10th ISCA Workshop on Speech Synthesis (SSW 10)
20-22 September 2019, Vienna, Austria

V2S attack: building DNN-based voice conversion from automatic speaker verification

Taiki Nakamura[1], Yuki Saito[2], Shinnosuke Takamichi[2], Yusuke Ijima[3], and Hiroshi Saruwatari[2]

[1] Faculty of Engineering, The University of Tokyo, Japan.
[2] Graduate School of Information Science and Technology, The University of Tokyo, Japan.
[3] Nippon Telegraph and Telephone Corporation, Japan.

supikiti22@gmail.com,
{yuuki_saito,shinnosuke_takamichi,hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp,
ijima.yusuke@lab.ntt.co.jp

Abstract

This paper presents a new voice impersonation attack using voice conversion (VC). Enrolling personal voices for automatic speaker verification (ASV) offers natural and flexible biometric authentication systems. Basically, the ASV systems do not include the users' voice data. However, if the ASV system is unexpectedly exposed and hacked by a malicious attacker, there is a risk that the attacker will use VC techniques to reproduce the enrolled user's voices. We name this the "verification-to-synthesis (V2S) attack" and propose VC training with the ASV and pre-trained automatic speech recognition (ASR) models and without the targeted speaker's voice data. The VC model reproduces the targeted speaker's individuality by deceiving the ASV model and restores phonetic property of an input voice by matching phonetic posteriorgrams predicted by the ASR model. The experimental evaluation compares converted voices between the proposed method that does not use the targeted speaker's voice data and the standard VC that uses the data. The experimental results demonstrate that the proposed method performs comparably to the existing VC methods that trained using a very small amount of parallel voice data.

Index Terms: automatic speaker verification, voice conversion, voice impersonation, automatic speech recognition, phonetic posteriorgrams

1. Introduction

Automatic speaker verification (ASV), which offers natural and flexible biometric authentication systems, has been actively studied in recent decades [1, 2]. Because the ASV systems identify the speaker of the input voice without using other biometrics, they are preferred for use in keyword spotting [3] and voice search implemented in smartphones. Among the ASV systems, text-independent ASV has the potential for highly portable speaker verification.

With deployments of ASV systems, we need to discuss the possibility of *voice impersonation attack* via the ASV systems. Specifically, if a malicious attacker exposes and hacks the ASV models, voices of the enrolled speakers risk being reproduced by the attacker. Voice conversion (VC) [4, 5, 6], which converts voices into the targeted speaker's ones, is a possible way for this type of attack. We call this attack *verification-to-synthesis (V2S) attack* that builds VC models from the pre-trained ASV model. Since ASV systems basically do not include voice data of the targeted speaker, we cannot perform the standard VC training using the targeted speaker's voice. However, since the ASV model learns the speaker's individuality, deceiving the model has some possibility of reproducing the targeted speaker's individuality by VC.

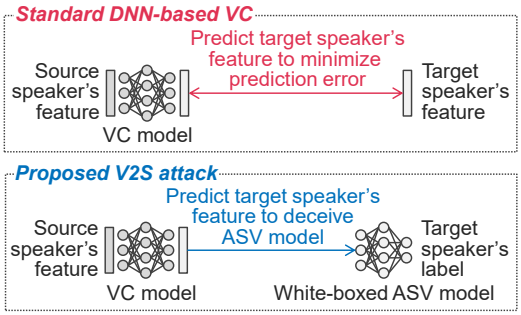

Figure 1: *Comparison of standard DNN-based voice conversion (VC) described in* **Section 2** *and proposed verification-to-synthesis (V2S) attack described in* **Section 3**. *ASV indicates automatic speaker verification.*

This paper proposes a V2S attack using a VC model trained with a ASV model. In this paper, we use a "white-boxed" ASV model, which means the attacker knows a deep neural network (DNN) architecture and the targeted speaker's label. Since the ASV model does not use phonetic property of the input voice, training the VC model using only the ASV model will lose phonetic property of the converted voice. Therefore, we further use the automatic speech recognition (ASR) model prepared by the attacker for restoring the phonetic property. The VC model is trained by not only deceiving the ASV model but also matching the output of the ASR model (i.e., phonetic posteriorgrams [7]) predicted from the input and converted voices. In the experimental evaluation, we evaluate the performance (i.e., naturalness and speaker individuality of the converted voice) of the proposed V2S attack with the existing VC methods (**Section 2**) because their performances are the upper limit of the proposed method. The experimental results demonstrate that the proposed method performs comparably to the existing VC methods trained using a very small amount of parallel voices.

This paper is organized as follows. **Section 2** briefly reviews conventional VC methods that require the targeted speaker's voice data. **Section 3** introduces the V2S attack that constructs the VC model with the ASV model and without the targeted speaker's voice. **Section 4** presents the experimental evaluations. **Section 5** concludes this paper with a summary.

10.21437/SSW.2019-29

2. Building voice conversion using targeted speaker's voice

This section describes standard VC techniques using a targeted speaker's voice: parallel and non-parallel VC (shown in the upper half of Fig. 1). They are references to evaluate the performances of the proposed V2S attack.

2.1. One-to-one parallel VC [8]

Let $\boldsymbol{G}(\cdot)$ be a VC model (a.k.a., an acoustic model), and let $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_T^\top]^\top$ and $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ be the source and targeted speakers' acoustic feature sequences extracted from a parallel speech corpus, respectively. t and T denote the frame index and total frame length, respectively. $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are the source and targeted speakers' acoustic feature vectors at frame t, respectively. A converted acoustic feature sequence $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ is given as $\hat{\boldsymbol{y}} = \boldsymbol{G}(\boldsymbol{x})$. $\boldsymbol{G}(\cdot)$ is trained to minimize a prediction error: e.g., mean squared error (MSE) between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ defined as

$$L_{\mathrm{MSE}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = \frac{1}{T}\left(\hat{\boldsymbol{y}} - \boldsymbol{y}\right)^\top \left(\hat{\boldsymbol{y}} - \boldsymbol{y}\right). \tag{1}$$

2.2. One-to-many non-parallel VC

In one-to-many VC, i.e., VC from a source speaker to any arbitrary speakers, $\boldsymbol{G}(\cdot)$ is often trained using multi-speaker corpora in advance, and is adapted using the specific targeted speaker's voice data. This paper utilizes the d-vector-based model adaptation method [9]. The method employs another DNN that estimates a d-vector (i.e., DNN-based continuous speaker representation [2]) of the targeted speaker, and feeds it to the VC model to convert the source speaker's features into the targeted speaker's ones.

3. V2S attack: building voice conversion without using targeted speaker's voice

This section proposes a novel voice impersonation attack named a V2S attack (shown in the lower half of Fig. 1). Unlike methods described in **Section 2**, the VC model is trained from a white-boxed ASV model but without the targeted speaker's voice data. Besides the ASV model, we use an ASR model for the VC model training. Deceiving the ASV model helps to reproduce the targeted speaker's individuality and using the ASR model helps to restore the phonetic property of the input voice.

3.1. ASV model to be attacked

We assume that the attacked ASV system has an ASV model $\boldsymbol{V}(\cdot)$ that extracts a latent variable of the speaker identity from input speech. In this paper, we construct a d-vector-based ASV model [2] and train it to recognize one of the enrolled S speakers. The s_yth speaker is identified by the one-hot speaker code $\boldsymbol{l}_y = [l_y(1), \cdots, l_y(s), \cdots, l_y(S)]^\top$ whose element is defined as

$$l_y(s) = \begin{cases} 1 & \text{if } s = s_y \\ 0 & \text{otherwise} \end{cases} \quad (1 \le s \le S), \tag{2}$$

where s is the speaker index. At run time, $\boldsymbol{V}(\cdot)$ outputs a frame-level posterior probability of the specific targeted speaker. Let $\boldsymbol{V}(\boldsymbol{y}) = [\boldsymbol{v}_1^\top, \cdots, \boldsymbol{v}_t^\top, \cdots, \boldsymbol{v}_T^\top]^\top$ be the probability sequence. $\boldsymbol{v}_t = [v_t(1), \cdots, v_t(s), \cdots, v_t(S)]^\top$ is the probability vector at frame t. $v_t(s)$ is the probability that $\boldsymbol{y}_t$ is uttered by the sth speaker. $\boldsymbol{V}(\cdot)$ is trained to minimize the softmax cross-entropy (SCE) loss defined as

$$L_{\mathrm{SCE}}\left(\boldsymbol{l}_y, \boldsymbol{V}(\boldsymbol{y})\right) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{S} l_y(s) \log v_t(s). \tag{3}$$

As described in **Section 1**, we set a situation in which the DNN architecture of $\boldsymbol{V}(\cdot)$ and the targeted speaker's label are given. Therefore, we can estimate the difference between the targeted speaker's individuality and that of the converted acoustic features, $\hat{\boldsymbol{y}}$, as $L_{\mathrm{SCE}}\left(\boldsymbol{l}_y, \boldsymbol{V}(\hat{\boldsymbol{y}})\right)$ that can be backpropagated to other DNNs for their training.

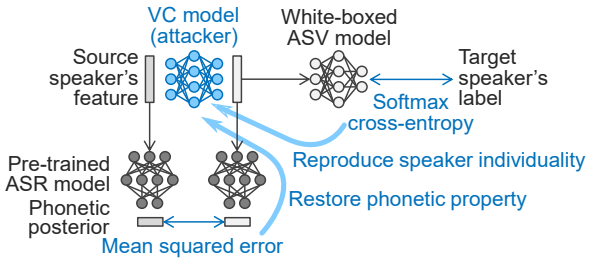

Figure 2: *Conceptual diagram of proposed V2S attack.*

3.2. ASR model to restore phonetic property

Training the VC model to deceive the ASV model, i.e., to minimize $L_{\mathrm{SCE}}\left(\boldsymbol{l}_y, \boldsymbol{V}(\hat{\boldsymbol{y}})\right)$, can be expected to reproduce the targeted speaker's individuality. However, it does not guarantee that the phonetic property of the input voice will be restored during the VC process. Therefore, we use the ASR model $\boldsymbol{R}(\cdot)$ to compute the discrepancy between the phonetic properties of the input and converted voices. As $\boldsymbol{R}(\cdot)$ predicts frame-level phonetic posteriorgrams of the input voice, we can estimate the discrepancy as the MSE between $\boldsymbol{R}(\boldsymbol{x})$ and $\boldsymbol{R}(\hat{\boldsymbol{y}})$, i.e., $L_{\mathrm{MSE}}\left(\boldsymbol{R}(\boldsymbol{x}), \boldsymbol{R}(\hat{\boldsymbol{y}})\right)$, and can use it for the proposed VC model training.

3.3. Training of VC model

A loss function $L(\cdot)$ for the proposed VC model training is given as

$$L\left(\boldsymbol{x}, \hat{\boldsymbol{y}}, \boldsymbol{l}_y\right) = L_{\mathrm{SCE}}\left(\boldsymbol{l}_y, \boldsymbol{V}(\hat{\boldsymbol{y}})\right) + \omega L_{\mathrm{MSE}}\left(\boldsymbol{R}(\boldsymbol{x}), \boldsymbol{R}(\hat{\boldsymbol{y}})\right), \tag{4}$$

where ω is a hyperparameter that controls the weight of the second term. Note that Eq. (4) does not include the targeted speaker's acoustic features $\boldsymbol{y}$ and therefore we can construct the VC model without using them. Figure 2 illustrates a conceptual diagram of the proposed V2S attack.

3.4. Discussion

The V2S attack is regarded as one kind of *voice spoofing attack* [10]. The aim is the same: masquerading oneself as another targeted speaker. So far, studies on voice spoofing attacks have focused on spoofing attacks (spoofing by synthesized voices) [10] and replay attacks (spoofing by replayed voices) [11]. The V2S attack follows the former and tries to reproduce the targeted speaker's voice from the ASV model. A similar idea was used in Saito et al.'s work [6] that incorporated a voice anti-spoofing (i.e., a discriminative model to detect spoofing attacks) into training of a VC model for reproducing fine structures of the synthesized voice.

From the above perspectives, our approach is close to that of Kinnunen et al. [12]. They proposed algorithms to select an utterance from public-domain corpora to deceive the ASV model. On the other hand, we aim to build the VC model for synthesizing every utterance of the targeted speaker.

An adversarial attack (e.g., for automatic speaker recognition [13] and image classification [14]) is the common attack for pre-trained recognition models. The purpose of the V2S attack is completely different from that of the adversarial attack: the adversarial attack aims to let the discriminator misclassify the input sample, but the V2S attack aims to reproduce attributes of the targeted sample (e.g., speaker individuality) from the discriminator.

From the above perspectives, a classifier-to-generator attack [15] and a membership inference attack [16] are approaches related to our V2S attack. They attack a pre-trained model to estimate the training data distribution or the data themselves from the model. Using these attacks is expected to improve naturalness in our converted voice, e.g., estimating utterances close to the training data of the ASV model will help deception by the VC model.

The proposed training algorithm does not use pre-stored speakers' voice data, unlike one-to-many non-parallel VC [9] and Kinnunen et al.'s work [12]. One of our future directions is to use the pre-stored speakers' voice data to improve speaker individuality of the converted voice (e.g., using transfer learning [17]). Also, integrating high-fidelity waveform generation [18] and using end-to-end ASV and ASR models [19, 20] is remaining challenges for the V2S attack in more realistic situations.

In this paper, the VC model is trained by attacking an white-boxed ASV model. A more realistic situation is an attack against the black-boxed ASV model, i.e., features, DNN architectures, and targeted speaker's label are unknown for the attacker. For instance, black-box optimization based on generative adversarial networks [21, 22] or reinforcement learning [23] can be introduced to the proposed V2S attack. Performances of the white-boxed case described in this paper will be the reference for the black-boxed case.

4. Experimental evaluation

4.1. Experimental setup

Acoustic features used in VC, ASV, and ASR were 39-dimensional (1st-through-39th) mel-cepstral coefficients and their delta features. The STRAIGHT [24] vocoder systems were used for the feature extraction. The frame shift was 5 ms. We set two types of conversion: male-to-male and male-to-female. One male was the source speaker, and two males and two females were the targeted speakers. We conducted evaluation for each pair of source and targeted speakers using their 25 parallel voices as the evaluation data.

The VC model for the V2S attack was a Feed-Forward neural network with 78 input units, $\{256, 128\}$ ReLU [25] hidden units, and 78 output units. The ASV model was a Feed-Forward neural network with 78 input units, $\{1024 \times 3, 8 \times 1\}$ sigmoid hidden units, and 260 output units. The ASR model was a Feed-Forward neural network with 78 input units, 1024×4 sigmoid hidden units, and 56 output units (56 is the number of phonemes). Utterances by 260 Japanese speakers (130 males and 130 females) were used for training the ASV and ASR models. The four targeted speakers (two males and two females) were in the 260 speakers since we performed the proposed V2S attack towards the white-boxed ASV model. AdaGrad [26] with the learning rate setting to 0.1 was used for training the VC model. The VC model training was performed with 25 epochs using 200 utterances of the source speaker. The weight ω in Eq. (4) was set to 0.01. In the V2S attack, the spectral parameters were converted by the VC model. The 0th mel-cepstral coefficients and band-aperiodicity were not converted, i.e., the original speaker's features were directly used for waveform generation. For F_0 conversion, the targeted speaker's F_0 is not observed in the V2S attack. This problem is not solved in this paper. Therefore, we calculated F_0 statistics (mean and variance) from the targeted speaker's voice data, and performed linear conversion of F_0 [5].

In one-to-one parallel VC, we trained a DNN for spectral conversion. The DNN architecture was the same as that for the VC model of the V2S attack. The number of training data is discussed in the following section. The DNN was trained with 25 epochs. The F_0 and band-aperiodicity conversions are the same as those in the V2S attack. In one-to-many non-parallel VC, the DNN architectures, training data, and other hyperparameters are the same as those in Saito et al.'s work [9] that used variational autoencoders [27] conditioned by phonetic posteriorgrams [7] and d-vectors [2] as the VC model.

4.2. Experimental results

We compared converted voices of the proposed V2S attack with those of the existing VC methods. We conducted preference AB tests on the naturalness of the converted voice. Methods to be compared were listed as follows.

- **ParaVC**: parallel VC (**Section 2.1**) trained with 5, 10, or 30 utterances
- **NonparaVC**: Non-parallel VC (**Section 2.2**)
- **V2S**: proposed V2S attack (**Section 3**)

We presented a pair of converted voices in random order and had listeners select the voice sample that sounded more natural. Similarly, preference XAB tests on the speaker individuality were conducted using the natural voices of the targeted speakers as reference samples "X." These tests were done in our crowdsourcing evaluation system. Forty listeners participated in each evaluation, and each listener evaluated 10 samples randomly extracted from the evaluation data. The total number of listeners was 2 (AB or XAB) $\times$ 2 (male-to-male or male-to-female) $\times$ 4 (reference methods) $\times$ 40 (listeners) $=$ 640.

Table 1 and Table 2 list results of the preference AB tests for male-to-male and male-to-female conversion, respectively. Similarly, Table 3 and Table 4 list those of the preference XAB tests for male-to-male and male-to-female conversion, respectively. From Table 1 and Table 2, the proposed V2S attack has lower quality than "NonparaVC." On the other hand, it has comparable or superior quality to "ParaVC (5 utts)" for both male-to-male and male-to-female conversion. From Table 3 and Table 4, the V2S attack performed worse in terms of speaker individuality than most settings of the existing VC, but it performed comparably to "ParaVC (5 utts)" for male-to-male conversion. From these results, the proposed V2S attack is potentially comparable to the standard parallel VC with a very small amount of training data.

5. Conclusion

This paper presents a new voice impersonation attack using voice conversion (VC), named the verification-to-synthesis

Table 1: *Results of preference AB tests on naturalness (male-to-male).* ***Bold*** *indicates the method preferred more with p-value* < 0.05.

A	Scores	p-value	B
ParaVC (5 utts)	0.388 vs. **0.612**	1.221×10^{-10}	V2S
ParaVC (10 utts)	0.475 vs. 0.525	0.158	V2S
ParaVC (30 utts)	0.458 vs. **0.542**	0.016	V2S
NonparaVC	**0.598** vs. 0.402	2.694×10^{-8}	V2S

Table 2: *Results of preference AB tests on naturalness (male-to-female).* ***Bold*** *indicates the method preferred more with p-value* < 0.05.

A	Scores	p-value	B
ParaVC (5 utts)	0.490 vs. 0.510	0.572	V2S
ParaVC (10 utts)	**0.593** vs. 0.407	1.365×10^{-7}	V2S
ParaVC (30 utts)	**0.610** vs. 0.390	3.174×10^{-10}	V2S
NonparaVC	**0.538** vs. 0.462	0.034	V2S

(V2S) attack. The VC model was trained to deceive the white-boxed automatic speaker verification (ASV) model for reproducing the targeted speaker's individuality and to restore phonetic property of the input voice by using pre-trained automatic speech recognition (ASR) model. The experimental results indicated that the proposed V2S attack can synthesize voice that has naturalness and speaker individuality comparable to an existing parallel VC with a very small amount of training data. In future work, we will evaluate the V2S attack that uses pre-stored speakers' voice data and investigate the dependency of the input speaker in our method.

Acknowledgements: Part of this work was supported by the SECOM Science and Technology Foundation.

6. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 788–798, 2011.

[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 *Proc. ICASSP*, Florence, Italy, May 2014, pp. 4080–4084.

[3] R. Prabhavalkar, R. Alvarez, C. Parada, P. Nakkiran, and T. Sainath, "Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4704–4708.

[4] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[5] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[6] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 755–767, Jun. 2018.

[7] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. ICME*, Seattle, U.S.A., Jul. 2016.

[8] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3893–3896.

Table 3: *Results of preference XAB tests on speaker individuality (male-to-male).* ***Bold*** *indicates the method preferred more with p-value* < 0.05.

A	Scores	p-value	B
ParaVC (5 utts)	0.530 vs. 0.470	0.090	V2S
ParaVC (10 utts)	**0.615** vs. 0.385	$< 10^{-10}$	V2S
ParaVC (30 utts)	**0.675** vs. 0.325	$< 10^{-10}$	V2S
NonparaVC	**0.660** vs. 0.340	$< 10^{-10}$	V2S

Table 4: *Results of preference XAB tests on speaker individuality (male-to-female).* ***Bold*** *indicates the method preferred more with p-value* < 0.05.

A	Scores	p-value	B
ParaVC (5 utts)	**0.585** vs. 0.415	1.324×10^{-6}	V2S
ParaVC (10 utts)	**0.713** vs. 0.287	$< 10^{-10}$	V2S
ParaVC (30 utts)	**0.705** vs. 0.295	$< 10^{-10}$	V2S
NonparaVC	**0.588** vs. 0.412	$< 10^{-10}$	V2S

[9] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5274–5278.

[10] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. ICASSP*, Vancouver, Canada, May. 2013, pp. 7234–7238.

[11] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification - a study of technical impostor techniques," in *Proc. EUROSPEECH*, Budapest, Hungary, Mar. 1999, pp. 1211–1214.

[12] T. Kinnunen, R. G. Hautamki, V. Vestman, and M. Sahidullah, "Can we use speaker recognition technology to attack itself? enhancing mimicry attacks using automatic target speaker selection," in *Proc. ICASSP*, Brighton, United Kingdom, May 2019.

[13] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," *arXiv:1810.11793*, 2018.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv:1412.6572*, 2014.

[15] K. Kusano and J. Sakuma, "Classifier-to-generator attack: Estimation of training data distribution from classifier," 2018. [Online]. Available: https://openreview.net/forum?id=SJOl4DlCZ

[16] S. Hisamoto, M. Post, and K. Duh, "Membership inference attacks on sequence-to-sequence models," *arXiv:1904.05506*, 2019.

[17] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," vol. abs/1806.04558, 2018.

[18] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," vol. abs/1609.03499, 2016.

[19] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with flexibility in utterance duration," in *Proc. ASRU*, Okinawa, Japan, Dec. 2017, pp. 584–590.

[20] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *arXiv*, vol. abs/1701.02720, 2017.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Montreal, Canada, Dec. 2014, pp. 2672–2680.

[22] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," vol. abs/1905.04874, 2019.

[23] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements," in *Proc. ICASSP*, Orleans, U.S.A., Mar. 2017, pp. 81–85.

[24] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. ICML*, Haifa, Israel, June 2010, pp. 807–814.

[26] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, Jul. 2011.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv*, vol. abs/1312.6114, 2013.