as follows:

$$L_1 = -\frac{1}{K}\sum_{i=1}^{K} \log \frac{e^{W_{y_i}^{\mathrm{T}} x_i}}{\sum_{j=1}^{n} e^{W_j^{\mathrm{T}} x_i}} \tag{1}$$

The angular softmax loss fixes $\|x_i\| = 1$ and $\|W_j\| = 1$ by L2-normlization. And $\|x_i\|$ is re-scaled to s, so that $W_j^{\mathrm{T}} x_i = s\cos(\theta_{i,j})$ where $\theta_{i,j}$ is the angle between x_i and W_j. This forces θ_{i,y_i} to be smaller than $\theta_{i,j} (j \neq y_i)$. To further enhance the intra-class compactness, an additive angular margin penalty m is used to penalize the large θ_{i,y_i}. The additive angular margin softmax is presented as follows:

$$L_2 = -\frac{1}{K}\sum_{i=1}^{K} \log \frac{e^{s(\cos(\theta_{i,y_i}+m))}}{e^{s(\cos(\theta_{i,y_i}+m))} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_{i,j}}} \tag{2}$$

As illustrated in Figure 1(a), the additive margin penalty compresses the speaker embeddings and the cosine distance between decision boundaries of different speakers is linearly controlled by the hyper-parameter m.

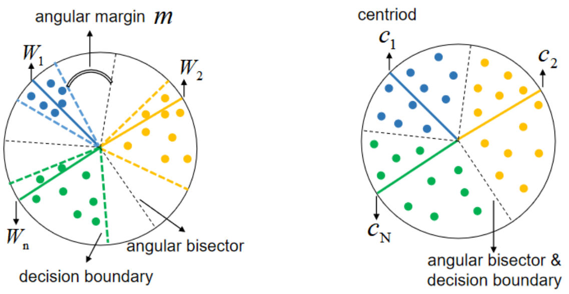
2.2. Generalized End-to-end Loss

Let a mini-batch consist of N speakers and M utterances per speaker. We use $x_{ij} (1 \leq i \leq N, 1 \leq j \leq M)$ to denote the speaker embedding extracted from speaker i utterance j. The embedding centroid of speaker k is defined as follows:

$$c_k = \frac{1}{M}\sum_{j=1}^{M} x_{kj}. \tag{3}$$

The GE2E loss defines the similarity matrix $S_{ij,k}$ to represent the scaled cosine similarity between each speaker embedding x_{ij} and all centroids c_k $(1 \leq i, k \leq N$ and $1 \leq j \leq M)$ as follows:

$$S_{ij,k} = w \cdot \cos(\theta_{x_{ij},c_k}) + b, \tag{4}$$



(a) AM-Softmax Loss (b) GE2E Loss

Figure 1: *Illustration of speaker embeddings learned by different loss functions, circle dots in color represents different identities. Black dotted lines are the angular bisectors between (a) weight vectors (b) centroids.*

where w, b are learnable parameters and θ_{x_{ij},c_k} refers to the angle between x_{ij} and c_k. Then the softmax activation function is used to normalize each row of S consisting of $\{S_{ij,k}\}(k = 1, 2, \cdots, N)$. The GE2E loss wants $softmax(S_{ij,k}) = 1$ when $i = k$, otherwise $softmax(S_{ij,k}) = 0$, so it is defined as follows:

$$L_3 = -\frac{1}{N\times M}\sum_{i,j} \log \frac{e^{S_{ij,i}}}{\sum_{k=1}^{N} e^{S_{ij,k}}}. \tag{5}$$

2.3. Angular Margin Centroid Loss

As shown in Figure 1(b), although GE2E loss encourages the embeddings belonging to speaker k to be closer to their own centroid c_k than other centroids, there still exists a large intra-class distance. Inspired by the concept of margin penalty proposed in the angular softmax loss, we want to penalize the angle between each speaker embedding and its corresponding centroid if the angle is large, so we redefine the similarity matrix $S_{ij,k}$ by replacing w with a scalar value s, setting $b = 0$ and adding a angular penalty m to the target angle as follows:

$$S_{ij,k} = \begin{cases} s \cdot \cos(\theta_{x_{ij},c_k} + m) & \text{if } i = k; \\ s \cdot \cos(\theta_{x_{ij},c_k}) & \text{otherwise.} \end{cases} \tag{6}$$

Note that a training trick to guarantee the stable convergence in [16] is to remove x_{ij} when calculating its corresponding embedding centroid as in Equation (7). We also integrate this into our loss, so the losses for a batch and for a embedding x_{ij} are defined by Equation (8) and Equation (9), respectively.

$$c_k^{(-j)} = \frac{1}{M-1}\sum_{h=1,h\neq j}^{M} x_{kh} \tag{7}$$

$$L_4 = \frac{1}{N\times M}\sum_{i,j} l(x_{ij}) \tag{8}$$

$$l(x_{ij}) = -\log \frac{e^{s\cdot\cos(\theta_{x_{ij},c_i^{(-j)}}+m)}}{e^{s\cdot\cos(\theta_{x_{ij},c_i^{(-j)}}+m)} + \sum_{k=1,k\neq i}^{N} e^{s\cdot\cos(\theta_{x_{ij},c_k})}} \tag{9}$$

The intra-class compactness and the inter-class separability are two key factors which are contributed to the distinguishability of speaker embedding. The additive angluar margin penalty takes effect by mainly influencing the former factor. We further consider the latter factor by enlarging the angle distance between different centroids $\{c_k\}, (k = 1, 2, \cdots, N)$. To do this, we impose a repulsive force on the embedding centroids to push them far away from each other. This repulsive force is formulated by Equation (10).

$$L_5 = \frac{N \times (N-1)}{2} \sum_{k=1}^{N-1} \sum_{g=k+1}^{N} \cos\left(\theta_{c_k, c_g}\right) \qquad (10)$$

The AM-Centroid loss is combined by (8) and (10) as Equation (11),where λ balances the intra-class loss and the inter-class loss. As illustrated in Figure 2, the AM-Centroid loss takes a batch as input, and calculates the embedding centroid for each speaker in the mini-batch. The loss L_4 focus on compressing the speaker embeddings from the same speaker to approach their centroid with the additive angular margin m, while the loss L_5 makes the centroids far apart. This joint effect can result in more discriminative speaker embeddings.

$$L_6 = L_4 + \lambda L_5 \qquad (11)$$

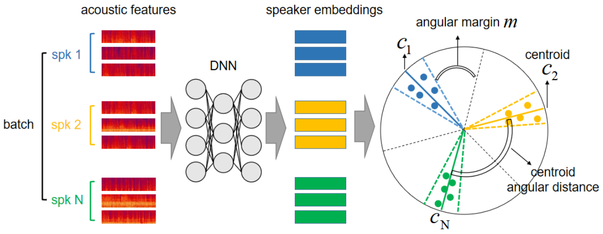


Figure 2: *The details of the AM-Centroid Loss.*

3. Experiments

3.1. Dataset

LibriSpeech [21] is a publicly available speech dataset which encompasses sufficient utterances annotated with speaker labels. We use the *train-clean-360* subset to train speaker models, which consists of 921 speakers, averaged 25 minutes speech per speaker and approximate 360 hours in total. We divide the *train-clean-100* subset containing 251 speakers and about 100 hours speech into two parts: a SID testing set and a SV testing set, with 125 and 126 speakers respectively.

We use fixed-length speech segments extracted from utterances of different speakers to build training mini-batches to ensure comparability of results, and use variable-length utterances for testing, which is a universal practice in most of the SR studies.

3.2. Evaluation Protocol

The evaluation follows the standard protocol: extracting speaker embeddings and calculating their cosine similarity to determine speaker identities. We report the accuracy (ACC) for SID and the equal error rate (EER) for SV.

3.3. Implementation Details

Input Features: We extract 40-dimensional log mel-filterbank energies for each speech frame of width 20 ms and step 10 ms. The training speech segments are set to 2 s and this generates a spectrogram with the size of 198 × 40 for each segment. The mean-and-variance normalization over a 3 s window is performed on each spectrogram.

Neural Network Architecture: The DNN architecture used to extract speaker embedding is constructed based on [22] with several modifications, as depicted in Figure 3. Three time delay neural network (TDNN) layers are stacked to learn short-term temporal context, which have the parameters (*the number of filters*, *the filter size*, *the dilated factor*). A bidirectional LSTM layer with 512 neural units is followed to capture long-term temporal dependence in the frame-level feature sequence. Next two full connected layers change the feature dimension to 512 and 1024 successively. We use a statistic pooling layer to generate the utterance-level feature vector by calculating the mean and standard deviation of input feature sequence and concentrating these statistics together. Final two fully connected layers with 512 and 256 nodes project the utterance-level feature vector into the 256-dimensional speaker embedding. We choose $GELU$ [23] activation function and use batch normalization [24] to accelerate training. Note that another fully connected layer with 921 nodes is attached to the last layer of this structure to act as the classification layer when training DNN with the angular softmax loss.

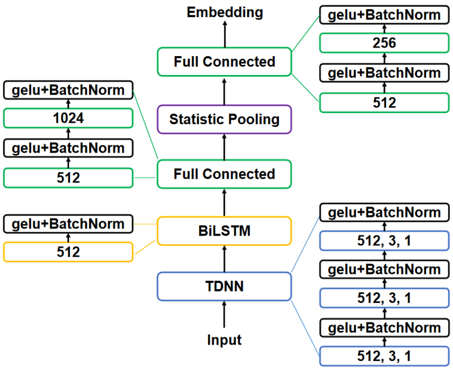


Figure 3: *The details of deep neural network used to extract speaker embedding.*

Training Setting: Experiment codes are implemented by Tensorflow[25]. The Adam algorithm[26] is used to optimized deep neural networks. We use a initial learning rate of $1e-3$ to train the softmax, AM-Softmax ($m = 0.0$) and GE2E losses. The models trained by the AM-Softmax ($m = 0.0$) loss and the GE2E loss are respectively used as pre-trained models to train the AM-Softmax ($m > 0.0$) loss and the AM-Centroid loss, and the learning rate is changed to $1e-4$. The mini-batch size is set to 128 for the softmax loss and the AM-Softmax loss. When training the Triplet loss, GE2E loss and AM-Centroid loss, each batch contains 64 speakers with 10 speech segments per speaker. For the Triplet loss, we use the *batch hard*[27] strategy to generate triplets for training.

4. Results

4.1. Speaker Verification and Speaker Identification

We firstly explore the recognition performance for different losses on both the SID and SV testing sets constructed in Section 3.1. For SID, we randomly picked 30 K utterances as enrollment utterances and selected 10 evaluation utterances (1 positive sample and 9 negative samples) for each enrollment utterance. For SV, we randomly generated 18 K positive pairs and 18 K negative pairs.

In the experiments, we train different losses based on the same DNN described in Section 3.3 for fair comparison. For the GE2E loss, the learable parameters w and b are initialized to 10 and -5 respectively as in [16]. For the triplet loss, we use the cosine distance as the similarity metric and set the margin α to 0.1. For the angular softmax loss and the proposed loss, we set the embedding scale s to 40. The balance factor λ in our loss is set to 0.1. The value of angular margin m significantly influences the recognition performance and the difficulty level of the convergence of the training process, so we tired multiple angular margin settings, including 0.3, 0.4 and 0.5. The values larger than 0.5 make the training process fail to converge well, so we don't list the corresponding evaluation results.

Table 1: *Speaker verification EER (%) and speaker identification accuracy (%).*

Loss	EER	Accuracy
Softmax	10.43	80.51
Triplet Loss ($cosine$, $\alpha = 0.1$)	8.41	82.62
GE2E Loss	8.30	83.74
AM-Softmax ($m = 0.0$)	11.36	79.12
AM-Softmax ($m = 0.3$)	9.85	81.53
AM-Softmax ($m = 0.4$)	8.01	84.61
AM-Softmax ($m = 0.5$)	7.38	85.28
AM-Centroid Loss ($m = 0.3$)	7.72	85.14
AM-Centroid Loss ($m = 0.4$)	6.59	86.37
AM-Centroid Loss ($m = 0.5$)	6.14	86.51

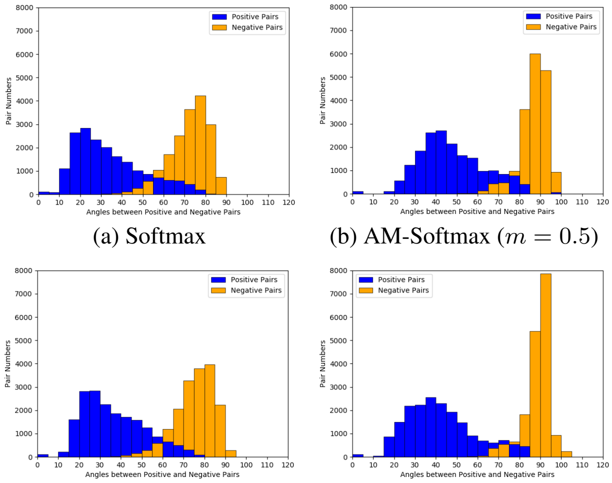
Table 1 shows the evaluation results, where *AM-Softmax* refers to the angular softmax loss with the additive angular margin penalty and *AM-Centroid* denotes the proposed loss. When $m = 0$, the AM-Softmax loss reduces to the angular softmax loss without any angular margin penalty, which achieves the worst performance in our experiment. Compared to the softmax loss and the triplet loss, the GE2E loss shows its superiority which potentially learns how to classify speakers based on the cosine similarity between speaker embeddings. However, the AM-Softmax loss ($m > 0.3$) surpasses the performance of the GE2E loss, which demonstrates the effectiveness of the angular margin penalty. The AM-Centroid loss considers the inter-class discrepancy and the intra-class density explicitly at the same time, which obtains the best performance on both the SV and SID trials when $m = 0.5$.

4.2. Angle Distributions

In order to intuitively understand the discrimination of speaker embedding. We give the detailed angle distributions of both positive pairs and negative pairs on the SV testing set. As in Figure 4, the verification performance is determined by the size of the overlap area between the intra-class and inter-class angles. The angular margin effectively makes them away from each other. The inter-class angles of the AM-Centroid loss trend to be larger than that of the AM-Softmax loss due to explicitly increasing the cosine distance between different speaker centroids during training.

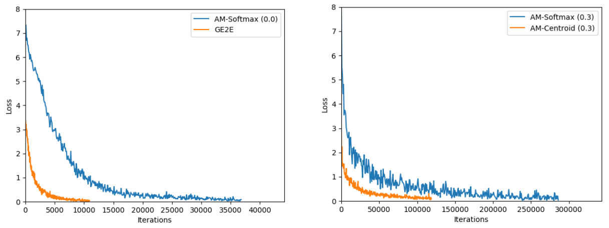
4.3. Training iterations

Figure 5 illustrates the training iterations required by model convergence when using the the AM-Softmax ($m = 0.3$) loss and the AM-Centroids ($m = 0.3$) loss. The AM-Centroids loss makes the training procedure more effective, because it handles a large number of utterances in a mini-batch at once.



(a) Softmax
(b) AM-Softmax ($m = 0.5$)
(c) Ge2E
(d) AM-Centroid ($m = 0.5$)

Figure 4: *Angle distributions of both positive and negative pairs. Orange bars denotes the angles between positive pairs, while blue bars indicates that between negative pairs. The X-axis is the angle presented in degree and the Y-axis is the count of corresponding pairs.*



(a)
(b)

Figure 5: *The number of training iterations when using different losses. (a) Am-Softmax ($m = 0.0$) and GE2E. (b) Am-Softmax ($m = 0.3$) and Am-Centroid ($m = 0.3$)*

5. Conclusions

In this paper, we propose the Angular Margin Centroid Loss (AM-Centroid Loss), which effectively enhances the inter-class separability and the intra-class compactness of speaker embedding simultaneously for text-independent SR when dealing with unseen speakers. We conducted comprehensive experiments to evaluate the performance of the proposed loss. Experiment results demonstrate that our loss outperforms other losses in performance and has the advantage of faster convergence.

6. Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities, the Innovation Fund of Xidian University, National Key Research & Development Program of China (#2018YFB1003605),Open Fund of State Key Laboratory of Computer Architecture (#$CARCH$201704).

7. References

[1] J. P. Campbell, “Speaker recognition: A tutorial,” *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[3] S. J. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[4] 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.

[5] 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.

[6] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[7] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, “Deep speaker: an end-to-end neural speaker embedding system,” *arXiv preprint arXiv:1705.02304*, 2017.

[8] H. Bredin, “Tristounet: triplet loss for speaker turn embedding,” in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 5430–5434.

[9] 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.

[10] 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.

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

[12] S. Yadav and A. Rai, “Learning discriminative features for speaker identification and verification,” in *Proc. Interspeech 2018*, 2018, pp. 2237–2241. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1015

[13] M. Hajibabaei and D. Dai, “Unified hypersphere embedding for speaker recognition,” *arXiv preprint arXiv:1807.08312*, 2018.

[14] S. Novoselov, A. Shulipa, I. Kremnev, A. Kozlov, and V. Shchemelinin, “On deep speaker embeddings for text-independent speaker recognition,” *arXiv preprint arXiv:1804.10080*, 2018.

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

[16] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[17] J. Wang, K. Wang, M. T. Law, F. Rudzicz, and M. Brudno, “Centroid-based deep metric learning for speaker recognition,” pp. 3652–3656, 2019.

[18] J. Snell, K. Swersky, and R. S. Zemel, “Prototypical networks for few-shot learning,” pp. 4077–4087, 2017.

[19] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, “Generalizing from a few examples: A survey on few-shot learning,” *arXiv: Learning*, 2019.

[20] K. Zhao, J. Xu, and M.-M. Cheng, “Regularface: Deep face recognition via exclusive regularization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1136–1144.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

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

[23] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.

[24] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv: Learning*, 2015.

[25] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: A system for large-scale machine learning,” in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

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

[27] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” 2017.