VoiceID Loss: Speech Enhancement for Speaker Verification

Suwon Shon, Hao Tang, James Glass

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA

{swshon,haotang,glass}@mit.edu

Abstract

In this paper, we propose VoiceID loss, a novel loss function for training a speech enhancement model to improve the robustness of speaker verification. In contrast to the commonly used loss functions for speech enhancement such as the L2 loss, the VoiceID loss is based on the feedback from a speaker verification model to generate a ratio mask. The generated ratio mask is multiplied pointwise with the original spectrogram to filter out unnecessary components for speaker verification. In the experiments, we observed that the enhancement network, after training with the VoiceID loss, is able to ignore a substantial amount of time-frequency bins, such as those dominated by noise, for verification. The resulting model consistently improves the speaker verification system on both clean and noisy conditions.

Index Terms: speech enhancement, speaker verification

1. Introduction

By exploiting large quantities of data, especially via data augmentation [1, 2], speaker embedding methods are now able to surpass the conventional i-vector [3] approach. Many variants, largely based on multiclass classification, have been proposed to extract robust embeddings from speech. The freely available speaker recognition dataset, Voxceleb [4, 5], also accelerated speaker verification improvements by having a common benchmark for different approaches to compare to.

While noise robustness is a general and hard problem for many speech processing tasks, there are relatively few studies about the use of speech enhancement for speaker recognition tasks. This is because, rather than having multiple preprocessing steps to remove noise, training speaker recognition systems using a large and diverse dataset is a simple and powerful solution. Speaker recognition systems naturally become robust to noisy environments when trained on a large dataset augmented with real or synthetic noise types. This approach is especially appealing for large, overparameterized neural networks. Besides, the objective of speech enhancement is to improve the speech quality by suppressing noise, and makes no guarantees to downstream tasks such as speaker verification. Even worse, the artifacts and distortions caused by speech enhancement might even deteriorate speaker verification performance [6].

For these reasons, only a few studies have explored speech enhancement for speaker verification [7, 6] and most recent studies are based on the i-vector approach [8, 9, 10]. They used a Denoising Autoencoder (DAE) to generate an enhanced signal from the noisy signal. As shown in Figure 1.(a), the objective of DAE is to minimize the L2 loss between the output of the model and the clean speech. During training, not only noisy-clean pairs but clean-clean pairs are also needed to prevent the DAE from deteriorating the quality of the clean signal. These studies show improvement on the noisy and mismatched conditions, but have marginal gains on the clean and matched conditions. This is ex-
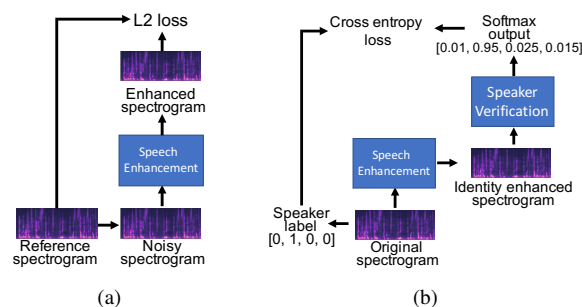

Figure 1: *A diagram for speech enhancement (a) using L2 loss, (b) using VoiceID loss.*

pected because the objective of DAE is to generate outputs that are closer to the inputs. Another study [11] avoids deterioration from artifacts and distortions by training an individual speaker verification system using enhanced clean speech for each enhancement approach.

To improve the effects of speech enhancement for speaker verification, we introduce a VoiceID loss that uses the error signal of a speaker verification model to train a speech enhancement model. The overall structure of this network is shown in Figure 1. Rather than minimizing the L2 loss between the output and the clean speech, we pass the enhanced signals to the speaker verification model, and compute the multiclass cross entropy between the output of the speaker model and the ground truth speaker label. The speech enhancement model is updated based on the cross entropy loss. Since the speaker verification system is a Deep Neural Network (DNN), the gradient can be backpropagated end to end.

VoiceFilter [12] is a similar approach to separate the voice of interest. They generate a ratio mask to filter out unwanted speaker's voice from the mixture of multiple speakers to obtain homogeneous speech of the target speaker. By pairing the training set with different noise types, this system could be also used for speech enhancement. However, to enable this, the system needs a strong prior, a speaker embedding, from the target speaker. Thus, the performance of separation and enhancement would heavily depend on the speaker embedding, which can be unreliable when there is only a small amount of speech from the target speaker. It is also not applicable to unseen speakers.

Another similar approach [13] is introduced for Automatic Speech Recognition (ASR). They have a DNN-based spectral mapper that extracts robust features from noisy speech. The spectral mapper is trained on a fidelity loss and a mimic loss. The fidelity loss is an L2 loss between the output of the spectral mapper on the clean speech and the output on the noisy speech. The mimic loss, however, is an L2 loss between the posterior probabilities of senones on the clean speech and the noisy speech. Ideally, the mimic loss should be an error between the ground truth senone class and the posterior probabilities of senones on the noisy speech. However, it is common that the
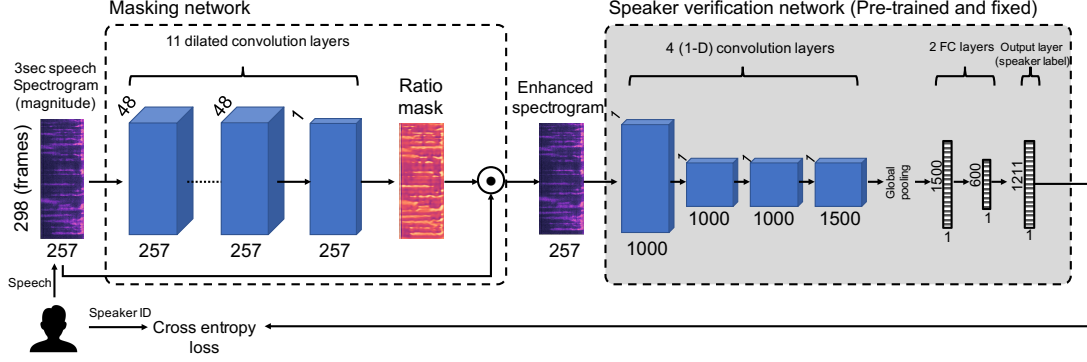
Figure 2: *A flow chart for the VoiceID loss.*

ground truth alignments on the speech are not accessible, so they use the posterior probabilities on the clean speech as the ground truth. For this reason, the mimic loss might not be sufficient to minimize the word error rate on the noisy speech, and the fidelity loss is required to compensate the mismatch. This is also shown in their empirical results, where the best scale between losses is achieved with 90% of fidelity loss and only 10% of mimic loss.

In this paper, we only use the feedback from the verification network, i.e., the VoiceID loss. In this way, the enhancement network has the flexibility to find the important time-frequency bins for speaker verification, though the quality of speech might not be improved. From the experiment, we observed this the VoiceID loss is able to remove the noisy bins to improve the speaker verification performance on both noisy and clean condition, even for the unseen type of noise. We describe a system architecture and experimental result in subsequent sections.

2. Speech enhancement using VoiceID loss

The system architecture is shown in Figure 2. At first, the verification network needs to be trained using a training dataset. After training, the weights of verification network are held fixed, i.e., not updated in the subsequent steps. At the second step, the masking network and the verification network are connected to each other to form a single network. Specifically, the masking network generates a ratio mask from the input spectrogram. Then the mask is multiplied pointwise with the input spectrogram. Finally, the masked spectrogram is fed into the verification network to generate a verification output. The cross entropy loss is computed based on the output and the ground truth speaker label of the input speech.

Table 1: *Masking network model structure.*

Layer	Filters / Size	Dilation	Context
conv1	48 / 1 x 7	1x1	1x7
conv2	48 / 7 x 1	1x1	7x7
conv3	48 / 5 x 5	1x1	11x11
conv4	48 / 5 x 5	2 x 1	19x11
conv5	48 / 5 x 5	4 x 1	35x11
conv6	48 / 5 x 5	8 x 1	67x11
conv7	48 / 5 x 5	1 x 1	71x15
conv8	48 / 5 x 5	2 x 2	79x23
conv9	48 / 5 x 5	4 x 4	95x39
conv10	48 / 5 x 5	8 x 8	127x71
conv11	1 / 1 x 1	1 x 1	127x71

2.1. Noisy dataset generation

First, we generate a set of data for multiple training and test settings. We use the Voxceleb1 development set ($\mathcal{D}$) to train the networks and the test set ($\mathcal{T}$) to validate the networks. Since the dataset is collected from YouTube, the dataset is moderately noisy, but we regard the original set as the clean set. We use the noise recordings from MUSAN [14] to generate corrupted versions of the Voxceleb1 development set ($\mathcal{D}^{\mathcal{N}}$) and the test set ($\mathcal{T}^{\mathcal{N}}$). Specifically, we divide the MUSAN dataset into two disjoint sets, each of which are used to augment the development and test set of Voxceleb1. We make sure we have the same types of noise in both the development and the test set, and the noise samples used to augment the test set are not seen in the development set. MUSAN consists of 4 categories of noise types: noise, music, babble, and reverberation. (The noise category contains multiple types of stationary and non-stationary noises. See [14]). For the development set ($\mathcal{D}^{\mathcal{N}}$), we corrupt each utterance with a Signal to Noise Ratio (SNR) randomly chosen between 0 and 20 in linear scale. For the test set ($\mathcal{T}^{\mathcal{N}}$), we consider all four types of noise and all SNRs (in db) in the set $\{0, 5, 10, 15, 20\}$.

The Voxceleb1 development set ($\mathcal{D}$) has a total of 147,935 utterances from 1,211 speakers. The noise augmented dataset ($\mathcal{D}^{\mathcal{N}}$) is generated with same amount as the original development set ($\mathcal{D}$). The Voxceleb1 test set ($\mathcal{T}$) has 18,860 verification pairs for each positive and negative test pair, i.e., a combination of 4,715 utterances from 40 speakers. Each noisy test set ($\mathcal{T}^{\mathcal{N}}$) is generated with same amount as the original test set ($\mathcal{T}$).

2.2. Speaker Verification Network

The speaker verification network uses 1-dimensional convolutions to consider all frequency bands at once. The network consists of 4 1D convolution layers (i.e., filters of size 40×5, 1000×7, 1000×1, 1000×1 with strides 1, 2, 1, and 1 and numbers of filters 1000, 1000, 1000, and 1500) and two fully connected (FC) layers (of size 1500 and 600). There is a global average pooling layer located between the last convolution layer and first FC layer. We use a similar structure in the previous study [15], except here spectrograms are used as inputs. We use 257 frequency-bin spectrogram as input with the 25ms window size and 10ms shift to represent the speech signal. We do not use any normalization on the spectrogram. We only use the magnitudes after the short-time Fourier transform and a power-law compression with p = 0.3 (i.e., $A^{0.3}$ where A is the magnitude spectrogram). In the training phase, we use a 298-frame fixed-length segment as input. We train two verification models using $\mathcal{D}$ and $\mathcal{D} + \mathcal{D}^{\mathcal{N}}$. Multicondition training using $\mathcal{D} + \mathcal{D}^{\mathcal{N}}$ is re-

Table 2: *EER results obtained on the Voxceleb1 test set.*

Verification network		Using original set $\mathcal{D}$						Using original and augmented set $\mathcal{D}+\mathcal{D}^{\mathcal{N}}$					
Enhancement		-		Proposed		DAE		-		Proposed		DAE	
Enhancement network		-		Using $\mathcal{D}^{\mathcal{N}}$		Using $\mathcal{D}+\mathcal{D}^{\mathcal{N}}$		-		Using $\mathcal{D}+\mathcal{D}^{\mathcal{N}}$		Using $\mathcal{D}+\mathcal{D}^{\mathcal{N}}$	
Type	SNR	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF
Original test set $\mathcal{T}$		7.73	0.608	**6.99**	**0.590**	7.73	0.608	7.01	0.592	**6.79**	**0.574**	6.93	0.589
Noise	20	10.34	0.761	**8.10**	**0.675**	10.02	0.738	8.08	0.659	**7.83**	**0.639**	8.28	0.671
	15	13.05	0.909	**9.32**	**0.699**	11.45	0.833	8.99	0.720	**8.69**	**0.686**	8.96	0.761
	10	17.71	0.987	**11.24**	**0.770**	14.00	0.943	10.36	0.770	**9.86**	**0.747**	10.73	0.869
	5	24.34	0.999	**14.78**	**0.885**	18.01	0.988	12.90	0.851	**12.26**	**0.830**	13.51	0.958
	0	31.76	1.000	**20.82**	**0.983**	23.87	0.998	17.68	0.945	**16.56**	**0.938**	18.32	0.994
Music	20	8.97	0.710	**7.54**	**0.666**	9.32	0.714	7.73	0.670	**7.48**	**0.635**	7.82	0.651
	15	10.60	0.764	**8.23**	**0.715**	10.27	0.743	8.43	0.695	**8.10**	**0.677**	8.42	0.692
	10	14.10	0.883	**9.72**	**0.760**	11.75	0.808	9.73	0.760	**9.13**	**0.733**	9.54	0.728
	5	20.37	0.992	**13.00**	**0.819**	15.15	0.941	12.28	0.833	**11.44**	**0.818**	11.76	0.846
	0	29.03	1.000	**18.89**	**0.937**	20.41	0.993	17.45	0.935	16.24	**0.913**	**15.96**	0.961
Babble	20	12.87	0.837	**10.16**	**0.781**	11.34	0.778	9.17	0.725	**8.99**	**0.705**	9.55	0.723
	15	18.83	0.931	**13.50**	**0.864**	14.45	0.881	11.68	**0.793**	**11.25**	0.807	12.10	0.801
	10	28.78	0.991	**21.18**	**0.944**	21.37	0.969	17.38	**0.922**	**16.66**	0.926	17.41	0.941
	5	38.74	1.000	**33.39**	**0.996**	33.14	0.997	28.21	**0.992**	**27.12**	0.996	29.19	0.992
	0	44.64	1.000	42.20	1.000	43.30	0.999	38.72	1.000	**37.96**	1.000	41.11	0.999
Reverb	Small room	13.81	0.835	**10.02**	**0.744**	13.54	0.831	10.52	0.725	**9.94**	**0.708**	11.52	0.814
	Large room	13.74	0.825	**10.11**	**0.756**	14.09	0.999	10.64	0.724	**10.17**	**0.691**	11.47	0.792

garded as the standard approach to train a noise-robust speaker verification model. We extract speaker embeddings from last FC layer.

2.3. Masking Network for Enhancement

The masking network consists of 11 dilated convolution layers, and Table 1 shows the configuration of each layer. We use the same setting for the extracting the spectrograms. To generate a ratio mask, a sigmoid function is used at the last convolution layer to have values between 0 and 1. For other layers, we use ReLUs for non-linear activations. Once we have the ratio mask from the last layer, the input spectrogram is multiplied with the mask and is then fed into the verification network. Training is done using the multiclass cross entropy objective with the ground truth speaker label and the verification network softmax output. The original Voxceleb1 test set ($\mathcal{T}$) is used as a validation set. We choose the masking network that has the best Equal Error Rate (EER) on the validation set.

3. Experiments

We use the original Voxceleb1 test set and the augmented test set to evaluate the Voxceleb1 verification task. Cosine similarity is used to measure the score between two utterances. Performance is evaluated using EER and the Detection Cost Function (DCF). In this paper, DCF is the average of two minimum DCF scores when P_{target}, a priori probability of the specified target speaker, is 0.01 and 0.001.

Performance comparison is done with DAE-based speech enhancement [8, 10, 9]. We use an 8-layer time-delay neural network (TDNN) [16] with 1000 hidden units per layer for enhancement. The architecture is the same as in [17], and the effective context size is 25 frames. We train the TDNN by minimizing the L2 loss for 10 epochs with step size 0.05 and gradient clipping of norm 5. The batch size is one utterance. After the first 10 epochs, we train the network for another 10 epochs starting from the model with the best L2 loss on the development set, with the same setting except that the step size is 0.00375 decayed by 0.75 after every epoch. The best model is chosen based on the L2 loss on the development set.

Since our proposed approach is the first to use only speaker identity for speech enhancement, we strongly recommend the

readers to listen to the samples on the demo page[1]. The inverse short-time Fourier transform is used to generate waveforms with the enhanced magnitudes and the original noisy phase. Note that the objective of the proposed enhancement is for the verification network, not for a human listener. Figure 3 shows example spectrograms.

3.1. Result

The performance (as shown in Table 2) is based on two types of verification model, a model trained using the clean dataset $\mathcal{D}$ and a model trained using the augmented dataset $\mathcal{D} + \mathcal{D}^{\mathcal{N}}$. In the clean setting, both the proposed masking approach and the DAE approach show significant improvement. The proposed masking approach shows improvement in all SNR settings consistently, but the DAE is only effective under low SNR settings. The relative improvement become marginal if we use the augmented model, as shown is the right part of the figure, but still, the proposed masking approach is effective in almost all cases. We also observe that the proposed approach shows remarkable performance under the reverberation compared to the DAE.

Table 3 shows the performance of models under unseen noise types. We exclude the musical noise from the augmented development set ($\mathcal{D}^{\mathcal{N}-\mathcal{M}}$). In this case, both the masking and verification networks are not exposed to musical noise during training. For the setting of unseen noise, the proposed approach also shows better performance than the DAE.

Objective measures for speech enhancement quality such as Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) are used for comparison in Figure 4. As expected, we observe that better speech quality does not imply better speaker verification. This indicates that speech enhancement should be customized to the eventual downstream task for maximum effectiveness.

3.2. Discussion

Interestingly, the proposed approach improves performance even in the clean setting for both verification models. This means that the VoiceID loss removes not only the noise but also unnecessary time-frequency bins from the spectrograms. The improvement also shows in the frame-level cosine similar-
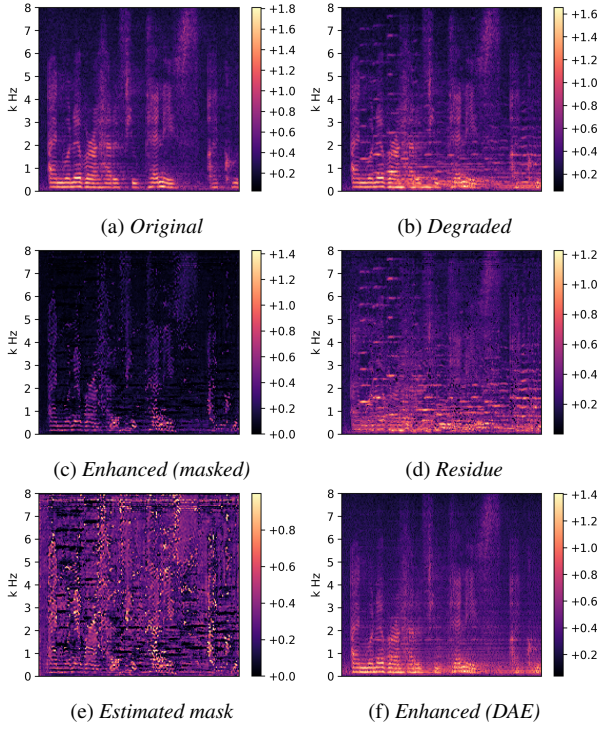
---

[1]https://people.csail.mit.edu/swshon/supplement/voiceid-loss

(a) *Original*   (b) *Degraded*

(c) *Enhanced (masked)*   (d) *Residue*

(e) *Estimated mask*   (f) *Enhanced (DAE)*

Figure 3: *Example spectrograms: (a) the original spectrogram from Voxceleb1 test set, a sample never seen during training (b) a degraded sample using a musical noise with an SNR of 0 (c) the result produced by the masking network (d) the residue between the masked and the original (e) the ratio mask from the masking network (f) the result from DAE.*

ity matrix, as shown in Figure 5. The analysis approach of using frame-level matrix is first introduced in [15]. We follow the same approach to compute the matrix before and after enhancement using two utterances from the same speaker in the TIMIT dataset, a clean and studio-level dataset. In Figure 5 (a), we see low scores between the different phonemes and high scores between the same phonemes. These low scores become close to 0 after enhancement. We hypothesize that the ambiguity of the different phonemes is removed by the mask and only similar phonemes with strong similarity have high scores.

A limitation of this study is that all experiments are done based on the speaker verification system in the previous study [15]. We do not use Angular Softmax [18, 19], Probabilistic Linear Discriminant Analysis (PLDA), ResNet [20, 21], and
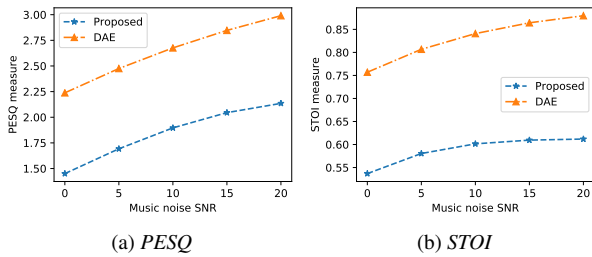

(a) *PESQ*   (b) *STOI*

Figure 4: *Speech quality measures after enhancement comparing the proposed approach and the DAE.*

Table 3: *Performance under unseen noise type (musical noise). Both verification and mask network trained noise augmented development set without any musical noise ($\mathcal{D}^{\mathcal{N}-\mathcal{M}}$).*

| Verification network		Using $\mathcal{D}$		Using $\mathcal{D} + \mathcal{D}^{\mathcal{N}-\mathcal{M}}$	
Unseen Noise SNR	Enhancement	EER	DCF	EER	DCF
20	-	8.97	0.710	8.00	0.702
	Proposed	**7.94**	**0.701**	**7.76**	**0.691**
	DAE	9.53	0.729	7.90	0.673
15	-	10.60	0.764	8.66	0.732
	Proposed	**8.74**	**0.747**	**8.49**	**0.712**
	DAE	10.49	0.777	8.52	0.732
10	-	14.10	0.883	10.32	0.774
	Proposed	**10.62**	**0.811**	9.95	**0.766**
	DAE	12.76	0.847	**9.93**	0.771
5	-	20.37	0.992	13.57	0.872
	Proposed	**14.19**	**0.864**	**12.39**	0.865
	DAE	16.92	0.953	13.05	**0.846**
0	-	29.03	1.000	19.73	0.970
	Proposed	**21.00**	**0.965**	**17.28**	0.958
	DAE	24.01	0.998	18.94	**0.955**

(a) *Original*   (b) *After enhancement*
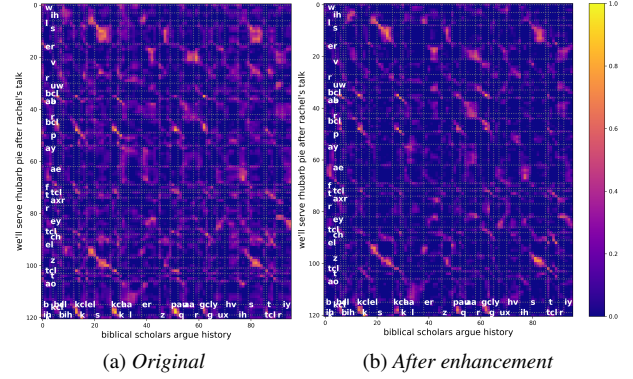
Figure 5: *A frame-level cosine similarity matrix between two sentences spoken by same person in TIMIT.*

various utterance aggregation approaches [22, 18, 23], which show better performance than the Softmax and Cosine similarity back-end. Also, we do not consider acoustic features such as log-Mel filter-banks. We believe these variants would give more robustness in overall performance with similar margin with and without enhancement. In the future, we will consider a further study thoroughly on the use of speech enhancement with the cutting-edge verification system.

4. Conclusion

Motivated by the discrepancy in objectives between speech enhancement and speaker verification, we propose a novel speech enhancement approach using the VoiceID loss to improve speaker verification. The proposed approach uses speaker identity information directly to generate ratio masks for emphasizing voice characteristics and filtering out unnecessary time-frequency bins, such as noise or even speech that does not carry strong voice characteristics. Experimental results show the effectiveness of the proposed approach combined with the speaker verification model in both clean and noisy settings.

5. References

[1] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, and Y. Carmiel, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Interspeech*, 2017, pp. 999–1003.

[2] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-End Text-Dependent Speaker Verification," in *ICASSP*, 2016, pp. 5115–5119.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[4] A. Nagraniy, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech*, 2017, pp. 2616–2620.

[5] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech*, 2018, pp. 1086–1090.

[6] S. O. Sadjadi and J. H. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Interspeech*, 2010, pp. 2138–2142.

[7] J. Ortega-García and J. González-Rodríguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP)*, 1996, pp. 929–932.

[8] O. Plchot, L. Burget, H. Aronowitz, and P. Matějka, "Audio enhancing with DNN autoencoder for speaker recognition," in *ICASSP*, vol. 2016-May, 2016, pp. 5090–5094.

[9] O. Novotny, O. Plchot, P. Matějka, and O. Glembek, "On the use of DNN Autoencoder for Robust Speaker Recognition," *ArXiv e-prints arXiv:1811.02938*, 2018.

[10] O. Novotny, O. Plchot, O. Glembek, J. H. Cernock, and L. Burget, "Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition," *ArXiv e-prints arXiv:1811.07629*, 2018.

[11] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Interspeech*, 2017, pp. 2008–2012.

[12] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," *ArXiv e-prints arXiv:1810.04826*, 2018.

[13] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral Feature Mapping With Mimic Loss," in *ICASSP*, 2018, pp. 5609–5613.

[14] D. Snyder, G. Chen, and D. Povey, "MUSAN : A Music , Speech , and Noise Corpus," *ArXiv e-prints arXiv:1510.08484*, 2015.

[15] S. Shon, H. Tang, and J. Glass, "Frame-level Speaker Embeddings for Text-independent Speaker Recognition and Analysis of End-to-end Model," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1007–1013.

[16] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, March 1989.

[17] H. Tang, W.-N. Hsu, F. Grondin, and J. Glass, "A study of enhancement, augmentation, and autoencoder methods for domain adasptation in distant speech recognition," in *Interspeech*, 2018, pp. 2928–2932.

[18] W. Cai, J. Chen, and M. Li, "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 74–81.

[19] M. Hajibabaei and D. Dai, "Unified Hypersphere Embedding for Speaker Recognition," *ArXiv e-prints arXiv:1807.08312*, 2018.

[20] 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.

[21] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, R. Dehak, L. P. G. Perera, D. Povey, P. Torres-Carrasquillo, S. Khudanpur, and N. Dehak, "State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18," in *Interspeech*, 2019.

[22] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level Aggregation For Speaker Recognition In The Wild," in *ICASSP*, 2019.

[23] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Interspeech*, 2018, pp. 2252–2256.