Voice Privacy - Leveraging Multi-Scale Blocks with ECAPA-TDNN SE-Res2NeXt Extension for Speaker Anonymization

Razieh Khamsehashari[1], Yamini Sinha[2], Jan Hintz[2], Suhita Ghosh[3], Tim Polzehl[4], Carlos Franzreb[4], Sebastian Stober[3], Ingo Siegert[1]

[1] Quality and Usability, Technical University of Berlin, Germany
[2] Mobile Dialog Systems, Otto von Guericke University Magdeburg, Germany
[3] Artificial Intelligence Lab (AILab), Otto von Guericke University Magdeburg, Germany
[4] Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI)

razieh.khamsehashari@tu-berlin.de
{yamini.sinha,jan.hintz,suhita.ghosh,stober,siegert}@ovgu.de,
{tim.polzehl,carlos.franzreb}@dfki.de

Abstract

This paper presents the ongoing efforts on voice anonymization with the purpose to securely anonymize a speaker's identity in a hotline call scenario. Our hotline seeks out to provide help by remote assessment, treatment and prevention against child sexual abuse in Germany. The presented work originates from the joint contribution to the VoicePrivacy Challenge 2022 and the Symposium on Security and Privacy in Speech Communication in 2022. Having analyzed in depth the results of the first instantiation of the Voice Privacy Challenge in 2020, the current experiments aim to improve the robustness of two distinct components of the challenge baseline. First, we analyze ASR embeddings, in order to present a more precise and resistant representation of the source speech that is used in the challenge baseline GAN. First experiments using wav2vec show promising results. Second, to alleviate modeling and matching of source and target speaker characteristics, we propose to exchange the baseline x-vectors speaker identity features with the more robust ECAPA-TDNN embedding, in order to leverage its higher resolution multi-scale architecture. Also, improving on ECAPA-TDNN, we propose to extend the model architecture by integrating SE-Res2NeXt units, as the expectation that by representing features at various scales using a cutting-edge building block for CNNs, the latter will perform better than the SE-Res2Net block that creates hierarchical residual-like connections within a single residual block, allowing them to represent features at multiple scales. This expands the range of receptive fields for each network layer and depicts multi-scale features at a finer level. Ultimately, when including a more precise speaker identity embedding we expect to reach improvements for future anonymization for various application cases.

Index Terms: VoicePrivacy challenge 2022, Speaker anonymization, Speech recognition

1. Introduction

Privacy-preserving data processing has developed into an active study subject in recent years as a result of the rising need for privacy protection. The European General Data Protection Regulation (GDPR) under EU law and comparable rules in many nations' national laws are two factors contributing to this.

Speech data contains a lot of personal information that may be revealed by listening or by automated systems even if there is no legal definition of privacy [1]. Age, gender, ethnicity, geographic region, physical or emotional condition, political leanings, and religious convictions are a few examples of this. The identity of the speaker can potentially be revealed through speaker recognition technologies. Therefore, it is not surprising that there is a greater interest in creating voice technology privacy preservation solutions. In the present application case, we pursue anonymization in order to hide the identity of a person calling a help hotline for remote assessment, treatment and prevention against child sexual abuse in Germany. Since help-seekers require a high degree of anonymity when self-reporting on individual inclination and preferences, anonymization is a prerequisite to foster engagement and trust towards the therapeutic offer connected to the service.

The VoicePrivacy [2] initiative is a concrete instantiation evolved out of the need to foster and benchmark anonymization performance and standardize the evaluation of these throughout the community. This effort intends to bring together a new community of researchers, engineers, and privacy specialists. As a part of this program, the initial VoicePrivacy challenge was held in 2020 [3].

Anonymization is the process of removing personally identifying information from voice signals while maintaining other features. In contrast to the methodologies mentioned above, it enables the use of the data for supervised machine learning and is simple to incorporate into current systems. Note that the word "anonymization" in the legal profession denotes the accomplishment of this objective. Even if the procedure under consideration failed in this instance, it still pertains to the task at hand. Anonymization entails changing not just the speaker's voice but also other characteristics and moods, the words used in spoken material, and background noises that, when taken into account in conjunction with one another and maybe with outside data, may reveal the speaker's identity.

The VoicePrivacy 2020 Challenge focuses on voice anonymization as the first step in achieving this objective. This involves changing the speaker's voice in order to conceal their identity as much as possible while preserving all other speech features, like traits, states, and spoken contents. Despite the attractiveness of voice anonymization, the amount of privacy protection demanded by these solutions is ambiguous because there is no formal task specification, no formal attack model, and no shared datasets, protocols, or measurements. All of these issues are addressed by the series of VoicePrivacy Challenges.

Obfuscation, encryption, distributed learning, and anonymization are some current methods for protecting speech privacy. The voice signal is suppressed or altered using

obfuscation techniques [4] to the point where no information about the original speaker can be retrieved. However, the derived data utilized for learning (such as model gradients) may still leak information about the original data [5, 6]. Decentralized or federated learning approaches learn models using distributed data without accessing it directly [7].

Noise addition [8], speech transformation [9, 10], voice conversion [11, 12, 13], and disentangled representation learning [14] are methods for voice anonymization.

2. Related Work

This section gives a brief overview of the results from the last Voice Privacy Challenge and the Baseline Provided by the challenge hosts.

2.1. VPC 2020

The first edition of the VPC was held in 2020, offering two baselines to be improved. The primary baseline performs anonymization with x-vectors. x-vectors [15] represent the speaker's individual characteristics. They result from training a time-delay neural network on the speaker classification task. The primary baseline anonymizes utterances by swapping the x-vector from the input utterance with a *pseudo x-vector*, which is the average of multiple x-vectors extracted from a separate pool. These vectors are the farthest away from the input's x-vector according to PLDA distance. Bottleneck features describe the utterance. They result from a time-delay neural network trained on the speech recognition task. Both these features are then concatenated with the vector containing pitch values of each time window. The resulting matrix is fed to a speech synthesis module that computes Mel-filterbanks, which are finally transformed into a speech signal by an NSF model.

Seven teams participated in the challenge [3]. Four of them modified the primary baseline, focusing on how x-vectors are anonymized. Some approaches revolved around optimizing the combination of existing x-vectors, while others attempted to generate new x-vectors with domain-adversarial training [16]. Increasing the size of the speaker pool from which the x-vectors are drawn to create a pseudo x-vector. Another model used a voice indistinguishability metric to select x-vectors from the pool, which they created with the test set instead of a separate dataset, as the baseline does. They also implemented a different approach for speech synthesis, comprising two modules: the end-to-end acoustic model ESPnet [17], which produces Mel-spectrograms from Mel-filterbank features and speaker x-vectors, and a module based on the Griffin-Lim algorithm [18] that transforms the Mel-spectrogram into a speech waveform.

The system was objectively evaluated on several datasets, which include trial and enrollment utterances of speakers. Trial utterances are the ones evaluated, whereas enrollment utterances serve as additional information for the attack system, which attempts to identify the original speaker behind each anonymized trial utterance. The participating anonymization systems are evaluated in two scenarios: one where only trial utterances are anonymized, and one where also enrollment utterances are anonymized, which makes it easier for the attack system. The results show that using x-vectors provides better anonymization than signal processing. However, no system was superior in both evaluation scenarios. None of the proposed systems achieved an equal error rate of 50 % on the harder scenario, meaning that they do not provide successful anonymization. On the other hand, many systems achieved successful anonymization in the easier scenario.

The quality of the speech produced by the anonymization systems was evaluated objectively, using the word error rate (WER). All anonymization systems degraded the quality of the original speech, as expected. Anonymization systems based on x-vectors again performed better than systems based on signal processing techniques. However, subjective experiments prove signal processing methods to output anonymized speech with higher naturalness and intelligibility.

2.2. New Baselines

The provided 2022 challenge Baselines follow two distinct approaches. The Baseline B1 focuses on anonymization using x-vectors and neural waveform models. B1.a was the primary baseline of the 2020 VoicePrivacy Challenge[2]. The model features three steps. The input speech is first passed through the feature extractors, obtaining F0, ASR acoustic model (AM) bottleneck features (BN) and x-vectors. The F0 extraction is done by pYAAPT[1]. In the second step, the x-vector is anonymized. This is done by choosing a new target speaker out of a pool of x-vectors. The last step is speech synthesis. Using an acoustic model and a neural waveform model, a speech waveform is produced based on the anonymized x-vector and the original BN and F0 data. The 2022 Version of the challenge [3] introduces a new baseline, B1.b. (see figure 1). This baseline replaces the acoustic- and neural waveform model with a HiFi-GAN [19].

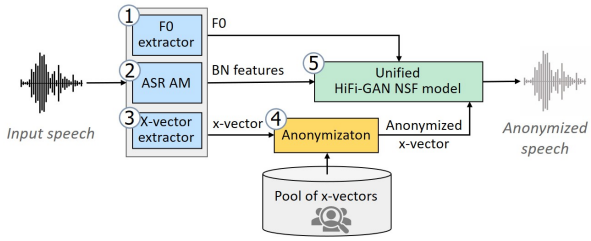

Figure 1: *Baseline B1.b [3]*

Baseline B2, is using the McAdams coefficient, sampling the McAdams coefficient for each source speaker in the evaluation set, from a uniform distribution with min α value 0.5 and max value 0.9.

2.3. Evaluation plan

Verifying that the speaker's anonymity and the content were successfully preserved is one of the most important components of anonymization. Both objective and subjective techniques can be used to accomplish this. Automated speaker verification (ASV) and automatic speech recognition (ASR) technologies can be used to validate, if the speaker's identity was successfully concealed and the content (and intelligibility) were retained [20]. The listening tests using human assessors, who rate content, speaker identification, and intelligibility, on the other hand, might yield a subjective assessment.

The challenge organizers use the Equal Error Rate (EER) as the privacy metric. This metric is applied to the two main evaluation scenarios: **unprotected** and **semi-informed**. In the first, the user is not anonymized, and the attacker has access to the original enrollment data. In the second scenario, the speaker is anonymized and the attacker uses enrollment data

[1] pYAAPT: http://bjbschmitt.github.io/AMFM_decompy/pYAAPT.html

with different pseudo-speakers, anonymized on utterance-level. Word Error Rate (WER) was picked as the primary utility metric and is evaluated on speaker- and utterance-level. The lower the WER the greater the utility [3] Pitch correlation and gain of voice distinctiveness are listed as secondary utility metrics.

The approach for the subjective evaluation is similar to that used for the VoicePrivacy 2020 Challenge [2]. Evaluators will be asked to grade a single original or anonymized trial utterance at a time for assessments of naturalness and intelligibility. The original utterance and an original or anonymized trial utterance obtained from the same or a different speaker will be used in pairs for assessments of speaker verifiably. The comparability of the voices in the enrollment and trial utterances will be graded by the examiners.[3]

3. Our Approach

The starting point of our approach is the baseline B1.b, which introduces the HiFi-GAN. According to the results presented in the evaluation plan [3], the baseline B1.b provides the lowest WER, meaning that the quality of the anonymized speech it outputs is the best among the baselines. The baseline B1.a, which extracts the same features from the input utterance but transforms them into a waveform differently, provides the best EER, meaning it provides the best anonymization. It is to be expected that the system with the best WER does not offer the best EER, as an increase in utility inevitably leads to a decrease in privacy. At least this was one of the findings of the first challenge. Therefore, we focus on improving the model that offers the best WER, namely the baseline B1.b.

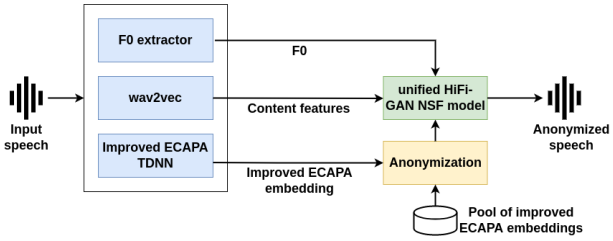

Figure 2: *Adaptation of Baseline B1.b*

Figure 2 shows the modifications we made to improve the baseline. We switched the components 2 and 3 to methods that have proven themselves as state of the art in their field. Component 2, the ASR AM is replaced with wav2vec 2.0 [21], to extract the bottleneck features. The x-vector extractor are replaced by an ECAPA [22] extractor, in the component 3.

3.1. Wav2vec 2.0

In the provided baseline architecture, bottleneck (BN) features are extracted from the final hidden layer of a factorized time delay neural network (TDNN-F) model architecture, which is trained to classify triphones. The BN features are used to encode the linguistic content of the speech. Instead, we use wav2vec 2.0 Base model, a semi-supervised method to extract speech representations directly from the audio. The model is pretrained and fine-tuned with Librispeech. The representations of speech are obtained from the last layer of the fine-tuned model. The model learns contextualized speech representations by randomly masking the feature vectors before feeding the latent speech representations to the transformers.

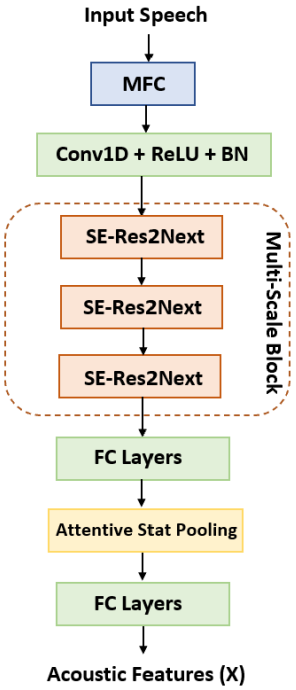

Figure 3: *The diagram of the proposed architecture.*

3.2. Anonymization Process

The neural network is used by the current speaker verification methods to derive speaker representations. The effective x-vector architecture uses TDNN to project variable-length utterances into fixed-length speaker characterization embeddings by applying statistics pooling. We aim at obtaining highly accurate x-vectors on the task of speaker verification and try to improve the performance of the original TDNN architecture [23]. In this regard, our basic architecture follows an established multi-scale architecture, ECAPA-TDNN [22]. Based on current trends, we suggest some improvements to the statistics pooling layer and TDNN design in this study.

Our proposed architecture, shown in Fig. 3, is based on ECAPA-TDNN architecture along with training and inferencing procedures with integrating SE-Res2Next units.

3.2.1. Multi-Scale Backbone Module

With a stack of convolutional layers that automatically learn coarse-to-fine features, multi-scale feature representation has been incorporated into the CNN architectural design from the outset [24]. Shortcut connections to residual networks and the bottleneck module both work well to reduce the number of parameters, successfully addressing the issue of gradient disappearance in deep CNN designs.

By substituting group convolution for standard convolution in order to enable more intricate transformations, ResNeXt-50 [25] added cardinal dimension to the bottleneck module to enable more complex transformations. In order to integrate the multi-scale capacity of the feature representation into the module, Gao et al. [26] replaced the 3×3 convolution with a series of 3×3 convolution with smaller filter groups that are connected in a hierarchical fashion. This might be considered a network inside of a network. The Res2NeXt, therefore, expands the range

of receptive fields for each network layer and depicts multi-scale features at a finer level. So by integrating hierarchical multi-scale feature representation inside the bottleneck module, Res2NeXt-50 [26] enhanced ResNeXt-50 by enabling multi-scale feature representation at both the global and local levels. In order to accomplish a channel-wise dynamic calibration of feature responses and provide a stronger feature representation capability, SE-Res2NeXt-50 integrated the SE block [27].

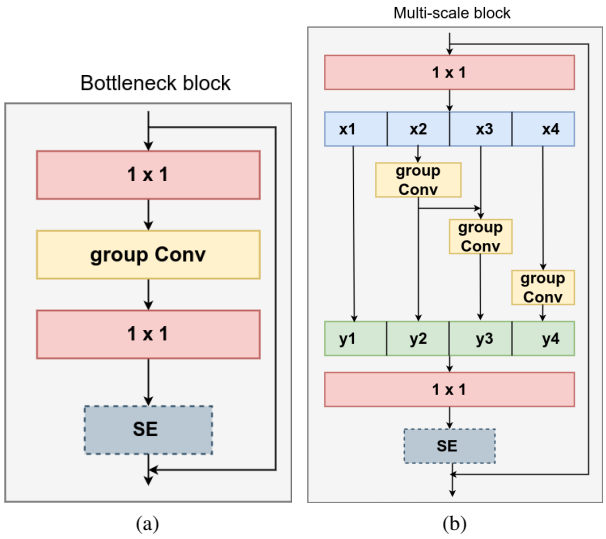

Figure 4: *Comparison of the (a) bottleneck block (SE-ResNeXt) and (b) multi-scale block (SE-Res2NeXt). Group convolution is used by ResNeXt and Res2NeXt with 8 groups.*

3.2.2. Res2NeXt Module

A group of 3×3 filters is substituted in the SE-Res2NeXt in Fig. 4 with smaller groups of filters, while connecting various filter groups in a hierarchical residual-like manner. Following a 3×3 convolution, the input is divided into s feature map subsets, indicated by the symbol X_i, where $i \in \{1, 2, ..., s\}$. In comparison to the input feature map, each feature subset X_i has the same spatial extent but 1/s number of channels. Each X_i has a matching 3×3 convolution, denoted by $K_i()$, with the exception of X_1, which is delivered directly to the output. The feature subset Xi and the output $K_{i-1}()$ from the previous 3×3 convolution are then fed into $K_i()$. Up till all groups have been processed, this procedure is repeated. The output of the module is produced by concatenating the outputs of all groups and passing them to a 1×1 convolution; hence, Y_i can be represented as:

$$Y_i = \begin{cases} X_i & i{=}1 \\ K_i(x_i) & i{=}2 \\ K_i(x_i + y_{i-1}) & 2 < i \leq s \end{cases}$$

4. Experiments

We plan to evaluate our propositions by running a series of experiments to compare against each other. As a reference, we use the Results from the Baseline B1.b. Running the standard settings without any adaption yields a weighted average EER on the original test-set of 3.786%, whereas, for a weighted average EER for the anonymized data is 9.854%. The average WER of the original test-set is 8.475%, while the average of the anonymized data is 7.330%.

We are testing the changes individually, first only replacing the x-vector embeddings with the standard ECAPA embeddings, and then with multiple versions of our extended implementation. We expect this to improve the EER in terms of anonymization since this architecture significantly outperforms state-of-the-art TDNN-based systems in ASV tasks[22]. We also run a test where only the ASR is exchanged with Wav2Vec, to determine which component has the greatest impact. We expect this to have a high impact on the WER, as the Wav2Vec architecture is state of the art in ASR[21]. Subsequently, the combination of these components is also tested. All of these experiments follow the provided evaluation plan (see. section 2.3).

4.1. Training the ECAPA embedding extractor

We evaluate the performance of proposed architecture on the ECAPA embedding on the development part of the VoxCeleb2 dataset with 5994 speakers as training data. For hyperparameter optimization, VoxCeleb1 test set aside as a validation set. All models are trained using a standard Adam optimizer with cyclical learning rates ranging between 1e-8 and 1e-3. Using AAM-softmax with a margin of 0.2 and softmax prescaling of 30 for 4 cycles, all systems are trained. We investigate the 1024 channel convolutional frame layer architecture of the proposed ECAPA-TDNN. The Res2XBlock takes into account various settings for the scale and cardinality dimensions, as indicated in Table 1. The final fully-connected layer has 192 nodes in total.

5. Results

We evaluate the effectiveness of the baseline model using varoius CNN dimensions, including scale and cardinality. As indicated in Table 1, a number of networks are trained and evaluated with different dimensions. The findings of the benchmark experiments [26] imply that scale is an effective dimension to enhance model performance. Moreover, scaling up is more efficient than other dimensions. In our case s = 8 performs better than scale 4 in terms of performance. However, the model with scale 16 is not successful, about which we assume that the size of input is not sufficient enough to support the many scales.

Table 1: *Top-1 test error (%) for the VoxCeleb dataset. The values of the parameters c and s indicate the cardinality and scale, respectively.*

Architecture	Model	Dimensions	$EER(\%)$
ECAPA-TDNN	Res2Net	s8	**1.10**
Extended ECAPA-TDNN	Res2NeXt	$s4 \times c8$	1.19
		$s8 \times c8$	**1.12**
		$s16 \times c16$	1.21

Our proposed architecture, with various values of c=8, 16, does not outperform the baseline in terms of the dimension of cardinality. The cause for this could be the format of the representation. In contrast to [26], ECAPA-TDNN use 1-dimensional TDNN-specific SE-blocks. As a next preliminary experiment, inspired by [28], we integrate a 2D convolutional stem in ECAPA-TDNN baseline. We use 2D convolutions based on Res2NeXt as the foundation for the initial network layers. Because the

Extended ECAPA-TDNN with scale s=8 and cardinality c=8 achieves the best performance in comparison to the other settings, we consider this structure for training the 2D ECAPA-TDNN. We evaluate the performance of proposed architecture with small subset of the development part of the VoxCeleb2 dataset. Table 2 demonstrates how our proposed architecture with 2D data representation outperforms the baseline. While these experiments only produce preliminary results, the direction of our future research is motivated by the indicative success of the 2D convolution stem for the small data set used in these experiments. As Table 2 indicates, using a 2D ECAPA-TDNN with Res2NeXt residual units improves the preliminary EER results by roughly 0.5% absolute.

Table 2: *top-1 test error* $(\%)$ *for the VoxCeleb dataset.*

Architecture	Residual Units	$EER(\%)$
ECAPA-TDNN	Res2Net	13.37
2D ECAPA-TDNN	Res2NeXt	**12.89**

6. Conclusion

This study presents an extended ECAPA-TDNN and 2D ECAPA-TDNN with Res2XBlock integration for speaker verification. In our experiments, extending ECAPA-TDNN with 1-dimensional TDNN-specific SE-blocks does not improve by adding an extra dimension of cardinality. However, changing to 2D ECAPA-TDNN we reach a relative improvement of roughly 0.5% absolute in EER over a strong baseline system applied to the VoxCeleb evaluation set.

In our upcoming study, we will keep evaluating the effectiveness of different types of residual units while integrating them with the 2D ECAPA-TDNN representation with more data utilizing additional datasets and generating extra samples for each utterance by data augmentation.

7. Acknowledgments

This research has been partly funded by the Federal Ministry of Education and Research of Germany in the project Emonymous (project number S21060A) and partly funded by the Volkswagen Foundation in the project AnonymPrevent (AI-based Improvement of Anonymity for Remote Assessment, Treatment and Prevention against Child Sexual Abuse).

8. References

[1] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The gdpr & speech data: Reflections of legal and technology communities, first steps towards a common understanding," *arXiv preprint arXiv:1907.03458*, 2019.

[2] 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 *Proc. Interspeech 2020*, 2020, pp. 1693–1697. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1333

[3] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien *et al.*, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.

[4] A. Cohen-Hadria, M. Cartwright, B. McFee, and J. P. Bello, "Voice anonymization in urban sound recordings," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2019, pp. 1–6.

[5] N. Tomashenko, S. Mdhaffar, M. Tommasi, Y. Estève, and J.-F. Bonastre, "Privacy attacks for automatic speech recognition acoustic models in a federated learning framework," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6972–6976.

[6] S. Mdhaffar, J.-F. Bonastre, M. Tommasi, N. Tomashenko, and Y. Estève, "Retrieving speaker information from personalized acoustic models for speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6767–6771.

[7] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6341–6345.

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

[9] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X. Li, Y. Wang, and Y. Deng, "Voicemask: Anonymize and sanitize voice input on mobile devices," *ArXiv*, vol. abs/1711.11460, 2017.

[10] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker Anonymisation Using the McAdams Coefficient," in *Proc. Interspeech 2021*, 2021, pp. 1099–1103.

[11] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. W. D. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *ArXiv*, vol. abs/1905.13561, 2019.

[12] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, "Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release," *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2020.

[13] B. M. L. Srivastava, N. A. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices for x-vector based speaker anonymization," in *INTERSPEECH*, 2020.

[14] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?" in *Proc. Interspeech 2019*, 2019, pp. 3700–3704.

[15] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[17] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech 2018*, 2018, pp. 2207–2211.

[18] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[20] I.-C. Yoo, K. Lee, S. Leem, H. Oh, B. Ko, and D. Yook, "Speaker anonymization for personal information protection using voice conversion techniques," *IEEE Access*, vol. 8, pp. 198 637–198 645, 2020.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[25] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[26] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[28] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating frequency translational invariance in TDNNs and frequency positional information in 2d ResNets to enhance speaker verification," in *Interspeech 2021*. ISCA, aug 2021. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2021-1570