A Study on Angular Based Embedding Learning for Text-independent Speaker Verification

Zhiyong Chen, Zongze Ren and Shugong Xu
Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China
Email: {bicbrv, zongzeren, shugong}@shu.edu.cn

***Abstract*—Learning a good speaker embedding is important for many automatic speaker recognition tasks, including verification, identification and diarization. The embeddings learned by softmax are not discriminative enough for open-set verification tasks. Angular based embedding learning target can achieve such discriminativeness by optimizing angular distance and adding margin penalty. We apply several different popular angular margin embedding learning strategies in this work and explicitly compare their performance on Voxceleb speaker recognition dataset. Observing the fact that encouraging inter-class separability is important when applying angular based embedding learning, we propose an exclusive inter-class regularization as a complement for angular based loss. We verify the effectiveness of these methods for learning a discriminative embedding space on ASV task with several experiments. These methods together, we manage to achieve an impressive result with 16.5% improvement on equal error rate (EER) and 18.2% improvement on minimum detection cost function comparing with baseline softmax systems.**

I. Introduction

Automatic speaker verification (ASV) is the task of determining the ID of a speaker given his/her voice. Lots of works have been focused on this open-set recognition task. In the past several years, the mainstream is to generatively model the speaker information base on manually designed acoustic features such as MFCC [1], [2]. I-vector based system enables us to estimate a low-level utterance level feature given an acoustic feature of a sentence, using factor analysis [3]. Probabilistic linear discriminant analysis (PLDA) [4], [5] models the speakers in the utterance level embedding space, enabling us to scoring the trials with a likelihood ratio.

Recently, the trend in this field is to use deep neural networks as the utterance level embedding extractor, exploiting the non-linearity and powerful representation power of DNN. This trend has lead to a wide range of studies that successfully improve the performance on ASV, such as the well-known x-vector/PLDA system [6], [7]. Moreover, many studies are now focusing on exploring an end-to-end speaker verification system [8]–[11], which use a neural network to model all component in a speaker verification system.

Learning a better embedding is important for all recent DNN based ASV systems no matter what front-end or back-end we use. The object of embedding learning is to find a discriminative feature representation at a specific hidden layer of a neuron network. This requires to learn an embedding space that has good intra-class compactness and inter-class separability. Softmax cross entropy loss has been widely used in recognition and classification task including ASV [6], [7], [12]. Triplet loss is designed to encourage the embeddings in the same class to have smaller Euclidean distance than embeddings from different classes [13], which also have successfully implemented in ASV tasks [14], [15]. However, there also remains some disadvantages to these methods. For softmax, the learned embeddings may be suitable for closed-set classification problems, but not discriminative enough for open-set verification problems. For triplet loss, triplet sampling method turns out to be tricky for effective model training.

In the face recognition field, the recently proposed large margin angular based embedding learning method is effective [16]–[19]. These methods encourage the embeddings to compactly distributed in the unit hypersphere by adding an angular margin between embeddings and their cluster centers. Some of these methods have been exploited for the ASV task [20]–[23] and better results have been observed.

In some more recent works, inter-class regularization is found to be effective when we are using angular based loss [24]–[27]. These studies show that angular based softmax tend to focus mainly on intra-class compactness but not inter-class separability. This makes the embedding centers from different classes not well distributed in the embedding space when the dimension of embeddings are high.

In this work, we give a more detailed study on the angular based embedding learning methods for text-independent ASV task comparing to existing works. We consider our contribution as follows: 1. Apply several different types of large margin angular based loss for ASV task and make an explicit comparative experiment for their performance. 2. Propose an angular based inter-class regularization which leads to improved results consistently, and verify the effect of such method with qualitative and quantitative experiments. We also share our findings of some effective training methods to give a successful implementation of angular based methods on ASV. These methods together lead to impressive improvements comparing with the baseline system.

II. Angular Based Embedding Learning

A. Angular based softmax

The most widely used classification loss function is known as softmax cross entropy, presented as follows:

$$L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathbf{w}_{y_i}^T\mathbf{x}_i+b_{y_i}}}{\sum_{j=1}^{C}e^{\mathbf{w}_j^T\mathbf{x}_i+b_j}}, \quad (1)$$

where $\mathbf{x}_i \in \Re^d$ denotes the last hidden layer output for the i-th sample, which is also known as an embedding. y_i is the label of $\mathbf{x}_i$. The embedding feature dimension d is set to 512 in this paper. $\mathbf{w}_j$ denotes the j-th column of the output linear matrix $\mathbf{W} \in \Re^{d \times C}$, and $b_j \in \Re^C$ is the bias term. N and C are batch size and class number. Following [16], [18] we fix $\mathbf{b}_j = 0$ and transform the logit as $\mathbf{w}_j^T \mathbf{x}_i = \|\mathbf{w}_j\| \|\mathbf{x}_i\| cos(\theta_{j,i})$. We further normalize $\|\mathbf{w}_j\| = 1$ to have modified softmax loss:

$$L_{modified} = -\frac{1}{N} \sum_{i=1} \log \frac{e^{\|\mathbf{x}_i\| cos(\theta_{y_i,i})}}{\sum_{j=1} e^{\|\mathbf{x}_i\| cos(\theta_{j,i})}}, \tag{2}$$

in which $\theta_{j,i} (0 \leqslant \theta_{j,i} \leqslant \pi)$ is the angle between vector $\mathbf{w}_j$ and $\mathbf{x}_i$. Note $\mathbf{w}_j$ can thus be seen as the *cluster centers*. Modified softmax is able to directly optimize angles by making the decision boundary only depends on the angles. We do not further normalize the length of $\mathbf{x}_i$ and give a fix scale constraint in our study, since that is parameter sensitive, as suggested in [26].

The recent studies show adding an angular margin to softmax is effective to learn more discriminative embeddings, and the connection with hypersphere manifold makes the learned features particularly suitable for open-set verification task. These methods can be concluded in the following equation:

$$L_a = -\frac{1}{N} \sum_{i=1} \log \frac{e^{\|\mathbf{x}_i\| (cos(m_1 \theta_{y_i,i} + m_2) - m_3)}}{e^{\|\mathbf{x}_i\| (cos(m_1 \theta_{y_i,i} + m_2) - m_3)} + \sum_{j \neq y_i} e^{\|\mathbf{x}_i\| cos(\theta_{j,i})}}, \tag{3}$$

in which m_1, m_2 and m_3 are hyper-parameters to set angular margins. More specifically, setting m_1 gives angular softmax (A-softmax or SphereFace) [16], m_2 gives additive angular margin softmax (AAM-softmax or ArcFace) [18] and m_3 gives additive margin softmax (AM-softmax) [19].

B. Annealing

Different from the implementation of angular based loss in face recognition, we observe some training difficulties when re-implementing these methods to ASV task. We find annealing is important for the convergence of all angular based loss in our study. For A-softmax system we use the following annealing as in [16]:

$$f_{y_i} = \frac{\lambda \|\mathbf{x}_i\| cos(\theta_{y_i,i}) + \|\mathbf{x}_i\| cos(m\theta_{y_i,i})}{1 + \lambda}, \tag{4}$$

where f_{y_i} is the y_i-th output logit given embedding $\mathbf{x}_i$, and λ need to be gradually reduced during training.

For AM-softmax and AAM-softmax systems, we find annealing is equivalently important to achieve neuron network convergence at the beginning of the training, we use a more straight forward method in these two cases:

$$L_a^{'} = (1 - \lambda^{'}) L_{modified} + \lambda^{'} L_a, \tag{5}$$

where we gradually increase $\lambda^{'}$ in first several epochs to gradually shift the loss from modified softmax to large margin angular based loss.

C. Inter class regularization

Inter-class separability and intra-class compactness are two key factors to learn a discriminative embedding space. However, as suggested in [24], [25], [27], current angular based embedding learning methods effectively encourage better intra-class compactness but the inter-class separability can be degraded due to feature redundancy. This may cause the *cluster centers* not so well distributed in the high-dimension embedding space. Following the effectiveness in face recognition, we thus propose an angular regularization that explicitly encourages the *cluster centers* in $\mathbf{W}$ to be uniformly distributed around the hypersphere. We consider the following separability measurement:

$$SEP_{\mathbf{W}} = \frac{1}{C} \sum_j \sum_{i, i \neq j} max[0, cos(\phi_{i,j})]^2, \tag{6}$$

where $\phi_{i,j}$ is the angle between $\mathbf{w}_i$ and $\mathbf{w}_j$, ideally this scalar should be minimize to zero. This is equivalent to the following criterion:

$$L_{inter} = \frac{1}{C} \|[\mathbf{W}_n^T \mathbf{W}_n]_+ - \mathbf{I}\|_F^2, \tag{7}$$

where l2-normalized columns of $\mathbf{W}$ consist $\mathbf{W}_n$, $[\cdot]_+$ denotes clamping the matrix elements below zero, and squared Frobenius norm can be considered as calculating the energy of the matrix. We thus denote this angular regularization as *hyperspherical energy* following [25]. We use the following loss function to combine the inter-class regularization:

$$L_{a+inter} = (1 - \lambda_{inter}) L_a + \lambda_{inter} L_{inter}, \tag{8}$$

where λ_{inter} is the hyper-parameter.

III. Experiments

A. Experimental Setup

Our experiment is based on the Voxceleb I & II dataset [28] [29]. To enhance the efficiency of experiments, we randomly sample 100 examples for each of the 5994 speakers in Voxceleb II development set to make a small training set for our experiment. Note we did not use any activity detection and data augmentation techniques to our training set following the baseline implementation [28]. For testing, we use Voxceleb I verification trial list, consisting of 37720 non-target or target trials from 40 speakers. For each example, we use 512-point STFT as input feature. Mean and variance normalization on each frequency bin is performed.

The backbone deep neural network is similar to the baseline Resnet structure used in [29], but we use a more efficient Resnet18 structure to speed up training. Preactivate Resnet blocks [30] are used in our implementation.

The major testing criterion is to evaluate equal error rate and minimum detection cost function (minDCF) with target prior set to 0.01. Since our focus is on the discriminative power from embedding learning, we use cosine distance scoring as a simple back-end. The length variability is dealt with averaging pooling on the hidden layer.

TABLE I
RESULTS OF SYSTEMS ON VOXCELEB

Systems	Margin	EER	MinDCF(0.01)
Resnet34 Softmax [29]	-	5.04	0.543
Resnet18 Softmax (Baseline)	-	5.33	0.489
Resnet18 Modified Softmax	$m_1 = 1$	5.74	0.515
Resnet18 AM-Softmax	$m_3 = 0.1$	5.17	0.485
Resnet18 AM-Softmax	$m_3 = 0.2$	4.56	0.402
Resnet18 AM-Softmax	$m_3 = 0.3$	4.81	0.456
Resnet18 AM-Softmax + inter	$m_3 = 0.2$	**4.45**	**0.400**
Resnet18 AAM-Softmax	$m_2 = 0.2$	4.63	0.446
Resnet18 AAM-Softmax	$m_2 = 0.3$	4.55	0.443
Resnet18 AAM-Softmax	$m_2 = 0.4$	4.99	0.485
Resnet18 AAM-Softmax + inter	$m_2 = 0.3$	4.49	0.441
Resnet18 A-Softmax	$m_1 = 2$	4.58	0.429
Resnet18 A-Softmax	$m_1 = 3$	4.64	0.431
Resnet18 A-Softmax	$m_1 = 4$	4.75	0.468
Resnet18 A-Softmax + inter	$m_1 = 2$	4.46	0.427

B. Training Details

We train our neural network with SGD optimizer, the initial learning rate is set to 0.1 in all of our experiments and step decay to 0.001, annealing techniques are used in all the angular based training. Note besides annealing, we find the choice of the large learning rate and sample balance is also critical to a successful implementation of angular based systems on ASV. We set λ_{inter} to 0.01 and our implementation is based on Pytorch.

C. Results and Analysis

In Table I, the performance of the systems using softmax and different kinds of angular based loss functions are compared. We use the corresponding type-1 results given in [29] as the baseline. From our tested results, the Resnet18 and our small training set are sufficient to achieve reasonable results and conduct the experiments. The modified softmax system perform worse than softmax, due to the pruned bias term. AM-softmax, AAM-softmax and A-softmax systems with different margins are explicitly compared. The angular based systems perform consistently better than softmax, showing the effectiveness of the adding large angular margin. Comparing different angular based systems, we see that their improvements to the baseline are similar. Note that adding relatively small margin is enough to get ideal improvement on our ASV task. We get better performance if all these angular based losses are regularized with inter-class separability. We manage to achieve the best performance with AM-softmax with inter-class regularization at margin 0.2 among all the system we implement.

To further explore the effectiveness of angular based systems, we plot the embeddings of 40 different classes draw from the training set, using t-SNE to reduce dimension Fig. 1. We see that by training the neural network we obtain a linear separable embedding space. Within each class, the samples distribute more compactly when we use AM-softmax (m=0.2) comparing with the softmax, which clearly shows that angular based loss encourage better intra-class compactness.

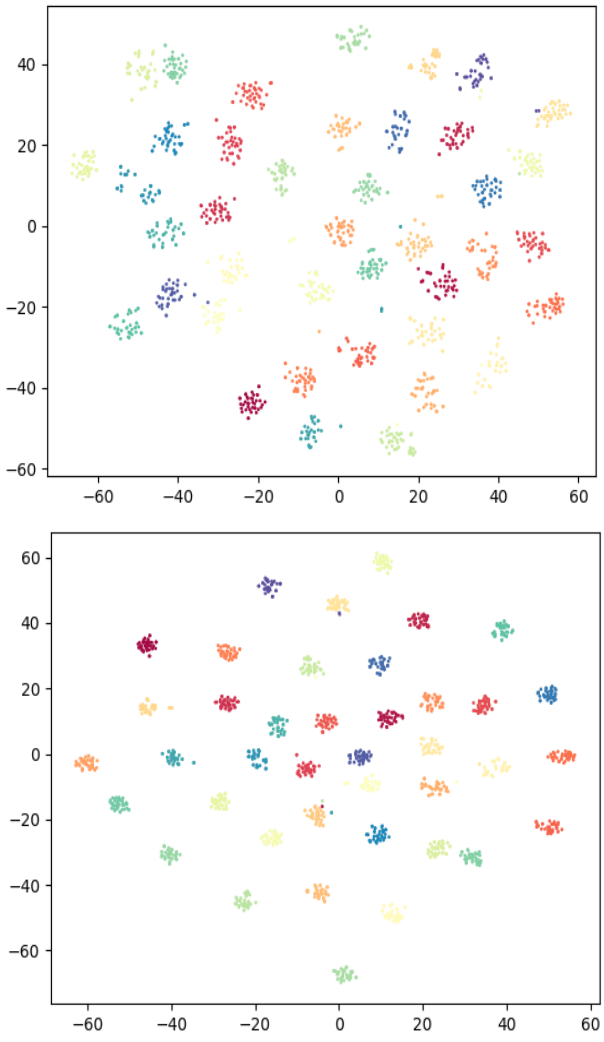

Fig. 1. Embedding plot of 40 classes from Voxceleb training data. Top: Softmax, Bottom: Angular based softmax. The embeddings are dimension reduced using t-SNE.

TABLE II
COMPARING ON INTER-CLASS SEPARABILITY

System	$SEP_{\mathbf{W}}$	S_b
Resnet18 Softmax	25.8	0.910
Resnet18 A-Softmax	39.3	0.907
Resnet18 A-Softmax + inter	**22.8**	**0.942**

We also explore the effectiveness of adding the inter-class regularization. In Talbe II, we first compare the *hyperspherical energy* in equation 6, which reflect the separability between the *cluster centers*. We see that using A-softmax, the *cluster centers* become less separably distributed in the embedding space, even comparing with the softmax system, which is not ideal to fully utilize the embedding space to learn better class separable embeddings. This motivate us to explicitly combining the inter-class regularization as a complement with A-softmax, which lead to reduced *hyperspherical energy*. This shows such regularization encourages inter-class separability. We also evaluate the between-class angular variance S_b like [16] given by,

$$S_b = \frac{1}{N}\frac{1}{C-1}\sum_{i}^{C} n_i \sum_{j,j\neq i}^{C}(1 - cos\langle m_i, m_j\rangle), \tag{9}$$

on the testing set. N is the sample number, C is the class number of the testing set, m_i is the mean vector from class i. A-softmax with inter-class regularization has larger between-class angular variance comparing to other systems. As shown in Fig. 2, we plot the score distribution from the Voxceleb non-target trials, which is a good indicator for the distance between examples from different classes in the testing set. We calculate their cosine distances and use a Student's t-distribution to fit results of each system. The A-softmax system has large score variance and many embeddings from different classes stay too close on the hypersphere. With inter-class regularization, the scores of non-target trails tend to distribute closer to zero with a smaller variance, indicating a large proportion of

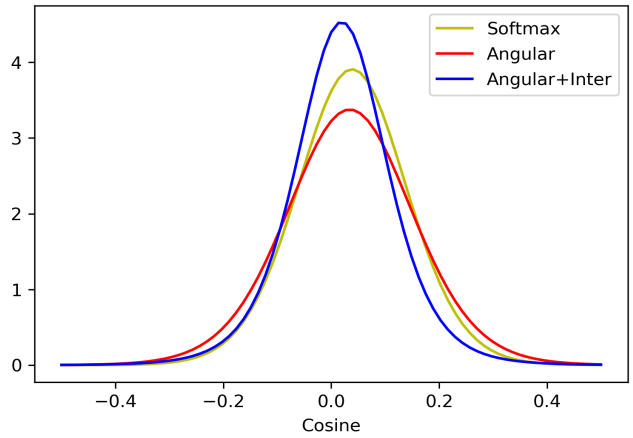

Fig. 2. Score distribution plot of the Voxceleb non-target trials, fit by Student's t-distributions. Inter-class regularization makes the scores compactly distributed near zero.

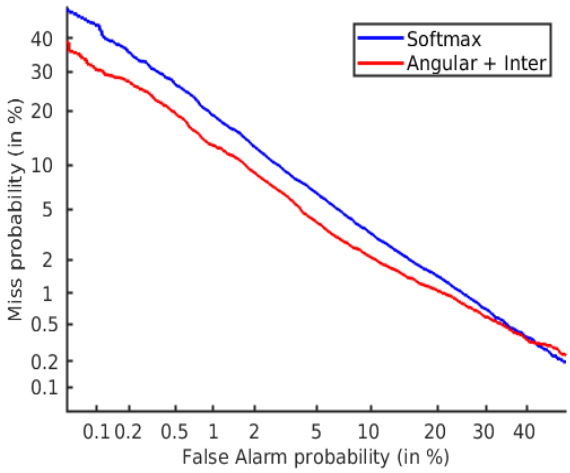

Fig. 3. DET plot for softmax system and angular based softmax system.

the angels between embeddings from different classes are 90 degrees. These experiments show that inter-class regularizing not only lead to a better-separated *cluster centers* but also well generalize to the class separability on the testing set.

We finally draw the detection error trade-off (DET) plot of the softmax system and inter-class regularized AM-softmax system in Fig. 3. We see that the proposed system outperform the baseline softmax system in almost all operation points. The gap between the two systems is especially large at lower false alarm range, suggesting that the angular based system can have superior performance on open-set speaker verification tasks.

IV. CONCLUSIONS & FUTURE WORKS

We have explored a sort of angular based embedding learning methods on ASV task. By adding a large angular margin, we manage to achieve better intra-class compactness. And with inter-class regularization, we effectively learned a more class-separable space. All these methods lead to more discriminative speaker embeddings and better speaker verification performance. The performance of several angular based learning methods has been compared on Voxceleb dataset. Angular based loss with inter-class regularization achieves apparently better results comparing with baseline softmax. The effectiveness of the inter-class regularization has also been studied, which improves the inter-class separability of training cluster centers and generalizes to testing data. In the future, we plan to conduct more detailed experiments on different strategies to add inter-class regularization. We will also conduct experiments on more datasets to see the robustness of these methods.

REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[3] 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.

[4] 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.

[5] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth annual conference of the international speech communication association*, 2011.

[6] 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.

[7] 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.

[8] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[9] J.-W. Jung, H.-S. Heo, I.-H. Yang, H.-J. Shim, and H.-J. Yu, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5349–5353.

[10] ——, "Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification," *extraction*, vol. 8, no. 12, pp. 23–24, 2018.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[20] S. Novoselov, V. Shchemelinin, A. Shulipa, A. Kozlov, and I. Kremnev, "Triplet loss based cosine similarity metric learning for text-independent speaker recognition," *Proc. Interspeech 2018*, pp. 2242–2246, 2018.

[21] Z. Huang, S. Wang, and K. Yu, "Angular softmax for short-duration text-independent speaker verification," *Proc. Interspeech, Hyderabad*, 2018.

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

[23] Y. Li, F. Gao, Z. Ou, and J. Sun, "Angular softmax loss for end-to-end speaker verification," *arXiv preprint arXiv:1806.03464*, 2018.

[24] 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.

[25] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, and L. Song, "Learning towards minimum hyperspherical energy," in *Advances in Neural Information Processing Systems*, 2018, pp. 6222–6233.

[26] W. Liu, Y.-M. Zhang, X. Li, Z. Yu, B. Dai, T. Zhao, and L. Song, "Deep hyperspherical learning," in *Advances in neural information processing systems*, 2017, pp. 3950–3960.

[27] Y. Duan, J. Lu, and J. Zhou, "Uniformface: Learning deep equidistributed representation for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3415–3424.

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

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

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.