SYNVOX2: TOWARDS A PRIVACY-FRIENDLY VOXCELEB2 DATASET

Xiaoxiao Miao[1,2], *Xin Wang*[1], *Erica Cooper*[1], *Junichi Yamagishi*[1],
Nicholas Evans[3], *Massimiliano Todisco*[3], *Jean-François Bonastre*[4], *Mickael Rouvier*[4]

[1]National Institute of Informatics, Japan [2]Singapore Institute of Technology, Singapore
[3]Digital Security Department, EURECOM, France [4]LIA, University of Avignon, France

ABSTRACT

The success of deep learning in speaker recognition relies heavily on the use of large datasets. However, the data-hungry nature of deep learning methods has already being questioned on account the ethical, privacy, and legal concerns that arise when using large-scale datasets of natural speech collected from real human speakers. For example, the widely-used VoxCeleb2 dataset for speaker recognition is no longer accessible from the official website. To mitigate these concerns, this work presents an initiative to generate a privacy-friendly synthetic VoxCeleb2 dataset that ensures the quality of the generated speech in terms of privacy, utility, and fairness. We also discuss the challenges of using synthetic data for the downstream task of speaker verification.

***Index Terms*—** Privacy-friendly data, speaker anonymization, language-robust orthogonal Householder neural network

1. INTRODUCTION

Large-scale speech data and powerful computing resources are key to the success of deep learning methods in automatic speaker verification (ASV). However, the use of speech as a form of biometric data is governed by a set of legal restrictions, such as the General Data Protection Regulation (GDPR) [1].

Related legal and ethical issues have already led to the withdrawal of well-known large-scale datasets used for face recognition research, namely the VGGFace2 [2] and MS-Celeb-1M [3] databases both of which were constructed by crawling facial images from the web. Researchers have hence begun to explore the potential of using synthetic images for face recognition research [4, 5, 6]. The ASV field has faced similar problems due to privacy issues. For instance, the widely-used large-scale VoxCeleb2 dataset [7], which contains speech data collected from 5,994 speakers, has become a standard ASV benchmark, though the database is *no longer available* from the official website[1]. The withdrawal of other popular biometric databases will likely soon follow. It is then inevitable that the community will have no choice but to consider alternatives.

As synthetic data is a promising option, this work aims to explore the creation and utilization of a privacy-friendly synthetic VoxCeleb2 dataset for ASV model training. Figure 1 illustrates the general idea of the synthetic VoxCeleb2 called SynVox2. The original VoxCeleb2, which is also referred to as *authentic VoxCeleb2*, is fed to a speech generator to create a *synthetic VoxCeleb2* database, which can meet two primary criteria: adequately protect speaker privacy while maintaining utility comparable to the authentic database. One sub-criteria is to mitigate data bias and increase fairness. The synthetic VoxCeleb2 dataset is subsequently used to train ASV models, with the goal of achieving comparable results to the same models trained with the authentic VoxCeleb2.

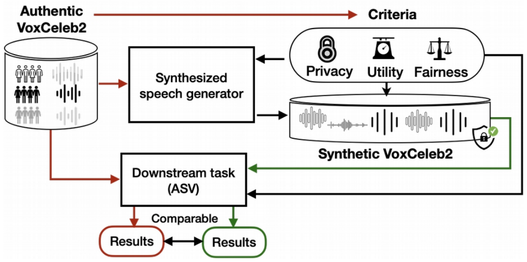

Fig. 1. Overview of the creation of the synthetic Voxceleb2 dataset. The scenario assumes a reliable party holds authentic data for generating and evaluating shareable synthetic data, with the assurance that the authentic data will not be released.

One possible solution to protect speaker privacy while still allowing the sharing of speech data is through speaker anonymization [8, 9, 10, 11, 12, 13]. This method hides the speaker's identity (privacy) while preserving other speech characteristics, such as content and emotion. Anonymization techniques could also be used to generate synthetic speaker voices and data. By using speaker anonymisation for this purpose, we aim to generate privacy-friendly synthetic dataset, which can strike a balance between protecting privacy and supporting research.

In this work, we employ the recently proposed language-robust orthogonal Householder neural network (OHNN)-based speaker anonymization technique [12] to create a privacy-friendly VoxCeleb2 dataset called SynVox2. With the same number of speakers as the authentic VoxCeleb2 dataset, SynVox2 can be shared with far fewer privacy concerns compared to the sharing of the authentic VoxCeleb2 database. In addition, we define several metrics for evaluating the use of SynVox2 in terms of privacy, utility, and fairness. These metrics may serve as a protocol for future research, enabling researchers to assess whether a synthetic dataset is suitable for their ASV research. Furthermore, we conduct an in-depth analysis of intra-/inter-speaker variations in SynVox2, aiming to improve the utility of SynVox2. Specifically, we propose methods for increasing intra-/inter-speaker variations, such as modeling background noise from authentic speech and incorporating it into the generated speech.

2. REQUIREMENTS FOR A PRIVACY-FRIENDLY SYNTHETIC SPEECH DATABASE

This section describes the three requirements a privacy-friendly synthetic speech database should satisfy: privacy, utility, and fairness. It also introduces the evaluation metrics used to assess the degree to

[1] https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html

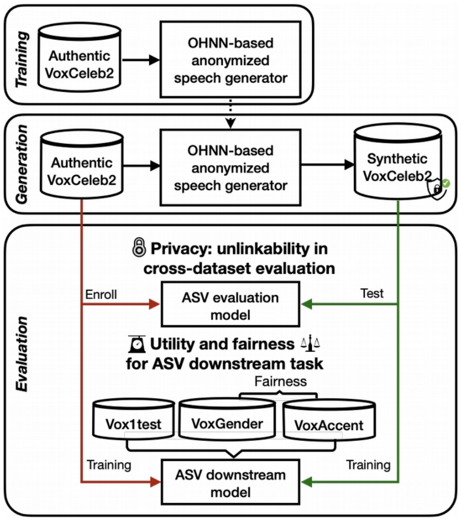

Fig. 2. Flowchart of OHNN-based synthetic VoxCeleb2 generation and evaluation.

which these requirements are fulfilled.

2.1. Ensuring Privacy through Unlinkability

Privacy-sensitive information in speech extends beyond that related to speaker identity. Nonetheless, privacy can still be preserved to a great extent by obfuscating the speaker identity since any remaining privacy-sensitive content cannot be linked to the original speaker. We hence consider a privacy-friendly synthetic speech database to be able to protect the speaker's identity. The privacy of a speaker's identity in a synthetic database is protected if it is *unlinkable* to its original identity in the authentic database [14][2]. Given two speech samples of a speaker from the synthetic and the authentic databases, respectively, privacy is protected if it is difficult to determine that the two samples have been uttered by the same speaker.

This study evaluates privacy protection performance using an ASV evaluation model. The enrollment data is from the authentic VoxCeleb2 database, while the test trial is from SynVox2, as shown in Figure 2. A privacy-friendly speech database should achieve a as high as possible ASV equal error rate (EER), which indicates that the ASV evaluation model has difficulty linking an authentic enrollment utterance and a protect test utterance.

2.2. Maintaining Data Utility

Downstream models trained on a privacy-friendly synthetic database with high utility are expected to perform similarly to models trained using authentic data. This study considers ASV as a downstream task, shown to the bottom part of Figure 2. Utility is assessed in terms of the ASV EER which is estimated from experiments performed on authentic test sets. ASV models trained using either authentic VoxCeleb2 or SynVox2 should exhibit similar performance.

2.3. Reducing Data Bias and Increasing Fairness

While data utility mainly measures the overall downstream performance on test sets, it is essential to ensure that downstream models trained on a privacy-friendly synthetic database do not disfavor any particular group in the test set, e.g., genders, dialects, and ethnicities.

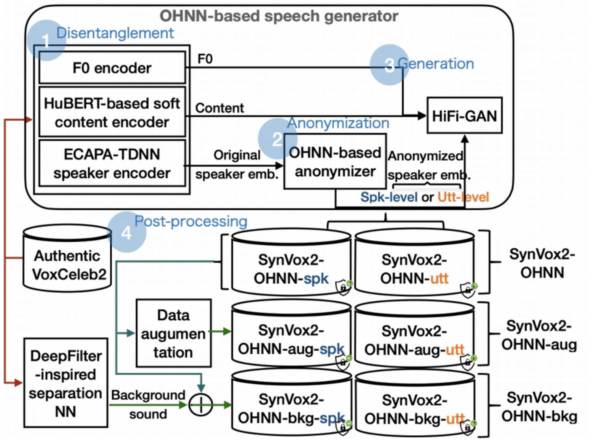

Fig. 3. Architecture of OHNN-based anonymized speech generator.

With ASV as the downstream task, this study uses the Fairness Disrepancy Rate (FDR) [15, 16] to assess the fairness. Given decision threshold τ, the FDR considers the largest distance between false alarm rates FAR (i.e., non-target speaker trials being classified as target) and false reject rates FRR (i.e., target trials being classified as non-target) over multiple groups. Given a set of groups $D = \{d_1, d_2, ..., d_n\}$, the FDR is defined as:

$$\begin{aligned}\mathrm{FDR} = 1 - \Big(&\alpha \times \max(|\mathrm{FAR}^{d_i}(\tau) - \mathrm{FAR}^{d_j}(\tau)|)\\ &+ (1-\alpha) \times \max(|\mathrm{FRR}^{d_i}(\tau) - \mathrm{FRR}^{d_j}(\tau)|)\Big).\end{aligned} \quad (1)$$

where FAR^{d_i} and FAR^{d_j} are the false alarm rates for groups d_i and d_j, $\forall d_i, d_j \in D$, respectively. FRR^{d_i} and FRR^{d_j} are the false reject rates of groups d_i and d_j, respectively. $\alpha \in [0, 1]$ is a design choice and set to 0.95 so as to give greater importance to disparities in the rate of more costly false alarms than to the rate of false rejects. FDR $= 1$ means the system is perfectly fair. FAR^{d_i}, FAR^{d_j}, FRR^{d_i}, and FRR^{d_j} of each subgroup are calculated using a fixed decision threshold $\tau = 10e - 1$, when considering trials of all the groups together.

We study fairness across gender and accent groups. The gender group consists of female and male speakers. The accent group includes English speakers haling from Austria, Canada, France, Germany, India, Italy, the Netherlands, Spain, the UK and the USA.

3. SYNVOX2 GENERATION METHODS

This section presents the process to create a privacy-friendly synthetic version of the VoxCeleb2 database that satisfies the constraints and criteria presented in section 2.

3.1. Language-robust OHNN-based speaker anonymization

Speaker anonymization is one approach to create privacy-friendly synthetic datasets. Because VoxCeleb2 is a multilingual dataset, we use a recently proposed language-robust OHNN-based anonymization method [12] to generate different speaker-anonymised versions of VoxCeleb2. It uses a self-supervised learning (SSL)-based content encoder and an OHNN-based anonymizer, which supports the generation of speaker-distinctive anonymized speech even in languages unseen in training. The generation process involves three steps, as shown to the top of Figure 3:

1) Disentanglement: The YAAPT algorithm [17] is used to extract F0. The ECAPA-TDNN speaker encoder is trained on the *VoxCeleb2* [7] datasets and provides 192-dimensional speaker identity

[2] Again, the authentic database will not be released in our assumed scenario, shown in Figure 1

representations. The HuBERT-based soft content encoder is fine-tuned on *LibriTTS-train-clean-100* [18] from a pre-trained HuBERT Base model[3] to capture the speech contents.

2) Anonymization: The OHNN-based anonymizer [12] rotates the original speaker embeddings to corresponding anonymized speaker embeddings using multiple orthogonal Householder transformation [19] layers. The weights of the OHNN are randomly initialized and trained with classification and distance losses that prevent anonymized speakers from overlapping with other original and anonymized speakers. Hence, the anonymized speaker embeddings generated by the trained OHNN-based anonymizer are expected to have distinctive pseudo-speaker identities.

3) Generation: Finally, the content features, F0, and anonymized speaker embeddings are passed to a HiFi-GAN model [20] for audio waveform generation. The HiFi-GAN model is trained using the *LibriTTS-train-clean-100* database [18].

3.2. Synthetic Datasets

The dataset generated directly by the OHNN-based speech generator is referred to as *SynVox2-OHNN*, as illustrated at the top of Figure 3.

Because VoxCeleb2 was collected under diverse real-world conditions and contains various types of background noise, the OHNN-based speech generator pre-trained solely on clean data may not reproduce the background noise of the authentic VoxCeleb2 database. This would reduce the variations among utterances in *SynVox2-OHNN*. To alleviate the reduced intra/inter-speaker variation, we generate different versions of SynVox2 using the post-processing methods shown at the bottom of Figure 3.

1) SynVox2-OHNN-aug: One straightforward approach is directly adding noise, reverberation to *SynVox2-OHNN*. Similar to the standard speech data augmentation method [21], we use room impulse responses and background noise from [21] and the MUSAN database [22]. Although this seems to be redundant to the data augmentation in downstream ASV training, our experiments demonstrated that ASV downstream models benefited from the double augmentation.

2) SynVox2-OHNN-bkg: Inspired by techniques used for preserving background sound in voice conversion [23], we use a pre-trained DeepFilter-inspired model [24] to separate background sound and clean speech from authentic speech. The background sound is then added to *SynVox2-OHNN* to ensure consistent ambient characteristics. DeepFilterNet-inspired model was developed in-house and trained on a large-scale dataset comprising noise extracted from YouTube videos and clean speech.

Another factor affecting the intra-speaker variations stems from the manner in which we extract speaker embeddings. One strategy involves extraction of embeddings by grouping together the set of all utterances corresponding to each speaker (denote *-spk* in Figure 3). Although this strategy ensures a consistent pseudo speaker identity, it results in the reduced intra-speaker variations. We hence explore an alternative utterance-level approach (denote *-utt* in Figure 3), in which anonymized speaker embeddings are extracted individually from each authentic input utterance. The anonymized embeddings extracted from utterances corresponding to the same speaker may then differ and better capture intra-speaker variability. We explored both approaches each combined with the two post-processing approaches.

Table 1. EERs (%) achieved by the ASV evaluation model in cross-dataset evaluation. EER in first row is calculated on authentic VoxCeleb2 dataset for reference. Other rows are results in various cross-dataset conditions. Higher EERs in cross-dataset conditions indicate that the synthetic and authentic speakers are hardly associated.

Cross datasets	EER(%)
Authentic vs Authentic	2.07
Authentic vs. SynVox2-OHNN-spk	32.66
Authentic vs. SynVox2-OHNN-aug-spk	34.43
Authentic vs. SynVox2-OHNN-bkg-spk	27.84
Authentic vs. SynVox2-OHNN-utt	32.81
Authentic vs. SynVox2-OHNN-aug-utt	34.76
Authentic vs. SynVox2-OHNN-bkg-utt	27.81

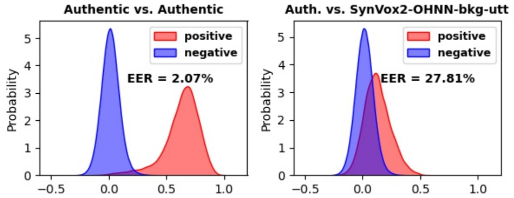

Fig. 4. EERs (%) and score distribution achieved by ASV evaluation model in cross-dataset evaluation for authentic VoxCeleb2 and SynVox2-OHNN-bkg-utt datasets.

4. EXPERIMENTS

4.1. Setup

OHNN Training: The language-robust OHNN-based anonymized speech generator is trained on authentic VoxCeleb2 used random orthogonal Householder reflections, with a random seed of 50 for parameter initialization. We used an additive loss function which combines weighted angular margin softmax and cosine similarity. Full details of the training procedure can be found in [12].

SynVox2 Evaluation: **Unlinkability**: We generate enrollment-test pairs, where enrollment utterances are sourced from the authentic VoxCeleb2 database, whereas test utterances are selected from the SynVox2 database. Specifically, for each speaker, we randomly selected four utterances to form two same-speaker trials. For each speaker, we also generated 100 different-speaker trials by pairing one of the 4 utterances with 100 utterances of different speakers. This resulted in a total of $611388 = (2 + 100) * 5994$ trials. We used a publicly available ECAPA-TDNN[4] ASV system designed using the SpeechBrain toolkit [25] to compute the EER. **Utility**: First, to establish a benchmark, we trained a downstream ASV model (ECAPA-TDNN with 512 channels in the convolution frame layers [26]) using the SpeechBrain recipe with data augmentation. Then, we trained multiple alternative downstream ASV models using the same recipe as the benchmark ASV model, but with each of the SynVox2 databases The utility of each model was then estimated in terms of the EER using the test partition of the VoxCeleb1 database [27]. **Fairness**: We built gender and accent test sets from the VoxAccent dataset [16] to further assess fairness. VoxAccent is a subset of the VoxCeleb2 training set and include utterances collected from 157 male and 112 female speakers who together cover 10 different English accents. The number of trials for each accent

[3] https://github.com/pytorch/fairseq/tree/main/examples/hubert

[4] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

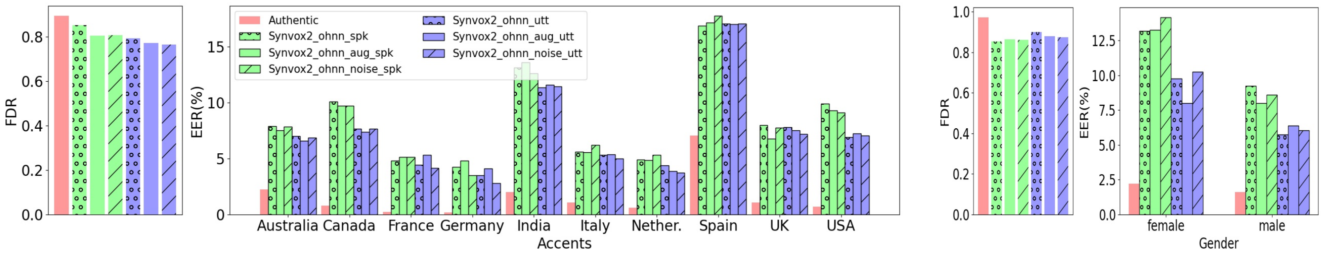

Fig. 5. EER and FDR for accent (left two plots) and gender (right two plots) groups, respectively.

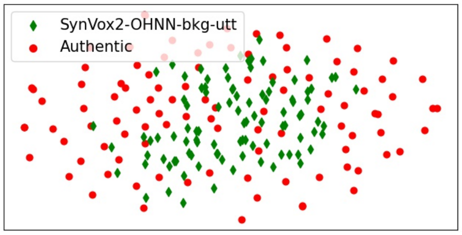

Fig. 6. Visualization of speaker embeddings for samples from authentic VoxCeleb2 and SynVox2-OHNN-bkg-utt datasets. Each point corresponds to a distinct speaker.

varies between 204k and 871k. The numbers of trials for female and male speakers was 12,882 and 24,806, respectively.

4.2. Results

Do SynVox2 datasets protect speaker identity information? Table 1 shows EERs for the unlinkability evaluation. The EER is 2.07% when both enrollment and test utterances come from the authentic VoxCeleb2 database. However, when the test utterances comes from any one of the six anonymized SynVox2 databases, EERs increase to levels in the order of 30%. ASV score distributions in Figure 4 confirm the effect of anonymization. The distributions for positive (same speaker) and negative (different speaker) trials show greater overlap for the "Authentic vs. SynVox2-OHNN-bkg-utt" anonymized setting than for the "Authentic vs. Authentic" baseline. These results confirm improvements[5] to privacy through anonymization, indicating greater difficulty to link authentic and anonymized utterances/voices.

Can SynVox2 datasets be used to train an ASV model? The results of utility assessments are shown in Table 2. The ASV model trained with authentic data gives an EER of 1.33%. Models trained with anonymised data produce EERs of 7% and above, with results for utterance-level embeddings being consistently superior to those for speaker-level embeddings. These observations highlight the importance of protecting intra-speaker variation. The introduction of additive noise through both *aug* and *bkg* approaches is beneficial, though EERs remain substantially higher than that of the baseline

Are the ASV models trained using SynVox2 datasets fair in terms of gender and accent? Figure 5 shows EERs and FDRs for each gender and accent groups. The trends are similar for SynVox2 and VoxCeleb2 databases. For instance, EERs for speakers with Spanish, Indian, and Australian accent are the highest. For gender groups, the EER for female speakers is consistently higher than those for male speakers. We nonetheless acknowledge that both the utility (EERs) and fairness (FDRs) degrade with the use of synthetic data.

Table 2. EERs (%) on official VoxCeleb1 test set achieved by ASV downstream models (ECAPA-TDNN) trained with different datasets. Results in first row are reported using the ASV model trained on authetic VoxCeleb2 dataset to give an indication. Remaining rows were obtained using different SynVox2 datasets.

Training dataset	EER(%)
Authentic	1.33
SynVox2-OHNN-spk	11.40
SynVox2-OHNN-aug-spk	10.67
SynVox2-OHNN-bkg-spk	10.64
SynVox2-OHNN-utt	7.74
SynVox2-OHNN-aug-utt	7.38
SynVox2-OHNN-bkg-utt	7.58

Inter-speaker variation: To shed light upon the impact of inter-speaker variation, we selected 100 speakers at random and plotted their corresponding embeddings using t-SNE plots [28] when extraction is performed using utterances from the Authentic or SynVox2-OHNN-bkg-utt databases. The results shown in Figure 6 show a reduction in inter-speaker variation for the SynVox2 database. This is a likely cause of the degradation to speaker verification performance

5. CONCLUSIONS

We present in this paper our attempt to create privacy-friendly synthetic VoxCeleb2 datasets for ASV training. By employing the OHNN-based speaker anonymization technique, we generate new SynVox2 substitute. The goal is to provide a better balance between privacy protection and utility. We also introduce metrics for evaluation in terms of privacy, utility, and fairness. Results show that anonymization is reasonably successful in protecting speaker identities. However, while the use of utterance-level embeddings and the addition of additive noise is somewhat successful in compensating for reductions to intra-/inter-speaker variation, the EER of a downstream speaker verification system increases from 1.33% to 7%. While this result might seem disappointing, the use of anonymized datasets may be compulsory in some settings and scenarios; increasing privacy legislation might mean that, one day, there is no alternative. We hence expect research in this direction to continue and to attract greater attention in the future. Further work should investigate the reduction in inter-speaker variation stemming from anonymization. Our results show that there is potential for compensation strategies to reduce the gap in utility between authentic and privacy-friendly databases.

Acknowledgement This study is supported by JST CREST Grants (JPMJCR18A6 and JPMJCR20D3), MEXT KAKENHI Grants (21K17775, 21H04906, 21K11951, 22K21319), and the VoicePersonae project (ANR-18-JSTS-0001). Thanks Xingwei Sun for providing DeepFilter-inspired model.

[5] Perfect anonymization corresponds to EERs of 50%, i.e. fully overlapping score distributions.

6. REFERENCES

[1] "General data protection regulation (GDPR)," https://gdpr.eu/what-is-gdpr.

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[3] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 87–102.

[4] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao, "Synface: Face recognition with synthetic data," in *Proc. ICCV*, 2021, pp. 10880–10890.

[5] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer, "Sface: Privacy-friendly and accurate face recognition using synthetic data," in *Proc. IJCB*. IEEE, 2022, pp. 1–11.

[6] Fadi Boutros, Vitomir Struc, Julian Fierrez, and Naser Damer, "Synthetic data for face recognition: Current state and future prospects," *Image and Vision Computing*, p. 104688, 2023.

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[8] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, et al., "The VoicePrivacy 2020 challenge: Results and findings," *Computer Speech & Language*, 2022.

[9] Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Hubert Nourtel, Pierre Champion, Massimiliano Todisco, Emmanuel Vincent, Nicholas Evans, Junichi Yamagishi, and Jean François Bonastre, "The VoicePrivacy 2022 Challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[10] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko, "Language-Independent Speaker Anonymization Approach Using Self-Supervised Pre-Trained Models," in *Proc. Odyssey*, 2022, pp. 279–286.

[11] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko, "Analyzing Language-Independent Speaker Anonymization Framework under Unseen Conditions," in *Proc. Interspeech*, 2022, pp. 4426–4430.

[12] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko, "Language-independent speaker anonymization using orthogonal householder neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (in press)*, 2023.

[13] Sarina Meyer, Pascal Tilli, Pavel Denisov, Florian Lux, Julia Koch, and Ngoc Thang Vu, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 912–919.

[14] Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Amine Hmani, Aymen Mtibaa, Mohammed Ahmed Abdelraheem, Alberto Abad, Francisco Teixeira, Driss Matrouf, Marta Gomez-Barrero, Dijana Petrovska-Delacrétaz, Gérard Chollet, Nicholas Evans, Thomas Schneider, Jean-François Bonastre, Bhiksha Raj, Isabel Trancoso, and Christoph Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019.

[15] Tiago de Freitas Pereira and Sébastien Marcel, "Fairness in biometrics: a figure of merit to assess biometric verification systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 19–29, 2021.

[16] Mariel Estevez and Luciana Ferrer, "Study on the fairness of speaker verification systems across accent and gender groups," in *Proc. ICASSP*, 2023, pp. 1–5.

[17] Kavita Kasi and Stephen A Zahorian, "Yet another algorithm for pitch tracking," in *Proc. ICASSP*, 2002, vol. 1, pp. I–361.

[18] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[19] Alston S Householder, "Unitary triangularization of a nonsymmetric matrix," *Journal of the ACM (JACM)*, vol. 5, no. 4, pp. 339–342, 1958.

[20] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020, pp. 17022–17033.

[21] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[22] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[23] Jixun Yao, Yi Lei, Qing Wang, Pengcheng Guo, Ziqian Ning, Lei Xie, Hai Li, Junhui Liu, and Danming Xie, "Preserving background sound in noise-robust voice conversion via multi-task learning," in *Proc. ICASSP*, 2023, pp. 1–5.

[24] Hendrik Schroter, Alberto N Escalante-B, Tobias Rosenkranz, and Andreas Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *Proc. ICASSP*, 2022, pp. 7407–7411.

[25] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[26] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[27] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A large-scale speaker identification dataset," .

[28] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE.," *Journal of machine learning research*, vol. 9, no. 11, 2008.