Cascade of Phonetic Speech Recognition, Speaker Embeddings GAN and Multispeaker Speech Synthesis for the VoicePrivacy 2022 Challenge

Sarina Meyer, Pascal Tilli, Florian Lux, Pavel Denisov, Julia Koch, Ngoc Thang Vu

Institute for Natural Language Processing (IMS), University of Stuttgart, Germany
sarina.meyer@ims.uni-stuttgart.de

Abstract

Speaker anonymization is the task of modifying speech recordings to hide the identity of the original speaker by changing the voice in the audio. Simultaneously, the anonymized audio should remain usable for downstream tasks and thus keep other information of the original audio like the linguistic content. This typically creates a privacy-utility trade-off of anonymization techniques. In our submission to the VoicePrivacy 2022 Challenge, we aim to reduce this trade-off by creating a speech-to-speech pipeline that (a) eliminates all clues about speaker identity by reducing the audio to phonetic transcriptions, (b) generates a new, non-existent voice using a Generative Adversarial Network, leading to artificial yet natural-like and distinctive speakers, and (c) synthesizes an anonymous version of the original utterance based on the transcriptions, anonymous speaker embedding, and estimated pitch. According to the objective evaluation, this anonymization method leads to almost perfect privacy and voice distinctiveness, and clearly outperforms all baseline systems for these two metrics. For the speech recognition utility metric, we achieve similar good results on LibriSpeech and much better ones on VCTK as compared to the baselines and the original non-anonymized data. Solely for pitch correlation, we only just meet the required threshold because our system does not use the original pitch trajectory for synthesis. Overall, our approach successfully hides the speaker identity while keeping the linguistic content, proving to be generally more effective than any of the baselines of the VoicePrivacy 2022 Challenge.

Index Terms: speaker anonymization, voice privacy, generative adversarial networks, speech synthesis, speech recognition

1. Introduction

In this paper, we describe our submission to the VoicePrivacy 2022 Challenge [1]. Our system is based on the anonymization pipeline proposed by [2] and extends it by the generation of artificial speaker embeddings using a Generative Adversarial Network (GAN). In this way, the anonymous speaker embeddings follow a similar distribution as the ones of the original speakers which allows us to sample non-existent but still realistic new voices. The whole pipeline consists then of a high-quality speech recognition model that converts the acoustic utterance into phonetic transcriptions, the GAN-based speaker embedding generation, and a high-quality speech synthesis system that uses the speaker embedding and transcription to generate an anonymous version of the original utterance.

We show improvements of our approach as compared to the baseline systems of the challenge on almost all objective metrics. The secondary utility metric of pitch correlation is the only one in which our system performs worse and only just meets the required threshold of 0.3. However, we believe it to be unfavorable to keep the pitch of the original recording because this can reveal information about the speaker. By explicitly not using the original pitch and instead estimating it from the transcribed speech, we achieve almost perfect anonymization that cannot be evaded by training the speaker verification attacker on anonymized data. Simultaneously, the anonymization results in almost the same voice distinctiveness as the data originally had, and, according to speech recognition, produces intelligible and content-preserving audio recordings.

2. Speaker Anonymization Pipeline

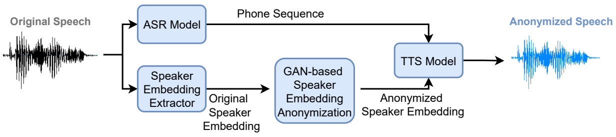
Original Speech
ASR Model
Phone Sequence
TTS Model
Anonymized Speech
Speaker Embedding Extractor
Original Speaker Embedding
GAN-based Speaker Embedding Anonymization
Anonymized Speaker Embedding

Figure 1: *Architecture of the speaker anonymization pipeline.*

The pipeline in its general outline is the same as described by [2] and shown in Figure 1. It consists of four models: (i) a speech recognition (ASR) model to transcribe the speech into phonetic sequences, (ii) a speaker embedding extractor, (iii) an anonymization module that exchanges the original speaker embedding by an anonymous one, and (iv) a text-to-speech (TTS) system to convert content and speaker embedding into spoken utterances. Each component will be described in more detail in the following.

2.1. Speech Recognition Module

Our ASR model is based on the hybrid CTC/attention architecture [3] with a Conformer as encoder [4] and a Transformer decoder. It is implemented in the ESPnet2 toolkit [5]. The neural network follows the standard configuration with $d_{\text{model}} = 512$, $d_{\text{ff}} = 2048$, $d_h = 8$, $E = 12$, $D = 6$ and Conv kernel size of 31. 80-dimensional log-mel Filterbank features are extracted from the input speech in order to be processed by the ASR model. The output of this module is not text, as typical for speech recognition, but phone sequences. We use SentencePiece [6] to learn 100 unigram language model [7] subword units from the phonemized training data transcriptions. Transcriptions are phonemized using the IMS Toucan toolkit [8]. Label smoothing with a penalty of 0.1, as well as SpecAugment [9] and 3-way speed perturbation [10] data augmentation methods are utilized during training. The training is performed with Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and warmup learning rate scheduler with 40,000 steps. Batch size is set to 10M bins resulting in the average batch size of 81 utterances. Gradients are accumulated for 8 steps before the model update is performed. The initial model is trained on a combination of train-clean-100 and train-other-500 subsets of

LibriTTS corpus [11] with the total duration of 363 hours. Validation data consisting of LibriSpeech [12] dev-clean and VCTK [13] dev subsets with the total duration of 15 hours is used for the early stopping and checkpoint selection. Learning objective coefficient γ is set to 0.6 and base learning rate is set to 0.005 during the initial training. The initial model is used to label the whole VoxCeleb 1 and 2 corpora [14, 15, 16] and is finetuned on the resulting labeled data combined with the original LibriTTS train-clean-100 and train-other-500 subsets with the total duration of 2954 hours. Learning objective coefficient γ is set to 0.1 and base learning rate is set to 0.003 during the finetuning. This process is repeated twice before the final ASR model is obtained. Our intention is to expose the ASR model to more diverse training data in order to make it more robust to recording conditions, speaking styles and accents. This is particularly important in a cascaded system because all ASR errors are propagated to the TTS module. The Phone Error Rate (PER) values on the LibriSpeech dev/test clean and VCTK dev/test subsets are 6.5%/6.4% and 5.0%/9.4% for the initial model, 6.3%/6.1% and 3.8%/8.4% after the first finetuning, and 6.2%/6.1% and 3.6%/8.3% for the final model (it should be noted that the VCTK dev subset is used with non-normalized transcriptions here). The effect of finetuning is more evident on VCTK corpus and we link this to higher accent variability in this dataset.

2.2. Speaker Embedding Extraction Module

We use two speaker embedding methods to extract the speaker identity information from the recordings: x-vector [17] and the more recent ECAPA-TDNN [18]. As presented by [2], both vector types contain complementary information about the speaker, and it is beneficial to use the concatenation of both instead of just one of them for synthesis. Thus, we retrieve from every recording one x-vector and one ECAPA-TDNN embedding by applying the extractors provided by SpeechBrain [19] which have been trained on VoxCeleb 1 and 2 [14, 15, 16]. The resulting speaker embedding is then a concatenation of the $192d$ ECAPA-TDNN vector and the $512d$ x-vector, resulting in 704 dimensions in total.

2.3. Speaker Embedding Anonymization Module

We train a Wasserstein GAN with Quadratic Transport Cost (WGAN-QC) [20] to generate artificial speaker embeddings. It consists of a generator and a discriminator. Unlike the classic GAN [21] technique, the discriminator in a Wasserstein Generative Adversarial Network (WGAN) [22] is not optimized towards distinguishing between between real and fake data but to decrease the distance between real and fake distributions, and is therefore called *critic*. This is achieved by training the critic to compute the quadratic Wasserstein distance [20] between the distributions, and by training the generator to minimize that distance. The WGAN-QC furthermore includes the quadratic transport cost to improve the convergence of the model. The GAN is trained on the training subset of the clean-100 part of LibriTTS.

Both the generator and the critic networks are simplified versions of ResNet [23] as proposed to use by [20]. The ResNet models are reduced in size to 150,000 parameters because of the limited amount of data in the challenge. Input to the generator is a 16-dimensional random noise vector that is sampled from a standard normal distribution $\mathcal{N}(0, 1)$. We experimented with different sizes of $z = 32$ and $z = 64$. It then generates a 704-dimensional vector that should be similar to any speaker embedding extracted from the original data. This is then measured by the critic which computes the distance between the generated vectors and the ones retrieved by the speaker embedding extraction module. The important hyperparameter γ, which balances the regression and regularization terms, is tuned between 0.1 and 1. We train the model with a batch size of 128, the Adam optimizer [24] with an initial learning rate of $5e-5$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We found that a $\gamma = 1$ and $z = 16$ slightly decrease the Word Error Rate (WER). For simplification, we tested to exchange the ResNet models by a four layer Multilayer Perceptron (MLP) that matches the number of trainable parameters as generator and critic. However, this drastically decreases the performance of the system by a large margin and slows down the training and convergence process of the WGAN.

We sample a new random vector z for every new voice that we want to anonymize and make sure that its cosine similarity to the original embedding of that speaker is smaller than 0.7 to avoid unexpectedly sampling a too similar voice. Thus, for all data during evaluation, we perform speaker-level anonymization by sampling one generated embedding per speaker and using this for all utterances in that particular dataset split. This leads to different anonymized voices for the same speaker in the respective development and test data for trial and enrollment utterances. For anonymizing the train-clean-360 dataset that is used for training the evaluation models, we perform utterance-level anonymization by sampling a new speaker embedding for each utterance.

During this anonymization step, no information about the actual speaker is maintained, not even the gender. We experimented with training different GAN models for generating female and male voices but ran into issues with the lack of sufficient data to train the models. Therefore, we apply only one GAN for all voices regardless of the gender.

2.4. Speech Synthesis Module

The TTS module is implemented using the IMS Toucan toolkit [8] and was trained on the clean-100 part of LibriTTS. It uses a FastSpeech 2 [25] parallel synthesis approach for an emphasis on robustness and speed. The encoder and decoder are build according to the Conformer architecture [4]. The inputs, which are the phonetic transcriptions as produced by our ASR module, are transformed into articulatory feature vectors for synthesis [26], to reduce the impact of near-misses in the ASR output. Furthermore, this would help making the approach agnostic of the phoneme set used if more languages than just English were to be incorporated. The estimators for pitch and energy are used as presented by FastSpeech 2 [25] with the averaging of pitch and energy over the duration of a phoneme proposed by FastPitch [27]. This enables fine-grained control over the realizations of individual phonemes, as is shown in e.g. [28], which would also further allow us to make an exact clone of the prosody [29]. Because this however also makes the speech identifyable to an extend, we decided against using this approach in our TTS module for now. Getting accurate measurements for pitch and energy are very important with this approach. In preliminary experimentation we found that the algorithms within Praat [30] work best for extracting these values, which we do using an open-source interface to Praat's algorithms[1]. The transformation from spectrograms to waveforms is performed by a HiFi-GAN vocoder [31] and also implemented using the

[1] https://github.com/YannickJadoul/Parselmouth

Table 1: *Primary privacy evaluation results for* ***speaker verification****, as EER in %, in comparison to the original data and all baselines.*

Dataset	Gender	Weight	Orig.	B1.a	B1.b	B2	Our
LibriSpeech-dev	female	0.25	8.67	17.76	19.03	11.36	44.60
	male	0.25	1.24	6.37	5.59	1.40	43.63
VCTK-dev (different)	female	0.20	2.86	12.46	8.25	6.68	51.04
	male	0.20	1.44	9.33	6.01	6.35	50.97
VCTK-dev (common)	female	0.05	2.62	13.95	9.01	5.81	51.16
	male	0.05	1.43	13.11	9.40	8.83	51.57
Weighted average dev			3.54	11.74	9.93	6.53	47.60
LibriSpeech-test	female	0.25	7.66	12.04	9.49	7.12	44.89
	male	0.25	1.11	8.91	7.80	1.11	45.43
VCTK-test (different)	female	0.20	4.89	16.00	10.91	16.92	49.02
	male	0.20	2.07	10.05	7.52	7.69	49.54
VCTK-test (common)	female	0.05	2.89	17.34	15.32	10.98	50.58
	male	0.05	1.13	9.89	8.19	4.80	51.13
Weighted average test			3.79	11.81	9.18	7.77	47.38

IMS Toucan toolkit. To produce synthesized speech with different voices, the TTS has been conditioned and trained on the same concatenated speaker embedding setup as the anonymization module by concatenating the embeddings to the output of the encoder, as [32] propose.

It should be noted that since the synthesis is based only on the recognized transcription and the generated speaker embedding, no acoustic or prosodic information is used from the original audio. Thus, the resulting synthesized utterance will most likely differ from the original one in terms of pitch, energy, and speaking rate. However, due to punctuation and context, the estimated values especially of pitch are often relatively close to the original ones. This issue will be addressed again regarding the challenge evaluation in Section 3.3.

3. Results

3.1. Evaluation setup

The objective evaluation in the challenge is divided into primary and secondary, with two metrics in each category. The first part of the primary evaluation consists in measuring the privacy of the approach by computing the Equal Error Rate (EER) of an automatic speaker verification (ASV) attacker. Prior to evaluation, the attack system is retrained on the train-clean-360 part of the LibriSpeech corpus that has been anonymized on utterance-level by the same anonymization system as described above, leading to a strong attacker. For the privacy metric, we are aiming for EER scores close to 50% because this indicates a random prediction behavior by the attacker. The second metric in the primary evaluation is the WER in speech recognition as utility metric. As we are using the evaluation suite of the challenge, this ASR model is different to the one embedded in our anonymization system. However, similar to the ASV attacker, the ASR model is also retrained on the anonymized train-clean-360 data. Lower WER scores are better because they imply that the anonymized speech retained the original linguistic content with sufficient intelligibility.

In the secondary evaluation, the pitch correlation ρ^{F_0} and Gain of Voice Distinctiveness (GVD) are assessed for further utility of the system. The pitch correlation metric was introduced to ensure that intonation is not completely lost during anonymization. It measures the Pearson correlation between the pitch sequences of the original and anonymized utterances. For the challenge, a $\rho^{F_0} > 0.3$ must be reached for all datasets. The GVD metric also compares original and anonymized utterances but assesses their difference in terms of voice distinctiveness, i.e., how well the voices of different speakers can be distinguished. A GVD score of zero denotes the same voice distinctiveness in the anonymized as in the original data. Scores above or below zero indicate a increased or decreased distinctiveness, respectively. Thus, it is favorable to achieve a GVD either close to zero or above in order to avoid losing the discriminability of different voices in multi-speaker conversations.

For all metrics, we report the results of the baseline systems of the challenge as given in [1] and the results of our system for different datasets, as well as their weighted average over all datasets. In all scenarios but the speech recognition, results for female and male speakers are separated. For the primary evaluation, the ASV and ASR scores on the original, non-anonymized data are also presented.

3.2. Primary evaluation

The results of the primary privacy evaluation are presented in Table 1. For all datasets, our system reaches scores close to 50%, denoting almost perfect privacy. In contrast to this, the performance of all baseline systems is significantly worse. Our privacy results are close to the ones reported by [2] using the evaluation framework of the VoicePrivacy 2020 Challenge [33]. This suggests that this pipeline combined with the GAN-generated speaker embeddings is robust against different kinds of privacy attacks, even if the attacker has been trained on anonymized data of the same anonymization technique.

Table 2: *Primary utility evaluation results for* ***speech recognition****, as WER in %, in comparison to the original data and all baselines.*

Dataset	Orig.	B1.a	B1.b	B2	Our
LibriSpeech-dev	3.82	4.34	4.19	4.32	4.56
VCTK-dev	10.79	11.54	10.98	11.76	9.02
Average dev	7.31	7.94	7.59	8.04	6.79
LibriSpeech-test	4.15	4.75	4.43	4.58	4.53
VCTK-test (different)	12.82	11.82	10.69	13.48	7.81
Average test	8.49	8.29	7.56	9.03	6.17

The WER results of the primary utility evaluation are given in Table 2. For LibriSpeech, our system achieves a similar low

Table 3: *Secondary utility evaluation results for* ***pitch correlation*** ρ^{F_0}*, in comparison to the original data and all baselines.*

Dataset	Gender	Weight	B1.a	B1.b	B2	Our
LibriSpeech-dev	female	0.25	0.77	0.84	0.64	0.36
	male	0.25	0.73	0.76	0.53	0.34
VCTK-dev (different)	female	0.20	0.84	0.87	0.70	0.40
	male	0.20	0.78	0.76	0.59	0.39
VCTK-dev (common)	female	0.05	0.79	0.84	0.64	0.31
	male	0.05	0.72	0.72	0.54	0.31
Weighted average dev			0.77	0.80	0.61	0.37
LibriSpeech-test	female	0.25	0.77	0.85	0.61	0.35
	male	0.25	0.69	0.72	0.54	0.30
VCTK-test (different)	female	0.20	0.84	0.87	0.68	0.42
	male	0.20	0.79	0.77	0.66	0.39
VCTK-test (common)	female	0.05	0.79	0.85	0.65	0.33
	male	0.05	0.70	0.71	0.61	0.30
Weighted average test			0.77	0.80	0.62	0.36

WER as the baselines, which are all close to the intelligibility of the original data. VCTK, on the other hand, is a more challenging dataset because its speakers use different English accents. The anonymized speech of our system leads not only to a clear reduction in WER as compared to the baselines but also to the original data. This suggests that our approach eliminates the accent information in the recordings – which can give clues about the identity of the speaker – and that the ASR module in our pipeline is able to recognize the speech correctly regardless of the accent.

3.3. Secondary evaluation

As a measure of preserved intonation, the pitch correlation scores are shown in Table 3. Since we do not use the original pitch in our system, our results on this metric are poor, although meeting the requirement of 0.3 for all datasets. This poor performance can be seen as a disadvantage of our approach. However, we argue that it is not advantageous to keep the original pitch because this can reveal the identity of the original speaker. By reaching the required threshold, we show that the pitch estimation in our approach is good enough to approximate the pitch of the original recording without copying the speaker-specific prosody style. The reason why we meet the required threshold although not using the original pitch lies in the design of our ASR and TTS models: Since the ASR module is trained on LibriTTS, its outputs contain punctuation which gives important clues about the intonation of the original utterance. Using this information together with the context of the utterance including the (phonemized) word order, the pitch estimator in the TTS module generates an intonation pattern that is close to how the original speaker probably pronounced the utterance. This process is facilitated by the fact that the evaluation data contains only standard speech without much variation.

The results of the gain of voice distinctiveness metric are given in Table 4. A GVD score close to zero denotes a similar voice distinctiveness as in the original data, which our approach almost completely fulfills. As the baseline systems achieve smaller scores further away from zero, we show that our GAN-based speaker embeddings lead to more distinguishable voices which we believe are due to them being sampled from a natural-like embedding distribution. In fact, [2] show that generating speaker embeddings that match original ones in terms of value ranges is important for the TTS system in order to produce different voices for different input embeddings.

4. Conclusion

For the VoicePrivacy 2022 Challenge, we proposed a system that reduces the original speech to the linguistic content before synthesizing it back to audio with an artificial voice created by a Generative Adversarial Network. This method effectively eliminates all information about the original speaker identity and thus successfully anonymizes it, while keeping the content and intelligibility of the recorded speech. Regarding the primary privacy evaluation, we clearly outperform all baseline systems, achieving almost perfect anonymization. In the primary utility metric, measured by speech recognition, our results are similar to the baselines on LibriSpeech and distinctly better on VCTK. For VCTK which contains utterances in different English accents, we even significantly outperform the speech recognition results on the original, non-anonymized speech, suggesting that the accent of the original speaker is hidden in the anonymized utterance. In the secondary evaluation, we show that we, contrary to the baselines, keep almost the same voice distinctiveness as in the original data. The only metric for which our system produces worse results is the pitch correlation. We explicitly do not keep the pitch of the original voice recording because this reveals information about the speaker, making the anonymization less effective. Instead, we use a smart pitch estimation method for generating a sensible pitch trajectory for the transcribed speech which matches the original pitch enough for fulfilling the pitch correlation requirement.

5. References

[1] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, "The VoicePrivacy 2022 Challenge evaluation plan," 2022. [Online]. Available: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2022_Eval_Plan_v1.0.pdf

[2] S. Meyer, F. Lux, P. Denisov, J. Koch, P. Tilli, and N. T. Vu, "Speaker Anonymization with Phonetic Intermediate Representations," 2022.

[3] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[4] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," *Interspeech*, pp. 5036–5040, 2020.

Table 4: *Secondary utility evaluation results for* ***gain of voice distinctiveness*** G_{VD}*, in comparison to the original data and all baselines.*

Dataset	Gender	Weight	B1.a	B1.b	B2	Our
LibriSpeech-dev	female	0.25	-9.15	-4.92	-1.94	-0.03
	male	0.25	-8.94	-6.38	-1.65	-0.27
VCTK-dev (different)	female	0.20	-8.82	-5.94	-1.32	-0.27
	male	0.20	-12.61	-9.38	-2.18	-0.45
VCTK-dev (common)	female	0.05	-7.56	-4.17	-1.14	0.03
	male	0.05	-10.37	-6.99	-1.32	-0.11
Weighted average dev			-9.71	-6.44	-1.72	-0.23
LibriSpeech-test	female	0.25	-10.04	-5.00	-1.71	-0.15
	male	0.25	-9.01	-6.64	-1.74	-0.15
VCTK-test (different)	female	0.20	-10.29	-6.09	-1.56	0.35
	male	0.20	-11.69	-8.64	-1.56	-0.15
VCTK-test (common)	female	0.05	-9.31	-5.10	-1.59	0.17
	male	0.05	-10.43	-6.50	-1.36	0.03
Weighted average test			-10.15	-6.44	-1.63	-0.03

[5] S. Watanabe, F. Boyer, X. Chang, P. Guo, T. Hayashi, Y. Higuchi, T. Hori, W.-C. Huang, H. Inaguma, N. Kamo *et al.*, "The 2020 espnet update: new features, broadened applications, performance improvements, and future plans," in *2021 IEEE Data Science and Learning Workshop (DSLW)*. IEEE, 2021, pp. 1–6.

[6] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *EMNLP*, 2018, pp. 66–71.

[7] T. Kudo, "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates," in *ACL*, 2018, pp. 66–75.

[8] F. Lux, J. Koch, A. Schweitzer, and N. T. Vu, "The IMS Toucan system for the Blizzard Challenge 2021," in *Proc. Blizzard Challenge Workshop*, vol. 2021. Speech Synthesis SIG, 2021.

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *Interspeech*, pp. 2613–2617, 2019.

[10] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[11] 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 *Interspeech*, 2019, pp. 1526–1530.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[13] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," 2019.

[14] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[15] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Interspeech*, 2017.

[16] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech*, 2018.

[17] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *ICASSP*, 2018, pp. 5329–5333.

[18] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[19] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A General-Purpose Speech Toolkit," 2021, arXiv:2106.04624.

[20] H. Liu, X. Gu, and D. Samaras, "Wasserstein gan with quadratic transport cost," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4832–4841.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, vol. 27, 2014.

[22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *ICML*. PMLR, 2017, pp. 214–223.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *International Conference on Learning Representations*, 2020.

[26] F. Lux and T. Vu, "Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features," in *ACL*, 2022, pp. 6858–6868.

[27] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *ICASSP*. IEEE, 2021, pp. 6588–6592.

[28] J. Koch, F. Lux, N. Schauffler, T. Bernhart, F. Dieterle, J. Kuhn, S. Richter, G. Viehhauser, and N. T. Vu, "PoeticTTS - Controllable Poetry Reading for Literary Studies," 2022. [Online]. Available: https://arxiv.org/abs/2207.05549

[29] F. Lux, J. Koch, and N. T. Vu, "Prosody cloning in zero-shot multispeaker text-to-speech," *arXiv preprint arXiv:2206.12229*, 2022.

[30] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[31] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *NeurIPS*, vol. 33, 2020.

[32] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arXiv preprint arXiv:1806.04558*, 2018.

[33] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Interspeech*, 2020, pp. 1693–1697.