Analysis of PingAn Submission in the VoicePrivacy 2020 Challenge

Chien-Lin Huang

PAII Inc., Palo Alto, CA, USA

chiccocl@gmail.com

Abstract

This paper shows the post-evaluation analysis of our efforts in the VoicePrivacy 2020 Challenge. The VoicePrivacy 2020 Challenge focuses on the task of speech anonymization which is related to speech synthesis, voice conversion, automatic speech recognition, and speaker verification. In this study, we focus on speaker verification. Based on speaker embedding x-vectors, we study different front-end feature extraction, data augmentations, and neural network topologies. Score fusion is used to combine different system results. Our systems are compared with two official anonymization baselines and report objective evaluation results. We can greatly reduce the Equal Error Rate (*EER*) by using the proposed methods.

Index Terms: speaker verification, privacy, anonymization, x-vectors, speaker embedding

1. Introduction

Speaker verification is one of biometric authentication methods such as iris scanning, facial recognition and fingerprinting sensing [1]–[3]. Because speech based human machine interaction has become popular in smart home, mobile devices and automobiles, speaker verification indicates its important role in machine learning and artificial intelligent. Speakers can be verified or identified by using a small amount of their voice [2]. Speaker verification can be classified into text-dependent and text-independent tasks according to the applications. The neural network based speaker embedding methods demonstrate good performance compared the traditional approaches. However, techniques of noise addition, speech transformation, voice conversion and speech synthesis make speaker verification hard. The word "anonymization" means suppressing personally identifiable attributes of the speech signal and leaving all other attributes intact such as noise addition [4], speech transformation [5], voice conversion [6]–[8], speech synthesis [9, 10], adversarial learning [11], and so on. The VoicePrivacy 2020 Challenge considers the following scenario, where the words "user" and "speaker" are used interchangeably [12, 13]. The speaker want to hide their identity while still allowing all other downstream goals to be achieved. But, the attacker want to identify the speaker. It is a trade-off situation between anonymization task and attack models. In this study, we focus on the analysis of speaker verification based on neural network based speaker embeddings. Recently, researchers are working on training neural network speaker embedding methods. Different methods of data augmentations, neural network topologies, and loss functions are proposed [14]–[22]. We show the impact of different front-end feature analysis, training data, data augmentation, and back-end scoring. The main objective of this study is to provide a description and analysis of our submission to the VoicePrivay 2020 challenge.

This paper is organized as follows. Section 2 introduces our system setup including dataset, feature analysis and speaker embedding neural network topologies. Section 3 describes the experimental results and analysis. Finally, Section 4 concludes this work.

2. System Setup

We explore different data augmentations, feature extraction, and speaker embedding neural network topologies.

2.1. Training, development, and evaluation dataset

In the VoicePrivacy 2020, a fixed training condition is required which means systems can only be trained using a designated training set including VoxCeleb1 [23], VoxCeleb2 [24], LibriSpeech [25], and LibriTTS [26]. A 600 hours subsets of the LibriSpeech and LibriTTS are used including train-clean-100 and train-other-500. We only use the VoxCeleb dataset to train speaker embedding neural networks. The VoxCeleb dataset is collected by the University of Oxford, UK and extracted from videos uploaded to YouTube. The overall dataset involves two parts of VoxCeleb1 and Voxceleb2 which contains over 2,000 hours, over one million speech utterances for over 7,000 celebrities. The average number of utterances per speaker is about 170. All the audio samples are 16 kHz and 16-bit format wideband speech.

The development set comprises LibriSpeech dev-clean and a subset of the CVTK corpus [27] called VCTK-dev. With the attach models in mind, data are divided into trial and enrollment set. The speakers in the enrollment set are the subset of those in the trial set for LibriSpeech dev-clean. However, the same speakers are used for enrollment and trial for VCTK-dev. In addition, there are two trial subsets denoted as common and different. The common trial subset is composed of utterances #1–24 in the VCTK corpus that are identical for all speakers which mean the subjective evaluation of speaker verifiability/linkability in a text-dependent manner. In other words, the different trial subset is a text-independent speaker verification. Similarly, the evaluation set comprises LibriSpeech test-clean and a subset of VCTK denoted as VCTK-test.

2.2. Front-end feature analysis

Different types of front-end feature extraction are used to analyze speech from different signal aspects. Three speech feature sets are extracted from audio files, including the Mel-frequency cepstral coefficients (MFC), perceptual linear predictive (PLP) analysis of speech, and Mel-frequency cepstral coefficients with pitch (MFP). The bandwidth is limited between 20 Hz and 7600 Hz. Features are extracted from a 25 millisecond (ms) frame length and a 10ms frame-shift.

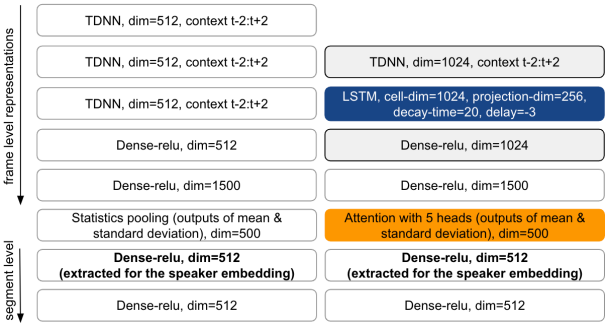

Figure 1: TDNN based x-vector which is modified by changing with long short-term memory, self-attention pooling, and bigger neurons in the frame level.

2.3. Data augmentations

Because the neural network based speaker embedding is a data greedy approach, data augmentation is used to increase the amount and diversity of the available training [28]–[31]. Not only create new data but also create new speaker, in this study, we propose a vocal tract length perturbation (VTLP) to create new speakers and explore the effects of adding different data augmentation methods for the short-duration speaker verification.

In speech signals, the female speaker's speech tends to have shorter vocal tract lengths and higher formant frequencies than male speakers [32]. One would expect to see more compressed spectra in female speech than in male speech [33]. In Mel-frequency cepstral coefficients, the frequency bins are computed with the perceptually motivated Mel-frequency scaling after the log-amplitude of the magnitude spectrum. To change different vocal tract lengths and create new speakers, the VTLP based the speaker-specific *Mel* scales are estimated as follows:

$$Mel(f) = 2959\log_{10}\left(1 + \frac{f}{700 \times \alpha}\right) \quad (1)$$

where the warping factor α is used to adjust a speaker-specific *Mel* scale. This frequency-warping procedure is implemented as a filter bank modification. $\alpha > 1.0$ results in a compressed spectrum, $\alpha < 1.0$ results in a stretched spectrum, and $\alpha = 1.0$ is for a non-warped spectrum. We use the spectral warping factor $\alpha = 0.9$ with 9ms frame-shift (the default is 10ms frame-shift) and the spectral warping factor $\alpha = 1.1$ with 11ms frame-shift to create two more copies of the original data (speakers) for training speaker embedding neural network, LDA/PLDA.

Besides of VTLP creating new speakers and data, we apply the MUSAN dataset [34] to corrupt the original audio files with additive noises, including babble noise, general noise, and music noise. The simulated room impulse responses (RIRs) is used to corrupt the original audio by convolving with simulated RIRs. The simulated room impulse responses include small and medium room size from the ranges of width and length of a room are uniformly sampled from 1m-10m and 10m-30m, respectively. In addition, we adopt the volume perturbation by using the sampled scaling factors. The scaling factor for each utterance is randomly chosen from a range (e.g. [0.5, 1.5]). Finally, the speed perturbation is used to create two more copies of the original signal with speed factors of 0.9 and 1.1 which is related to modify the speed to 90% and 110% of the original rate.

2.4. Voice activity detection

We used the non-parameter approach of energy-based voice activity detection (VAD) to estimate frame-by-frame speech activity. Without modeling, such as Gaussian mixture model (GMM) classifiers, the frames with silence or low signal-to-noise ratio in the audio samples are removed.

2.5. TDNN-LSTM-Attention speaker embedding

The neural network based speaker embedding technologies demonstrate sound performance and become the mainstream methods in speaker recognition. Variable-length utterances are converted to fixed-dimensional embedding vectors. TDNN-LSTM-Attention neural network topology is proposed by considering long short-term memory (LSTM), time-delay neural network (TDNN), and self-attention pooling as shown in Fig. 1. The long short-term memory recurrent neural networks (RNN) is applied to better capture the temporal information in speech than using TDNN alone as in x-vector [35]–[38]. The bigger hidden neurons (1,024 instead of 512) are considered in training speaker neural networks. The temporal average pooling layer in x-vector is replaced with an attention pooling layer is applied to automatically determine weights of the speaker's frame-level hidden vectors by an attention mechanism [39]–[42]. The self-attention pooling layer with 5 heads is used in this study. Mean and standard deviation from the variable-length inputs are estimated in the pooling layer. After the pooling layer, the speaker embedding representation is extracted from the first segment-level layers. To avoid swapping data in training neural networks, we create the training archives and shuffled training data. The number of maximum and minimum frames in each training example is 400 and 200, respectively. The number of repeats for each speaker is 50. The number of frames per iteration is one billion. According to the configuration, 309 archives are generated in training neural networks. The TDNN-LSTM-Attention is trained using the Stochastic Gradient Descent (SGD) optimizer. SGD with weight decay=1e-8 and momentum=0.5 is used for six epochs.

2.6. Backend LDA-PLDA scoring

The PLDA based classifier is used in speaker embedding scoring. Before scoring, the vectors of speaker embedding are centered, projected to 200 dimensionalities using LDA and applied length normalization. The LDA and PLDA are trained using the VoxCeleb1 and VoxCeleb2 datasets.

3. Results and Analysis

3.1. Evaluation metrics

The main metrics for the challenge are Equal Error Rate (*EER*) and the log-likelihood-ratio cost function (*Cllr* and *Cllr_min*). The *EER* means the point of the two detection error rates of false alarm and miss are equal. The log-likelihood-ratio (*LLR*) cost function, *Cllr*, is computed as follows:

$$Cllr = \left(\left(\frac{1}{N_{Target}}\right) + \sum_{i \in Target} \log_2\left(1 + e^{-LLR_i}\right)\right.$$
$$\left. + \left(\frac{1}{N_{NonTarget}}\right) + \sum_{j \in NonTarget} \log_2\left(1 + e^{-LLR_j}\right)\right)/2 \quad (2)$$

Table 1: Analysis of the results based on the Anonymization baseline-1 method of VoicePrivacy 2020 challenge. Official x-vector result and our submitted result are listed. Our result is bold face.

Dataset	Gender	Anonymization method 1		Development			Test		
		Enroll	Trial	*%EER*	*Cllr_min*	*Cllr*	*%EER*	*Cllr_min*	*Cllr*
LibriSpeech	female	original	original	8.67 (**2.42**)	0.30 (**0.10**)	42.86 (**1.01**)	7.66 (**0.73**)	0.18 (**0.02**)	42.86 (**0.34**)
			anonymized	50.28 (**39.20**)	1.00 (**0.95**)	146.01 (**20.90**)	48.54 (**39.42**)	1.00 (**0.94**)	146.01 (**18.25**)
		anonymized		35.09 (**33.38**)	0.88 (**0.85**)	15.19 (**36.06**)	29.74 (**29.38**)	0.80 (**0.80**)	15.19 (**48.84**)
	male	original	original	1.24 (**0.16**)	0.03 (**0.00**)	14.25 (**0.01**)	1.11 (**0.22**)	0.04 (**0.00**)	14.25 (**0.01**)
			anonymized	58.39 (**49.07**)	1.00 (**0.99**)	168.50 (**37.07**)	53.23 (**49.22**)	1.00 (**1.00**)	168.50 (**31.74**)
		anonymized		29.66 (**30.59**)	0.81 (**0.80**)	20.08 (**40.50**)	32.52 (**32.96**)	0.84 (**0.85**)	20.08 (**48.82**)
VCTK (different)	female	original	original	2.86 (**0.22**)	0.10 (**0.01**)	1.13 (**0.08**)	4.89 (**0.41**)	0.17 (**0.01**)	1.13 (**0.38**)
			anonymized	50.03 (**41.83**)	0.99 (**0.96**)	162.91 (**30.99**)	48.87 (**42.64**)	1.00 (**0.98**)	162.91 (**22.15**)
		anonymized		29.48 (**27.23**)	0.81 (**0.76**)	10.24 (**27.36**)	34.21 (**32.15**)	0.88 (**0.84**)	10.24 (**28.55**)
	male	original	original	1.44 (**0.05**)	0.05 (**0.00**)	1.16 (**0.01**)	2.07 (**0.17**)	0.07 (**0.01**)	1.16 (**0.04**)
			anonymized	55.33 (**43.13**)	1.00 (**0.97**)	166.50 (**33.63**)	53.73 (**47.24**)	1.00 (**1.00**)	166.50 (**40.29**)
		anonymized		26.10 (**28.83**)	0.76 (**0.81**)	18.81 (**40.07**)	25.83 (**29.56**)	0.74 (**0.81**)	18.81 (**40.67**)
VCTK (common)	female	original	original	2.62 (**0.29**)	0.09 (**0.01**)	0.87 (**0.17**)	2.89 (**0.29**)	0.09 (**0.00**)	0.87 (**0.19**)
			anonymized	49.42 (**38.37**)	1.00 (**0.94**)	165.44 (**27.19**)	50.00 (**40.17**)	1.00 (**0.95**)	157.81 (**21.68**)
		anonymized		25.29 (**20.35**)	0.74 (**0.63**)	7.96 (**32.54**)	30.92 (**24.28**)	0.83 (**0.71**)	9.49 (**34.39**)
	male	original	original	1.43 (**0.00**)	0.05 (**0.00**)	1.56 (**0.02**)	1.13 (**0.00**)	0.04 (**0.00**)	1.04 (**0.01**)
			anonymized	56.98 (**49.57**)	1.00 (**0.99**)	191.90 (**43.47**)	55.93 (**45.76**)	1.00 (**0.99**)	189.24 (**40.68**)
		anonymized		27.64 (**26.21**)	0.74 (**0.73**)	18.51 (**46.89**)	22.03 (**23.45**)	0.66 (**0.68**)	14.06 (**44.48**)

Table 2: Analysis of the results based on the Anonymization baseline-2 method of VoicePrivacy 2020 challenge. Official x-vector result and our submitted result are listed. Our result is bold face.

Dataset	Gender	Anonymization method 2		Development			Test		
		Enroll	Trial	*%EER*	*Cllr_min*	*Cllr*	*%EER*	*Cllr_min*	*Cllr*
LibriSpeech	female	original	original	8.81 (**2.42**)	0.31 (**0.10**)	42.90 (**1.01**)	7.66 (**0.73**)	0.18 (**0.02**)	26.81 (**0.34**)
			anonymized	35.37 (**12.78**)	0.82 (**0.45**)	116.89 (**15.66**)	26.09 (**9.31**)	0.69 (**0.29**)	115.57 (**13.59**)
		anonymized		23.44 (**8.67**)	0.62 (**0.29**)	11.73 (**12.68**)	15.33 (**7.66**)	0.49 (**0.22**)	12.55 (**14.02**)
	male	original	original	1.24 (**0.16**)	0.04 (**0.00**)	14.29 (**0.01**)	1.11 (**0.22**)	0.04 (**0.00**)	15.34 (**0.01**)
			anonymized	17.86 (**9.47**)	0.53 (**0.28**)	105.72 (**8.48**)	17.82 (**8.02**)	0.50 (**0.24**)	106.43 (**9.30**)
		anonymized		10.56 (**4.19**)	0.36 (**0.15**)	11.95 (**9.25**)	8.24 (**4.45**)	0.26 (**0.17**)	15.38 (**16.72**)
VCTK (different)	female	original	original	2.92 (**0.22**)	0.10 (**0.01**)	1.14 (**0.08**)	4.94 (**0.41**)	0.17 (**0.01**)	1.49 (**0.38**)
			anonymized	35.54 (**13.92**)	0.91 (**0.45**)	90.54 (**12.97**)	30.04 (**14.61**)	0.79 (**0.47**)	93.21 (**5.69**)
		anonymized		15.83 (**3.26**)	0.50 (**0.12**)	39.81 (**20.66**)	16.92 (**5.61**)	0.55 (**0.22**)	41.34 (**16.59**)
	male	original	original	1.44 (**0.15**)	0.05 (**0.00**)	1.16 (**0.01**)	2.067 (**0.17**)	0.07 (**0.01**)	1.82 (**0.04**)
			anonymized	28.24 (**3.67**)	0.74 (**0.13**)	98.42 (**9.76**)	28.24 (**5.17**)	0.72 (**0.19**)	101.70 (**11.49**)
		anonymized		11.22 (**2.88**)	0.38 (**0.10**)	23.09 (**15.07**)	12.23 (**2.81**)	0.40 (**0.09**)	25.06 (**12.95**)
VCTK (common)	female	original	original	2.62 (**0.29**)	0.09 (**0.01**)	0.87 (**0.17**)	2.89 (**0.29**)	0.09 (**0.00**)	0.86 (**0.19**)
			anonymized	34.30 (**11.63**)	0.88 (**0.36**)	85.90 (**11.59**)	30.64 (**15.90**)	0.81 (**0.48**)	93.97 (**9.13**)
		anonymized		11.63 (**2.33**)	0.37 (**0.10**)	43.56 (**20.43**)	14.45 (**2.89**)	0.47 (**0.10**)	42.73 (**20.96**)
	male	original	original	1.46 (**0.00**)	0.05 (**0.00**)	1.56 (**0.02**)	1.13 (**0.00**)	0.04 (**0.00**)	1.04 (**0.01**)
			anonymized	23.93 (**9.12**)	0.67 (**0.28**)	90.76 (**13.56**)	24.29 (**4.24**)	0.71 (**0.13**)	99.34 (**11.25**)
		anonymized		10.54 (**1.99**)	0.32 (**0.06**)	24.99 (**17.70**)	11.86 (**1.98**)	0.35 (**0.05**)	28.23 (**13.15**)

where N_{Target} and $N_{NonTarget}$ are the number of target and nontarget *LLR* scores in the evaluation set, respectively. The *Cllr_min* is estimated by the optimal calibration using monotomic transformation scores to their empirical *LLR* scores. To obtain the monotonoic transformation, the pool adjacent violators (PAV) to *LLR* method is used [43, 44].

3.2. Official baseline systems

Two anonymization baselines are provided officially including codes and corresponding objective results.

3.2.1. The baseline-1

The first baseline is anonymization using x-vectors and neural waveform models. There are three steps: In Step1, 256 dimensional bottleneck features encoding spoken content are extracted using automatic speech recognition (ASR) acoustic model trained on LibriSpeech train-clean-100 and train-other-500. A 512 dimensional x-vector encoding the speaker is extracted using a TDNN trained on VoxCeleb datasets. In Step2, for every source x-vector, an anonymized x-vector is estimated by finding the N=200 farthest x-vectors in an

external pool of LibriTTS train-other-500 based on the PLDA distance and averaging M=100 randomly selected vectors among them. In Step3, a speech synthesis acoustic model generates Mel-filterbank features given the F0, the anonymized x-vector, bottleneck features (BN), and a neural source filter (NSF) waveform model [45] produces a speech signal given the F0, the anonymized x-vector, and generated Mel-filterbank features. The speech synthesis acoustic model and NSF model are trained on LibriTTS train-clean-100.

3.2.2. The baseline-2

The second baseline is anonymization using McAdams coefficients [46]. In contrast to the baseline-1, there is no requirement of any training data for the baseline-2. The McAdams coefficient is used to achieve anonymization by shifting the pole positions derive from linear predictive coding (LPC) analysis of speech signals.

3.3. VoicePrivacy 2020 submission results

Our systems are compared with two official anonymization baselines and report objective evaluation results of the VoicePrivacy 2020 Challenge. The first baseline is anonymization using x-vectors and neural waveform models. The secondary baseline is anonymization using McAdams coefficients. Based on two official anonymization baselines, we focus on techniques of speaker verification. All results of baseline-1 and baselines-2 are shown in Table 1 and Table 2, respectively. Our result is bold face. Results of VCTK different and VCTK common denote tasks of text-independent and text-dependent speaker verification, respectively. Both anonymization methods of the trial data greatly increase the *EER* on all datasets. This shows that two anonymization baseline methods effectively increase the users' privacy. In addition, the anonymized enrollment data result in a lower *EER*, which suggests that F0+BN features retain information related to the speaker. If the attacker can have such enrollment data, they might be able to re-identify users. Compared with the first anonymization baseline approach, the secondary baseline is a simpler, formant-shifting approach. Therefore, the proposed speaker verification methods can greatly decrease the *EER* in the secondary baseline than the first baseline system, while interpretation of *Cllr* is more challenging due to non-calibration.

3.4. Greedy joint fusion

To select the best fusion combination, we use a greedy fusion scheme. First, we select the best one given the lowest *EER*. We fix that as the best system and evaluate all the two system fusions that include the best system. Thus, we select the best fusion of two systems. We fix two systems and then add a third system, and so on. To reduce the chances of overfitting, in each step, we prioritize fusions with only positive weights. Our results in Table 1 and Table 2 are fusion results by considering 4 systems including TDNN-LSTM speaker embedding using MFC feature, TDNN-LSTM-Attention speaker embedding using MFC, MFP, and PLP features. Experiments were implemented using the open-source Kaldi Speech Recognition Toolkit [47]. The experiments are tested on machines of NVIDIA DGX station equipped with Intel Xeon E5-2698 CPU 2.2 GHz, 256 GB RDIMM DDR4 and Tesla V100 GPUs. For training neural networks of speaker embeddings, it takes about 4-8 weeks depending on data augmentations and neural network topologies.

4. Conclusions

In this study, we proposed TDNN-LSTM-Attention based speaker embedding for the INTERSPEECH 2020 VoicePrivacy Challenge. Different types of front-end feature extraction are used to analyze speech from different signal aspects. Data augmentation is essential to boost the robustness of the speaker or acoustic model, and also to avoid overfitting during the training step. A vocal tract length perturbation (VTLP) is used to augment new data but also create new speakers for training speaker embedding neural networks, LDA, and PLDA. The proposed methods were trained on the VoxCeleb dataset including more than 2,000 hours of speech and 7,000 speakers, and evaluated on the LibriSpeech and VCTK datasets based on two official anonymization baselines.

5. References

[1] J. H. L. Hansen, and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[3] C.-L. Huang, H. Su, B. Ma, and H. Li, "Speaker characterization using long-term and temporal information," in *Proceedings of INTERSPEECH*, 2010.

[4] K. Hashimoto, J. Yamagishi, and I. Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5500–5504, 2016.

[5] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X.-Y. Li, Y. Wang, and Y. Deng, "Voicemask: Anonymize and sanitize voice input on mobile devices," *arXiv preprint arXiv:1711.11460,* 2017.

[6] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU),* pp. 529–533, 2009.

[7] M. Pobar and I. Ipši´c, "Online speaker de-identification using voice transformation," in *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* pp. 1264–1267, 2014.

[8] F. Bahmaninezhad, C. Zhang, and J. H. Hansen, "Convolutional neural network based speaker de-identification." in *Odyssey 2018: The Speaker and Language Recognition Workshop*, pp. 255–260, 2018.

[9] F. Fang, X.Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Speech Synthesis Workshop*, pp. 155–160, 2019.

[10] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, "Voice-indistinguishability: protecting voiceprint in privacy-preserving speech data release," *arXiv preprint arXiv:2004.07442*, 2020.

[11] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in ASR: Reality or illusion?" in *Proceedings of INTERSPEECH*, pp. 3700–3704, 2019.

[12] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans et al., "The VoicePrivacy 2020 challenge evaluation plan," 2020.

[13] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans et al., "Introducing the VoicePrivacy initiative," in *Proceedings of INTERSPEECH* 2020.

[14] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," in *Proceedings of INTERSPEECH,* 2019.

[15] L. Ferrer and M. McLaren, "Optimizing a speaker embedding extractor through backend-driven regularization," in *Proceedings of INTERSPEECH,* 2019.

[16] S. Shon, H. Tang, and J. Glass, "VoiceID loss: speech enhancement for speaker verification," in *Proceedings of INTERSPEECH,* 2019.

[17] Y. Zhu, T. Ko, and B. Mak, "Mixup learning strategies for text-independent speaker verification," in *Proceedings of INTERSPEECH,* 2019.

[18] S. Seo, D. J. Rim, M. Lim, D. Lee, H. Park, J. Oh, C. Kim, and J.-H. Kim, "Shortcut connections based deep speaker embeddings for end-to-end speaker verification system," in *Proceedings of INTERSPEECH,* 2019.

[19] A. Nautsch, J. Patino, A. Treiber, T. Stafylakis, P. Mizera, M. Todisco, T. Schneider, and N. Evans, "Privacy-preserving speaker recognition with cohort score normalisation," in *Proceedings of INTERSPEECH,* 2019.

[20] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," *IEEE Spoken Language Workshop (SLT),* 2016.

[21] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proceedings of INTERSPEECH,* 2017.

[22] 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. García-Perera, D. Povey, P. A. Torres-Carrasquillo, S. Khudanpur, and N. Dehak, "Deep neural network embeddings for text-independent speaker verification," in *Proceedings of INTERSPEECH,* 2019.

[23] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proceedings of INTERSPEECH*, 2017.

[24] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proceedings of INTERSPEECH*, 2018.

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

[26] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proceedings of INTERSPEECH*, pp. 1526–1530, 2019.

[27] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/3443

[28] Z. Wu, S. Wang, Y. Qian, K. Yu, "Data augmentation using variational autoencoder for Embedding based Speaker Verification," in *Proceedings of INTERSPEECH,* 2019.

[29] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," in *Proceedings of INTERSPEECH,* 2019.

[30] D. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of INTERSPEECH*, 2019.

[31] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of INTERSPEECH*, 2015.

[32] C.-L. Huang, C. Hori, H. Kashioka, and B. Ma, "Joint analysis of vocal tract length and temporal information for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[33] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, 1998.

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

[35] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[36] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," in *Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018.

[37] C.-L. Huang, "Exploring effective data augmentation with TDNN-LSTM neural network embedding for speaker recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[38] C.-P. Chen, S.-Y. Zhang, C.-T. Yeh, J.-C. Wang, T. Wang, and C.-L. Huang, "Speaker characterization using TDNN-LSTM based speaker embedding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[39] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Proceedings of INTERSPEECH*, 2018.

[40] T. Stafylakis, J. Rohdin, O. Plchot, and P. Mizera, L. Burget, "Self-supervised speaker embeddings," in *Proceedings of INTERSPEECH,* 2019.

[41] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," in *Proceedings of INTERSPEECH,* 2019.

[42] C.-L. Huang, "Speaker characterization using TDNN, TDNN-LSTM, TDNN-LSTM-Attention based speaker embeddings for NIST SRE 2019," in *Odyssey 2020: The Speaker and Language Recognition Workshop*, 2020.

[43] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[44] D. Ramos and J. Gonzalez-Rodriguez, "Cross-entropy analysis of the information in forensic speaker recognition," in *Odyssey 2008: The Speaker and Language Recognition Workshop*, 2008.

[45] X. Wang and J. Yamagishi, "Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis*,"* in *Speech Synthesis Workshop*, pp. 1–6, 2019.

[46] J. Patino, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," Eurecom, Tech. Report EURECOM+6190, 2020. [Online]. Available: http://www.eurecom.fr/publication/6190

[47] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.