INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Joint Feature Enhancement and Speaker Recognition with Multi-Objective Task-Oriented Network

Yibo Wu[1], *Longbiao Wang*[1,*], *Kong Aik Lee*[2,*], *Meng Liu*[1], *Jianwu Dang*[1,3]

[1]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, China
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan

yibo_wu@tju.edu.cn, longbiao_wang@tju.edu.cn, lee_kong_aik@i2r.a-star.edu.sg

Abstract

Recently, increasing attention has been paid to the joint training of upstream and downstream tasks, and to address the challenge of how to synchronize various loss functions in a multi-objective scenario. In this paper, to address the competing gradient directions between the speaker classification loss and the feature enhancement loss, we propose an asynchronous subregion optimization approach for the joint training of feature enhancement and speaker embedding neural networks. For the asynchronous subregion optimization, the squeeze and excitation (SE) method is introduced in the enhancement network to adaptively select important channels for speaker embedding. Furthermore, channel-wise feature concatenation is applied between the input feature and the enhanced feature to address the distortion of speaker information that is caused by enhancement loss. By using the proposed joint training network with asynchronous subregion optimization and channel-wise feature concatenation, we obtained relative gains of 11.95% and 6.43% in equal error rate on a noisy version of Voxceleb1 and VOiCES corpus, respectively.

Index Terms: far-field speaker verification, feature enhancement, squeeze and excitation, joint training.

1. Introduction

Automatic speaker verification (ASV) refers to the task of verifying whether a pair of utterances belong to the same speaker or not. With the advent of deep-speaker embedding [1, 2, 3, 4, 5], the performance of ASV over the telephone channel with a close-talking microphone has improved significantly. Although these deep-learning methods have made a lot of progress on clean speech signals, the performance could degrade dramatically in noisy and reverberant environments, which is typical in far-field smart home applications [6].

Different approaches have been proposed to compensate for the adverse impacts in far-field settings. At the signal and feature level, data augmentation [7, 8] and feature normalization [9] have been investigated for far-field ASV. At the model level, condition labels are used to train a multi-task adversarial network structure to extract noise-robust speaker embeddings [10, 11]. Furthermore, noise-robust speaker embeddings were extracted by reducing the distance between the embeddings of noisy utterance and its clean version [8].

Speech and feature enhancement methods have been studied so as to generate clean features from noisy speech. In the traditional speech enhancement approach for far-field ASV, the first step is to train a fronted-end enhancement model using pairs of clean and noisy speech. Then, the enhanced features are extracted for ASV. However, the direct use of enhancement loss based on mean square error (MSE) can lead to unwanted distortion of speaker information [12, 13, 14]. To address this problem, speaker-centric loss was proposed as a candidate solution [15, 6]. Using this method, first a fixed auxiliary speaker embedding network was trained. Then, the enhancement network was optimized using error propagated from the speaker embedding network. In [15], the voiceID loss of the speaker network is used to optimize the feature enhancement network and find the time-frequency bins that were most beneficial for ASV tasks. In [6], the deep feature loss calculated from the hidden layers of speaker network also showed considerable improvement over traditional methods.

However, treating the feature enhancement network and speaker verification network independently can cause information mismatch. This is because the weights in the speaker verification network remain frozen while updating the feature enhancement network, making it unaware of the distribution of the enhanced features. Therefore, we conjecture that a joint training approach is more appropriate, whereby the speaker loss and the enhancement loss are used to optimize the two networks simultaneously.

Nevertheless, two problems arise when considering a joint training approach. The first problem is due to the different convergence rates, orders of magnitude, and gradient directions of the speaker classification loss and the mean-square-error (MSE) loss. To address this problem, we propose the asynchronous subregion optimization technique. In particular, squeeze-and-excitation (SE) blocks are introduced in the feature enhancement network to filter important speaker information. The classification loss optimizes the SE blocks and the speaker verification network, while the MSE loss optimizes the remaining components in the enhancement network.

The second problem arising from a joint training approach is the potential for distortion of relevant speaker information. Even though the feature enhancement loss and the speaker loss are used to optimize the enhancement network simultaneously, some speech distortion caused by the enhancement loss is inevitable. This is especially true when the system only uses the enhanced feature as the input for the speaker network [6]. Therefore, channel-wise feature concatenation is applied between the input feature and the enhanced feature. By so doing, the first convolution layer of the speaker verification network can learn how to remove the noise signals from the enhanced features and learn the raw fine structures from the noisy features.

[*] corresponding authors

http://dx.doi.org/10.21437/Interspeech.2021-1978

2. Related Works

2.1. Mask based speech enhancement

Speech enhancement [16] aims to remove the additive noise and estimate a target clean speech from the noisy input. Let Y, X and N be the short-time Fourier transform (STFT) of the noisy input speech, target clean speech, and additive noise, respectively, such that

$$Y = X + N \tag{1}$$

The main goal of a mask-based feature enhancement network is to estimate a time-frequency (T-F) mask:

$$M = f_{enh}(Y) \tag{2}$$

where M is the estimated mask, and $f_{enh}(\cdot)$ denotes the feature enhancement network. The network is optimized with a mean square error (MSE) loss calculated between the enhanced and target clean features, as follows:

$$L_{enh} = \frac{1}{T}||Y \odot M - X||_2^2 \tag{3}$$

where T is the length of the input, the operator $\odot$ denotes pointwise multiplication, $Y \odot M$ is the enhanced feature, and X is the clean features. The MSE loss L_{enh} is computed over all T-F bins.

In this study, bidirectional long short-term memory (BLSTM) neural network were used in the feature enhancement model. In particular, we used 3 BLSTM layers, with 80 LSTM cells in each of the forward and backward directions. The BLSTM outputs were projected by a fully-connected linear layer and reshaped into the same size as the input features. After the fully-connected linear layer, a sigmoid function was used to generate the T-F mask with output values between 0 and 1.

2.2. ResNet for deep speaker embedding

Deep neural networks for speaker embeddings are typically based on a residual neural network (ResNet) [3, 17] or a time-delayed neural network (TDNN) [2]. We used a ResNet [18] for speaker embedding in this study.

There are four residual blocks in the ResNet, with a one-dimensional convolutional layer before each residual block. For each residual block, two convolutional layers with the same kernel size (3 × 3) and a stride of (1 × 1) were used. For the convolutional layers between blocks, the kernel size and stride were (5 × 5) and (2 × 2), respectively, and the number of channels varied from 64 to 512. After the residual blocks, a global average pooling layer was applied. Then, the speaker embedding was extracted with a fully connected layer. After that, the embedding was mapped into a number corresponding to the number of speakers in the training data. In this paper the speaker embedding neural network was trained by optimizing the angular softmax (A-softmax) [19] loss L_{spk}.

2.3. VoiceID loss for speech enhancement

VoiceID loss was proposed in [15] for training a speech enhancement network to find the time-frequency mask that is most beneficial for ASV tasks. The speaker embedding network was first trained. In the second step, the enhanced features that were estimated by the enhancement network was fed into the speaker network. The enhancement network is optimized by propagating the error from the speaker network while holding the weight of the speaker network fixed. Details are shown in Figure 1.

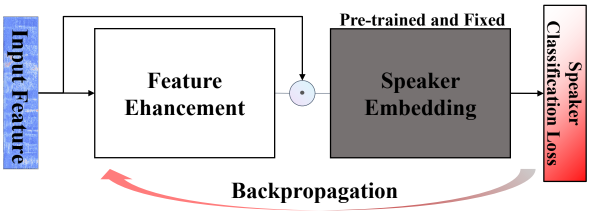

Figure 1: *VoiceID loss for speech enhancement.*

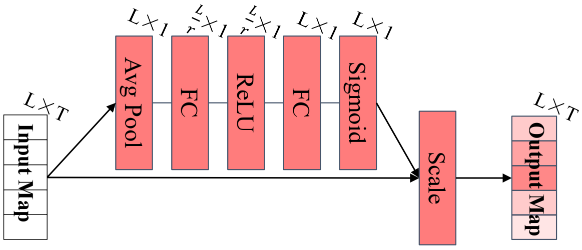

Figure 2: *SE block architecture for enhancement network.*

2.4. Joint training

A conventional joint training framework with a feature enhancement network and a speaker verification network was introduced as the joint training baseline. It included two parts: the BLSTM based feature enhancement network and the ResNet based speaker embedding neural network. Firstly, the input feature was fed into the feature enhancement network. Secondly, only the enhanced feature was utilized as the input feature for the speaker embedding neural network. Finally, the MSE loss optimized the enhancement network while the A-softmax optimized the enhancement network and speaker embedding network simultaneously.

3. Asynchronous Subregion Optimization

The MSE loss for feature enhancement and the A-softmax loss for speaker classification have different convergence rates, orders of magnitude, and gradient directions leading to a competing scenario when both losses are used to optimize the enhancement network jointly. To address this problem, asynchronous subregion optimization is proposed in this paper. The central idea is to update specific regions or layers of the neural network with one loss function, while updating the remaining regions with another loss function. In particular, we used the MSE loss to update the BLSTM layers of the feature enhancement network, and the A-softmax loss to update the SE blocks inserted after each BLSTM-FC layer. By doing so, we ensure that gradients derived from the MSE and A-softmax losses, which might be mutually competing, are applied on different sets of weights. We refer to this as the asynchronous optimization since the weights update are accomplished with separate error propagation. The input map dimension for each SE block is $L \times T$, where the T denotes the number of time bins. Different from conventional usage of SE block [20], we apply the SE block along the L dimension of the input map (the time dimension was variable in the test stage). Details are shown in Figure 2.

Speech enhancement aims to improve the speech quality by suppressing noise. However, the artifacts and distortions caused by speech enhancement loss might influence speaker verifica-

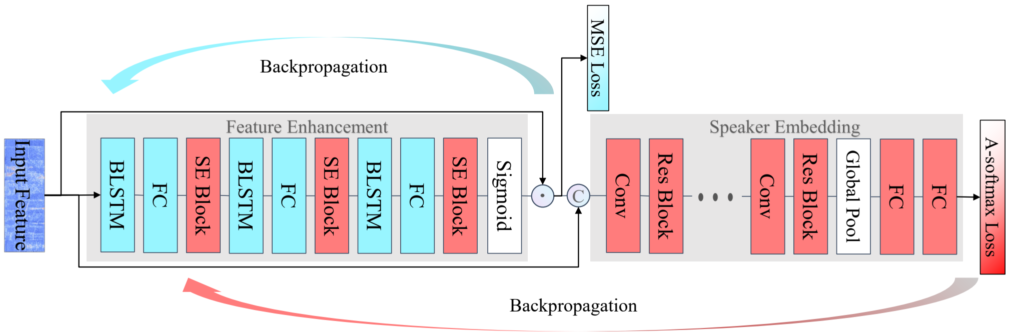

Figure 3: *Proposed joint training architecture with asynchronous subregion optimization and channel-wise feature concatenation. A-softmax loss optimizes the SE blocks and speaker network. MSE loss optimizes the remaining components in the enhancement network.* ⓒ *denotes channel-wise feature concatenation.*

Table 1: *Performance on the Voxceleb1 test set (EER[%]). A. denotes asynchronous subregion optimization, SE denotes using the enhancement network with SE blocks, and CFC denotes channel-wise feature concatenation.*

SV training set		D_C			$D_C + D_N$				
Noise Type	**SNR**	**Resnet**	**SA**	**Voiceid**	**Resnet**	**SA**	**Voiceid**	**Joint**	**Joint A. SE+CFC**
Original set		8.81	8.72	8.90	9.10	9.20	9.43	8.39	**7.60**
Babble	0	28.52	24.66	26.20	27.84	22.60	27.28	22.33	**20.11**
	5	20.79	17.58	18.08	18.87	15.59	18.70	14.59	**12.02**
	10	15.11	13.50	13.37	13.59	12.50	13.87	11.20	**9.63**
	15	11.93	11.32	11.08	11.22	10.94	11.27	9.75	**8.48**
	20	10.41	10.01	10.00	10.13	9.93	10.05	8.97	**7.99**
Music	0	30.61	25.46	25.27	17.92	18.62	17.85	14.95	**12.92**
	5	21.73	18.79	17.37	13.28	14.35	13.42	11.70	**10.10**
	10	15.20	13.91	13.04	11.14	11.76	11.41	9.97	**8.95**
	15	11.71	11.11	10.67	10.15	10.31	10.41	9.21	**8.35**
	20	10.08	9.69	9.65	9.64	9.58	9.81	8.78	**7.95**
Other	0	31.26	26.01	23.80	17.70	19.36	17.68	15.23	**13.12**
	5	24.32	20.32	17.80	14.34	15.76	14.11	12.05	**10.57**
	10	18.29	15.96	13.97	12.09	13.34	12.01	10.47	**9.28**
	15	14.00	12.83	11.72	10.85	11.34	10.76	9.49	**8.59**
	20	11.30	10.82	10.40	9.97	10.18	10.09	8.95	**8.10**
Average		17.75	15.67	15.08	13.61	13.46	13.63	11.63	**10.24**

tion performance as much as the noise itself especially for high SNR scenario. Therefore, we concatenated the input feature and enhanced feature in the channel dimension to resolve speech distortion. This augmentation method removed the noise signals from the enhanced features and learned the raw fine structures from the noisy features, so that it could alleviate the problems resulting from a lack of speaker information.

Figure 3 illustrates the network architecture of our proposed joint training framework for robust end-to-end ASV. Firstly, the input spectrogram is enhanced by the speech enhancement network. Secondly, the enhanced feature is concatenated with the input feature in the channel dimension. Then the concatenated features are used as the input for the end-to-end ASV network. Finally, we use asynchronous subregion optimization to optimize the joint training model.

4. Experiments

4.1. Dataset

Experiments were conducted on the Voxceleb 1 [21] and VOiCES [22] datasets. The development set contained 148,642 utterances from 1,211 speakers, and the test set contained 4,874 utterances from 40 speakers, which were used to construct 37,720 test trials. The MUSAN dataset [23] was used as the noise source. Since the Voxceleb1 dataset was collected from online videos, the audio contained in the dataset is not strictly in a clean condition, however we assumed the original data to be a clean dataset and generated noisy data based on that original data. We divided the MUSAN dataset into two disjoint sets, each of which was used to augment the development and test set of Voxceleb1 specifically. Therefore, the noise samples used to augment the test set are not seen in the development set. The Voxceleb1 development dataset was regarded as the clean training set (D_C). Moreover, for the noisy training set (D_N), we corrupted each utterance with a Signal to Noise Ratio (SNR) randomly chosen between 0 and 20 on a linear scale from the training subset of MUSAN. The amount of noise used in the augmented set (D_N) was the same as the clean training set (D_C). For the noisy test set, we generated all noise types and each type has all possible SNRs represented in the set $\{0, 5, 10, 15, 20\}$.

The Voxceleb1 set is a simulated test set and we chose the Voices Obscured in Complex Environmental Settings (VOiCES) corpus for evaluation our system with real data. The VOiCES development dataset consists of 15,904 audio segments from 196 speakers. The dataset represents different rooms, microphones, noise distractors, and loudspeaker angles.

4.2. Experimental setup

Speech signals were first converted to a 161-dimensional spectrogram, and each frame had a frame length of 20 ms and a frame shift of 10 ms. 64 utterances were grouped as one batch and fed into the systems, and the training epochs of each system was 50. During the training stage, 3s utterances ere selected from each raw waveform. For SE block-based feature enhancement network, the input feature dimension was (161 × 300). The dimension was (161 × 1) after the average pooling. The dimension was then compressed from 161 to 32 after the first fully-connected layer in the SE block. The whole utterance was used to extract speaker embeddings during the test stage. The equal error rates were used as the evaluation index in our study.

4.3. Experimental results

There were four types of models trained to validate our ideas. ResNet means directly test the ResNet model without any enhancements. SA denotes spectrum approximation, wherein we trained a mask-based BLSTM enhancement network using MSE loss to enhance the test sets. VoiceID model was the BLSTM enhancement network optimized by A-softmax loss. Joint denotes the model wherein we jointly optimized the feature enhancement network and speaker verification network.

Table 1 shows the results from the Voxceleb1 test set from eight different models. As expected, the ResNet trained with only the clean training data performed worst, and data augmentation strategies greatly improved the performance of the system. Joint training was more beneficial than separate training, whether using speaker loss or enhancement loss. It is important to note that our proposed model attains the best performance and achieved an 11.95% increase in average result when compared with the baseline joint training model.

Table 2: *Ablation analysis of channel-wise feature concatenation (EER[%]).*

Noise Type	SNR	Joint	Joint CFC	Joint A. SE	Joint A. SE+CFC
Original set		8.39	**8.05**	8.04	**7.60**
Babble	0	22.33	**21.65**	22.69	**20.11**
	5	14.59	**13.69**	14.02	**12.02**
	10	11.20	**10.36**	10.56	**9.63**
	15	9.75	**8.95**	9.11	**8.48**
	20	8.97	**8.41**	8.56	**7.99**
Music	0	14.95	**13.87**	14.13	**12.92**
	5	11.70	**10.95**	10.92	**10.10**
	10	9.97	**9.55**	9.65	**8.95**
	15	9.21	**8.78**	8.81	**8.35**
	20	8.78	**8.33**	8.47	**7.95**
Other	0	15.23	**14.10**	14.77	**13.12**
	5	12.05	**11.36**	11.76	**10.57**
	10	10.47	**9.89**	10.11	**9.28**
	15	9.49	**9.07**	9.28	**8.59**
	20	8.95	**8.54**	8.60	**8.10**
Average		11.63	**10.97**	11.22	**10.24**

In Table 2, we present the performance of the feature augmentation. The left half of the table shows results from the joint training network without SE blocks and the right half shows results from when the enhancement component was BLSTM with SE blocks. From the results, it is clear that the improvement of channel-wise feature concatenation was stable.

We analyzed the effect of asynchronous subregion optimization and SE blocks for enhancement network, as depicted in Table 3. The first model used the speaker loss and enhancement loss to optimize the whole enhancement network cooperatively. Differing from the first model, the second model added SE blocks in its feature enhancement component. For the third model, asynchronous subregion optimization was applied. From the comparison of the first and second models, using only the SE module proved effective. However, more importantly, the optimization method we proposed is more effective, when comparing the second and third models.

Table 3: *Ablation analysis of asynchronous subregion optimization (EER[%]).*

Noise Type	SNR	Joint	Joint SE	Joint A. SE
Original set		8.39	8.24	**8.04**
Babble	0	22.33	23.42	**22.69**
	5	14.59	14.58	**14.02**
	10	11.20	10.71	**10.56**
	15	9.75	9.25	**9.11**
	20	8.97	8.63	**8.56**
Music	0	14.95	14.56	**14.13**
	5	11.70	11.33	**10.92**
	10	9.97	9.81	**9.65**
	15	9.21	9.01	**8.81**
	20	8.78	8.53	**8.47**
Other	0	15.23	14.78	**14.77**
	5	12.05	12.	**11.76**
	10	10.47	10.40	**10.11**
	15	9.49	9.43	**9.28**
	20	8.95	8.88	**8.60**
Average		11.63	11.48	**11.22**

Table 4: *Performance on VOiCES set (EER[%]).*

ResNet	SA	Voiceid	Joint	Joint A. SE+CFC
14.71	15.18	13.57	13.37	**12.51**

In Table 4, we present the results from tests with the VOiCES dataset. For every model, the training data as mixed data ($D_C + D_N$). Similar to the previous results, our proposed model achieved the best performance of all models tested and obtained a 6.43% (from 13.37% to 12.51%) relative improvement for EER.

5. Conclusion

In this study, we proposed the asynchronous subregion optimization for the joint training frameworks involving feature enhancement and speaker verification. The proposed optimization strategy addresses the problem of competing gradients in multi-objective networks, targeting both speaker classification loss and enhancement loss. We also employed channel-wise feature concatenation between the enhanced feature and the input feature to address speaker information distortion caused by enhancement loss. The proposed joint training framework demonstrates accuracy and consistency in performance on simulated noisy Voxceleb dataset and the realstic VOiCES dataset.

6. Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grant 61771333 and the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

7. References

[1] G. Bhattacharya, M. J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification." in *Interspeech*, 2017, pp. 1517–1521.

[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] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[5] D. Zhou, L. Wang, K. A. Lee, Y. Wu, M. Liu, J. Dang, and J. Wei, "Dynamic Margin Softmax Loss for Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3800–3804.

[6] S. Kataria, P. S. Nidadavolu, J. Villalba, N. Chen, P. Garcia-Perera, and N. Dehak, "Feature enhancement with deep feature losses for speaker verification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7584–7588.

[7] 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.

[8] D. Cai, W. Cai, and M. Li, "Within-sample variability-invariant loss for robust speaker recognition under noisy environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6469–6473.

[9] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceedings of 2001 A Speaker Odyssey: The Speaker Recognition Workshop*. European Speech Communication Association, 2001, pp. 213–218.

[10] Z. Meng, Y. Zhao, J. Li, and Y. Gong, "Adversarial speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6216–6220.

[11] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang, and B. Xia, "Training multi-task adversarial network for extracting noise-robust speaker embedding," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6196–6200.

[12] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836–845, 2014.

[13] M. Kolbœk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *2016 IEEE spoken language technology workshop (SLT)*. IEEE, 2016, pp. 305–311.

[14] Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, and M. Iwahashi, "Dnn-based amplitude and phase feature enhancement for noise robust speaker identification." in *Interspeech*, 2016, pp. 2204–2208.

[15] S. Shon, H. Tang, and J. Glass, "VoiceID Loss: Speech Enhancement for Speaker Verification," in *Proc. Interspeech 2019*, 2019, pp. 2888–2892.

[16] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[17] K. Li, M. Akagi, Y. Wu, and J. Dang, "Segment-Level Effects of Gender, Nationality and Emotion Information on Text-Independent Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 2987–2991.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] 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.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

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

[22] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout *et al.*, "Voices obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

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