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

Speaker Re-identification with Speaker Dependent Speech Enhancement

Yanpei Shi, Qiang Huang, Thomas Hain

Speech and Hearing Research Group
Department of Computer Science, University of Sheffield, UK
{YShi30, qiang.huang, t.hain}@sheffield.ac.uk

Abstract

While the use of deep neural networks has significantly boosted speaker recognition performance, it is still challenging to separate speakers in poor acoustic environments. Here speech enhancement methods have traditionally allowed improved performance. The recent works have shown that adapting speech enhancement can lead to further gains. This paper introduces a novel approach that cascades speech enhancement and speaker recognition. In the first step, a speaker embedding vector is generated , which is used in the second step to enhance the speech quality and re-identify the speakers. Models are trained in an integrated framework with joint optimisation. The proposed approach is evaluated using the Voxceleb1 dataset, which aims to assess speaker recognition in real world situations. In addition three types of noise at different signal-noise-ratios were added for this work. The obtained results show that the proposed approach using speaker dependent speech enhancement can yield better speaker recognition and speech enhancement performances than two baselines in various noise conditions.
Index Terms: Speech Enhancement, Speaker Identification, Speaker Verification, Noise Robustness.

1. Introduction

The aim of speaker recognition is to recognize speaker identities from their voice characteristics [1]. In recent years, the use of deep learning technologies [2, 3, 4, 5] has significantly improved speaker recognition performance. However, speaker recognition in poor noise conditions is still a challenging task as some important acoustic information related to the speaker is often interfered. To tackle speech signals corrupted by noise, some methods have been developed. Previous studies [6, 7, 8] tended to recover original signals by removing noise. Other methods [9, 10, 11] focused on feature extraction from uncorrupted speech signals, and further methods [12, 13] tried to estimated speech quality by computing signal-to-noise ratios (SNRs).

In many previous studies, speech enhancement is often processed individually [14, 15, 16, 17]. However, the learned features or enhanced speech signals might not be able to match well to the information required by speaker recognition and verification. It is highly desirable that both the speech enhancement and the speaker processing models can work together and can be optimized jointly. In [18], Shon et al. tried to integrate speech enhancement module and speaker processing module into one framework. In this method, a speech enhancement module filters out the noise by generating a ratio mask, and then multiplying it by the original spectrogram. However, in [18], the speaker verification module was pertrained and fixed when training the speech enhancement. The two modules are not optimized jointly.

To improve speaker identification and verification performance, our proposed approach proposes a speech enhancement module cascaded to a speaker recognition module in order to reduce the impact caused by noise interference. The two modules are optimized jointly by computing the enhancement loss and identification loss simultaneously.

To our knowledge, although speaker information has been widely used for acoustic model adaptation in speech recognition [19, 20], it is still under-developed in speech enhancement. To further improve the robustness of speaker recognition against noise, two steps are taken in this work. The first step is to learn a speaker embedding vector, which will be then used as a prior knowledge to enhance speech quality in the second step. The details of the proposed approach will be described in the following sections.

The rest of the paper is organized as follow: Section 2 presents the model architecture of the proposed approach and how it is implemented in order to identify a speaker and enhance the speech quality simultaneously. The used data set and experiment set-up are introduced in Section 3. The obtained results and related analysis are given in Section 4, and finally conclusions are drawn in Section 5.

2. Speaker Re-Identification

2.1. Model Structure

Figure 1 shows the architecture of the proposed approach, consisting of two steps (Step1 and Step2). Each step contains two modules, a speech enhancement (*SE*) module and a speaker recognition (*SR*) module. Given an input spectrogram $\boldsymbol{X_N}$, the goal for Step1 is to generate a speaker embedding (X_{e1}) using the speech enhancement module (*SE*1) and speaker recognition module (*SR*1). In Step2, the speaker embedding X_{e1} is used as the prior knowledge to improve the speaker recognition and speech enhancement performances. The architecture of the speech enhancement module (*SE*2) and speaker recognition module (*SR*2) have similar architecture to the *SE*1 and *SR*1 modules in Step1. The only difference is *SE*2 takes X_{e1} into account.

2.2. Module of Speech Enhancement and Speaker Recognition

Figure 2 shows the structure of the speech enhancement (*SE*) module. It is based on the structure of a residual auto-encoder [21, 15, 6], where several convolutional layers are stacked. It can be viewed as one stack of several singlelayer auto-encoders. The residual connection could improve the quality of the reconstructed spectrogram and avoid the vanishing gradient problem [21]. The use of a bi-directional GRU layer inserted between the encoder and decoder is to improve performance of speech enhancement [22], as it takes context information into account. The speaker embedding is only used in SE2.

http://dx.doi.org/10.21437/Interspeech.2020-1772

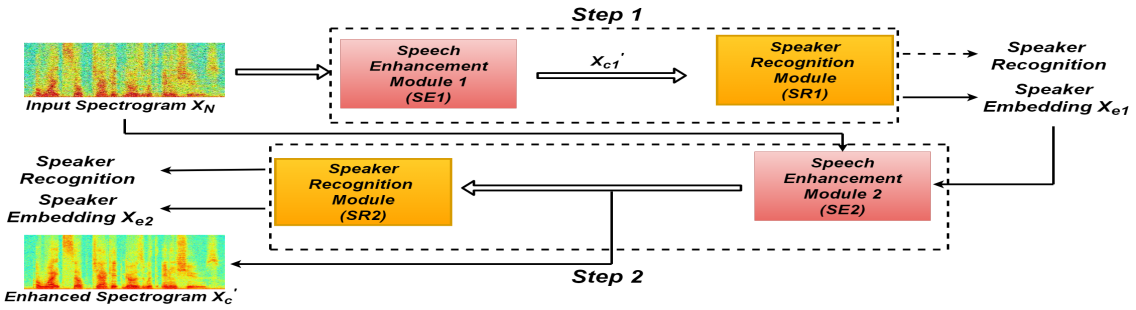


Figure 1: *Architecture of the proposed approach consisting of two steps (Step1 and Step2), each of which contains two modules: a speech enhancement (SE) module and a speaker recognition (SR) module. The input is a spectrogram corrupted by noise. A speaker identity and an enhanced spectrogram are the output.*

Figure 3 shows the structure of the used *SR* module, which is built on a Resnet-20 [23] structure. The following two fully-connected (FC) layers are used for speaker classification, and the output of the second to last FC layer is defined as a speaker embedding.

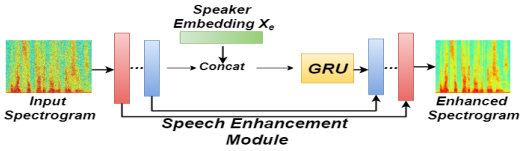


Figure 2: *Structure of the speech enhancement (SE) model is built on a residual/skip auto-encoder network and used in both Step1 and Step2. The speaker embedding is used only in SE2.*

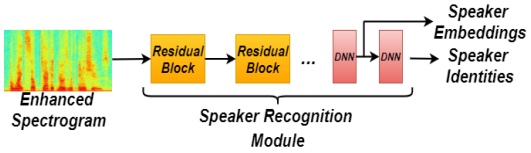


Figure 3: *Structure of speaker recognition (SR) module is built on a Resnet-20 network. SR1 aims to generate a speaker embedding, and SR2 is in charge of recognizing speaker ID.*

2.3. Speaker Embeddings (Step1)

As shown in Figure 1, *SE*1 and *SR*1 are cascaded. The noise corrupted spectrogram $\boldsymbol{X_N}$ is denoised using *SE*1, by which a speaker embedding is then yielded by the first fully connected layer of *SR*1.

For the *SE*1 module, mean absolute error (MAE) [24] is used to measure the difference between an input spectrogram $\boldsymbol{X_C}$ and an enhanced spectrogram $\boldsymbol{X'_{C1}}$, as it is more efficient compared to mean squared error (MSE) [21]. The loss function of this module is defined as:

$$\mathcal{L}_{SE} = \frac{1}{TF}||\boldsymbol{X_C} - \boldsymbol{X'_{C1}}||_1 = \frac{1}{TF}\sum_{i=0}^{T}\sum_{j=0}^{F}|x_{ij} - x'_{ij}| \quad (1)$$

where x_{ij} and x'_{ij} denotes each element in the clean and denoised spectrogram. T and F denote the dimension on time and frequency axes, respectively.

For SR1 module, the classifier is trained in terms of the difference between predictions $\hat{y}$ and corresponding targets y and uses the categorical cross entropy as its loss function:

$$\mathcal{L}_{SR} = -\sum_{i=0}^{N}\sum_{j=0}^{M} y_{ij} * \log \hat{y}_{ij} \quad (2)$$

where N denotes the number of samples and M denotes the number of classes (speakers).

The *SE*1 and the *SR*1 modules are firstly trained independently using the loss function introduced above and then fine-tuned together using Eq 3.

2.4. Speaker-Dependent Speech Enhancement (Step2)

In Step2, both the speech enhancement and speaker recognition modules use similar structures to the modules in Step1. However, unlike *SE*1, the *SE*2 module concatenates the speaker embedding vector $\boldsymbol{X_{e1}}$, with its own bottleneck vector and enhances the quality of $\boldsymbol{X_N}$.

In this work, the optimization of Step1 and Step2 are independent to each other. The parameters of *SE*1 and *SR*1 used in Step1 are fixed when training *SE*2. *SR*2 shares weights with *SR*1. Unlike Step1, a joint optimization is implemented on the two modules in Step2 by using Eq 1 and 2 simultaneously:

$$\mathcal{L} = \mathcal{L}_{SE} + \mathcal{L}_{SR} \quad (3)$$

3. Experiments

3.1. Data

All experiments were run on the Voxceleb1 dataset [25], which is a widely used large dataset for speaker identification and verification. The Voxceleb1 dataset contains 1251 speakers and more than 150 thousand "wild" utterances, extracted from YouTube videos.

In all experiments, spectrograms are used as the input acoustic features. Input speech streams were firstly segmented using a 25ms sliding window with a 10ms step. A 512-point Fast Fourier Transform (FFT) was then conducted on each segment which yielded a 257-D vector (a DC component is concatenated). The length of spectrogram covers 300 frame, about 3 seconds. No normalization techniques were used to preprocess the spectrograms.

To evaluate the robustness of the proposed model, additional noises from the MUSAN dataset were added. The MUSAN dataset contains three categories of noises: general noise, music and babble [26]. The general noise type contains 6 hours

of audio, including dialtones and fax machine noises etc. The music data contains 42 hours of music recordings from different categories, and the babble data contains 60 hours of speech, including read speech from public domain, hearings, committees and debates etc.

3.2. Experiment Setup

To evaluate the effectiveness of the proposed approach, two tasks, speaker identification and verification, were designed and tested on the Voxceleb1 dataset using the official train/test split [25].

For speaker identification, there are 1251 speakers in both training and test set [25]. Each utterance is randomly mixed with a type of noise at one of five SNR levels (from 0 to 20 dB). To evaluate the recognition performance, The Top-1 and Top-5 accuracies were computed [23].

For speaker verification, the same data configuration as the speaker identification task was set. A cosine score between two vectors was computed and used to measure the similarity [18]. Equal Error Rate (EER) [27] and Detection Cost Function (DCF) [28] were used as evaluation metrics. DCF represents the average of two minimum DCF score (DCF0.01 and DCF0.001) [28, 29].

To evaluate our proposed approaches, two baselines and two proposed approaches were tested.

SID : represents the baseline method using only a speaker recognition module (*SR*1) without any pre-processing and post-processing.

VoiceID-Loss [18]: represents a baseline from [18], where the speech enhancement and speaker recognition modules are cascaded, but without a joint training and the use of speaker embeddings.

SESR-Step1 : represents the proposed model where the *SE*1 and *SR*1 modules are jointly trained, but speaker embedding vectors are not being used.

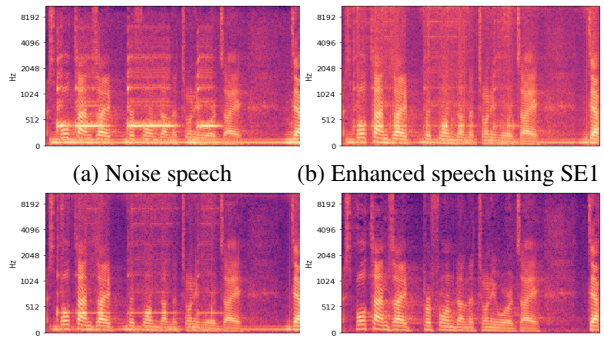
SESR2-Step2 : represents the model where the *SE*2 and *SR*2 modules are jointly trained with the learned speaker embedding vector being used in *SE*2. The loss function, defined by Eq 3), is employed for model optimization.

To verify the effectiveness of the proposed approach in speech enhancement, two metrics, Perceptual Evaluation of Speech Quality (PESQ) [30] and Short-Time Objective Intelligibility (STOI) [31], are used to assess the enhanced speech quality.

3.3. Network Structure

Table 1: *The encoder architecture of the proposed speech enhancement approach, where T, F, C represents the time, frequency and feature dimensions. The number of features and strides on each dimension are shown as Feature/Strides*

Operation	Structure	Input (T, F, C)	Output (T, F, C)
Encoder	16/(1,2)	(300,257,1)	(300,129,16)
	32/(2,2)	(300,129,16)	(150,65,32)
	64/(2,2)	(150,65,32)	(75,33,64)
	128/(2,2)	(75,33,64)	(38,17,128)
	256/(2,4)	(38,17,128)	(19,5,256)
Reshape	-	(19,5,256)	(19,1280)
Concatenation	-	(19,1280)	(19,1536)
DNN	512	(19,1536)	(19,512)
Bi-GRU	640	(19,512)	(19,1280)
Reshape	-	(19,1280)	(19,5,256)



(a) Noise speech (b) Enhanced speech using SE1

(c) Enhanced speech using SE2 (d) Original speech

Figure 4: *Comparison of spectrograms (a) speech corrupted by noise; (b) enhanced spectrogram obtained using the SE1 module; (c) enhanced spectrogram obtained using the SE2 module; (d) original speech.*

Table 1 shows the encoder architecture of the skip/residual auto-encoder employed by the speech enhancement module used in Step1 and Step2. Its decoder structure mirrors the encoder. For the speaker recognition module, the structure of Resnet-20 is used and the details can be found in [32].

For model optimization, The Adam optimizer [33] is used with the initial learning rate being set to 1e-3 and the decay rate being set to 0.9 for each epoch.

4. Results and Analysis

Table 2 shows the speaker identification performances obtained using two baselines (SID and VoiceID-Loss) and two proposed approaches (SESR-Step1, SESR-Step2). It is clear that the two proposed approaches, SESR-Step1 and SESR-Step2 can yield better performances than other baselines in various noise conditions, even if the SNR is 0dB. Moreover, after using speaker information learned by the SR1 module, the proposed approach SESR-Step2 can further improve the identification performance in comparison with SESR-Step1. This case is probably related to two factors. The first factor is the use of speech enhancement before speaker identification and a joint optimization, by which some noise interferences might be filtered out. The second factor is the implementation of the speaker dependent speech enhancement in Step2. Unlike speaker-independent speech enhancement, the use of speaker information can not only recover the noise corrupted speech signals to some extent, but also possible highlights speaker-specific features, which might be key to speaker recognition.

Table 3 shows the speaker verification performances obtained using the four methods. Similar to Table 2, the use of SESR-Step2 can achieve the best results in most conditions. When evaluating the verification performance, any further post-process, such as Probabilistic Linear Discriminant Analysis (PLDA) [34], was not employed, and only a cosine score was used to compare the similarities between enrolment and test data. This might be the reason that the improvement of using SESR-Step2 over SESR-Step1 on the speaker verification task is relative slight.

Table 4 and 5 show the speech enhancement performances evaluated using PESQ and STOI respectively. The second column in both tables show the quality of input speech corrupted by music noise at five different SNR levels. The third column indicates the obtained speech quality after using VoiceID-loss.

Table 2: *Comparison of speaker identification performances obtained using four different methods in various noise conditions.*

Noise Type	SNR	SID		VoidID-Loss[18]		SESR-Step1		SESR-Step2	
		Top1 (%)	Top5 (%)	Top1 (%)	Top5 (%)	Top1 (%)	Top5 (%)	Top1 (%)	Top5 (%)
Noise	0	74.1	86.9	75.6	88.0	76.4	88.9	77.4	89.4
	5	79.2	90.0	80.4	90.8	81.8	91.2	83.6	91.6
	10	83.2	93.2	84.7	94.3	85.4	94.7	87.1	95.7
	15	84.9	94.6	85.6	95.1	86.3	95.8	88.7	95.9
	20	87.9	95.4	88.7	96.0	89.5	96.4	90.3	96.8
Music	0	65.8	82.0	67.1	83.3	69.2	85.2	70.4	85.8
	5	76.9	89.1	78.2	89.9	80.1	90.6	81.3	90.6
	10	83.8	93.5	84.6	94.2	85.9	95.1	85.9	95.5
	15	86.1	93.9	87.3	95.0	88.4	95.7	89.0	96.3
	20	87.4	94.7	88.9	95.6	89.2	96.6	90.4	97.0
Babble	0	62.4	80.2	63.8	82.1	65.7	83.5	66.6	83.9
	5	76.2	87.3	77.6	88.7	79.4	89.1	80.1	90.3
	10	81.4	92.2	82.3	93.5	84.0	94.9	84.8	94.5
	15	84.0	92.6	86.1	94.0	87.2	95.2	89.1	95.7
	20	85.8	92.9	86.6	95.1	88.4	95.7	90.3	96.2
Original		88.5	95.9	89.7	96.4	90.2	96.8	91.1	97.7

Table 3: *Comparison of speaker verification performances obtained using four different methods in various noise conditions.*

Noise Type	SNR	SID		VoidID-Loss[18]		SESR-Step1		SESR-Step2	
		EER (%)	DCF	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
Noise	0	16.94	0.993	16.56	0.938	16.02	0.885	15.89	0.886
	5	12.48	0.855	12.26	0.830	11.87	0.794	11.83	0.786
	10	10.03	0.760	9.86	0.747	9.21	0.695	9.17	0.695
	15	8.84	0.648	8.69	0.686	8.18	0.625	7.99	0.616
	20	7.96	0.594	7.83	0.639	7.06	0.590	6.85	0.589
Music	0	17.04	0.940	16.24	0.913	15.69	0.893	15.70	0.904
	5	11.54	0.828	11.44	0.818	10.88	0.754	10.78	0.770
	10	9.69	0.749	9.13	0.733	8.76	0.690	8.94	0.704
	15	8.40	0.689	8.10	0.677	7.81	0.631	7.65	0.621
	20	7.70	0.665	7.48	0.635	7.09	0.606	7.03	0.592
Babble	0	38.90	1.000	37.96	1.000	37.18	0.999	37.52	0.994
	5	28.04	0.998	27.12	0.996	26.84	0.991	26.69	0.991
	10	17.34	0.917	16.66	0.926	16.38	0.878	16.93	0.901
	15	11.31	0.795	11.25	0.807	10.87	0.781	10.84	0.780
	20	9.12	0.720	8.99	0.705	8.76	0.679	8.72	0.685
Original		6.92	0.565	6.79	0.574	6.52	0.548	6.48	0.537

Table 4: *Comparison of the PESQ values obtained using the proposed approaches and baselines in various music noise conditions.*

SNR	Noisy	VoiceID-loss [18]	SESR-Step1	SESR-Step2
0	1.53	1.48	1.62	1.90
5	1.78	1.72	1.89	2.14
10	1.86	1.83	1.97	2.35
15	2.16	2.06	2.21	2.58
20	2.39	2.20	2.53	2.89

Table 5: *Comparison of the STOI values obtained using the proposed approaches and baselines in various music noise conditions.*

SNR	Noisy	VoiceID-loss [18]	SESR-Step1	SESR-Step2
0	0.53	0.50	0.56	0.63
5	0.60	0.58	0.64	0.71
10	0.65	0.61	0.67	0.75
15	0.67	0.62	0.69	0.77
20	0.68	0.64	0.70	0.78

The use of proposed approach, SESR-Step2, shows clear advantages over VoiceID-loss and SESR-Step1 in various noise conditions.

To further verify the robustness of the proposed approach against noise, Figure 4 shows fours spectrograms: The noise corrupted speech by music noise at 0 dB; The enhanced speech obtained using SESR-Step1; The enhanced speech obtained using SESR-Step2; The original speech. It can be found that the music noise can be removed to a certain extent from the spectrograms shown in Figure (b) and (c) after using speech enhancement, and the spectrogram shown in figure (c) is closer to the original spectrogram shown in figure (d).

5. Conclusion and Future Work

In this paper, a novel speaker-dependent speech enhancement for speaker recognition approach is presented and tested on Voxceleb1. The obtained results show that the use of the proposed approach can yield better performances in speaker recognition and enhance speech quality in various noise conditions.

To further improve speaker identification performance and its robustness against noise, more advanced deep learning technologies, such as capsule networks and the vector quantizer variable auto-encoder (VQVAE) will tested in this framework. In addition, some post-process methods, such as PLDA, will be also taken into account.

Acknowledgement

This work was in part supported by Innovate UK Grant number 104264.

6. References

[1] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, 2017.

[2] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *ICASSP*. IEEE, 2014.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*. IEEE, 2018.

[4] F. R. rahman Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, "Attention-based models for text-dependent speaker verification," in *ICASSP*. IEEE, 2018.

[5] N. N. An, N. Q. Thanh, and Y. Liu, "Deep cnns with self-attention for speaker identification," *IEEE Access*, 2019.

[6] S. Leglaive, U. Şimşekli, A. Liutkus, L. Girin, and R. Horaud, "Speech enhancement with variational autoencoders and alpha-stable distributions," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 541–545.

[7] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoder," *arXiv preprint arXiv:1908.02590*, 2019.

[8] 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.

[9] I. Jang, C. Ahn, J. Seo, and Y. Jang, "Enhanced feature extraction for speech detection in media audio." in *INTERSPEECH*, 2017, pp. 479–483.

[10] G. Farahani, S. M. Ahadi, and M. M. Homayounpour, "Robust feature extraction of speech via noise reduction in autocorrelation domain," in *International Workshop on Multimedia Content Representation, Classification and Security*. Springer, 2006, pp. 466–473.

[11] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[12] L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, "An adaptive a priori snr estimator for perceptual speech enhancement," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, p. 7, 2019.

[13] R. Yao, Z. Zeng, and P. Zhu, "A priori snr estimation and noise estimation for speech enhancement," *EURASIP journal on advances in signal processing*, vol. 2016, no. 1, p. 101, 2016.

[14] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5756–5760.

[15] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *Proc. Interspeech 2017*, pp. 3642–3646, 2017.

[16] P. Muhammed Shifas, N. Adiga, V. Tsiaras, and Y. Stylianou, "A non-causal fftnet architecture for speech enhancement," *Proc. Interspeech 2019*, pp. 1826–1830, 2019.

[17] J. Yao and A. Al-Dahle, "Coarse-to-fine optimization for speech enhancement," *Proc. Interspeech 2019*, pp. 2743–2747, 2019.

[18] S. Shon, H. Tang, and J. Glass, "Voiceid loss: Speech enhancement for speaker verification," *Proc. Interspeech 2019*, pp. 2888–2892, 2019.

[19] Z. Meng, Y. Gaur, J. Li, and Y. Gong, "Speaker adaptation for attention-based end-to-end speech recognition," *arXiv preprint arXiv:1911.03762*, 2019.

[20] Y. Zhao, J. Li, S. Zhang, L. Chen, and Y. Gong, "Domain and speaker adaptation for cortana speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5984–5988.

[21] A. Pandey and D. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[22] J.-Y. Liu and Y.-H. Yang, "Denoising auto-encoder with recurrent skip connections and residual regression for music source separation," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 773–778.

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

[24] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

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

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

[27] J.-M. Cheng and H.-C. Wang, "A method of estimating the equal error rate for automatic speaker verification," in *2004 International Symposium on Chinese Spoken Language Processing*. IEEE, 2004, pp. 285–288.

[28] D. A. Van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker classification I*. Springer, 2007, pp. 330–353.

[29] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[32] 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.

[33] D. P. Kingma and J. L. Ba, "Adam: Amethod for stochastic optimization."

[34] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7649–7653.