INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Two-Stage Voice Anonymization for Enhanced Privacy

*Francesco Nespoli*[1,3], *Daniel Barreda*[1], *Jöerg Bitzer*[2], *Patrick A. Naylor*[3]

[1]Microsoft, London, UK
[2]Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg, Germany
[3]Imperial College, London, UK

fnespoli@microsoft.com, daniel.almendrobarreda@microsoft.com,
joerg.bitzer@idmt.fraunhofer.de, p.naylor@imperial.ac.uk

## Abstract

In recent years, the need for privacy preservation when manipulating or storing personal data, including speech , has become a major issue. In this paper, we present a system addressing the speaker-level anonymization problem. We propose and evaluate a two-stage anonymization pipeline exploiting a state-of-the-art anonymization model described in the Voice Privacy Challenge 2022 in combination with a zero-shot voice conversion architecture able to capture speaker characteristics from a few seconds of speech. We show this architecture can lead to strong privacy preservation while preserving pitch information. Finally, we propose a new compressed metric to evaluate anonymization systems in privacy scenarios with different constraints on privacy and utility.

**Index Terms**: privacy, speaker anonymization, speech recognition, speaker recognition, voice conversion

## 1. Introduction

Speech is a primary modality for humans to communicate. In recent years, speech technologies enabled effective human-machine interaction making it possible to control systems and devices with speech [1]. If, on one side, voice assistants and smart speakers facilitate daily tasks [2], then, on the other, these technologies raise privacy concerns for the public and policy makers [3]. Such concerns come from the fact that speech contains significant personal identifiable information (PII) both in the semantic and acoustic domain. Specifically, personal identifiers such as full name, social security number or geographical positioning can alone allow speaker identification. Moreover, voice characteristics such as prosody, speaking rate, accent and intonation inherently contain a variety of PII such as personality, physical characteristics, emotional state, age and gender that can be identified [4] and therefore used for malicious privacy attacks. In this context, suppressing PII in speech signals would improve privacy. Considering a situation in which personal identifiers are not present or has been obfuscated, initial acoustic privacy protection approaches explored several research directions such as extracting privacy-preserving features [5], working with encrypted speech signals [6], learning adversarial features [7], or performing federated learning [8]. However, feature or model-level privacy protection have limitations. Privacy preserving features can, in principle, be used for any downstream application but there is no guarantee they retain any useful information to efficiently address the new task. Encryption, although allowing data manipulation in the encrypted domain, introduces significant computational overheads. Adversarial features have been reported to increase privacy in closed-set classification scenarios but lack in generalization [7]. Federated models, despite avoiding direct access to the data, can leak original data information through the higher level representation used for learning (e.g., local gradients) [9]. Therefore, state-of-the-art anonymization systems are based on the idea of disentangling the speaker identity information from the linguistic and prosodic content thus producing synthetic utterances in which the speaker identity has been altered [10], [11], [12] while other speech characteristics, possibly important for downstream tasks, are preserved. Although this general approach has been shown to be effective [10], the potential for concrete attacks is quite large. Specifically, speaker information can leak into linguistic and prosodic features [13], propagate to the modified speech and be used by an attacker to identify the speaker.

One effective technique for speech anonymization employs automatic speech recognition (ASR) to transcribe speech followed by a text-to-speech (TTS) system that re-synthesizes audio signals from text transcriptions. This ASR+TTS method protects the vocal characteristics of the speaker's identity but completely destroys the original prosodic attributes such as intonation, stress and rhythm [14]. Moreover, the incorrect linguistic content induced by any ASR error and the very limited variability of TTS-synthesized speech outputs, mainly due to few available voices, lead to poor results when applying ASR+TTS on downstream tasks [15], [16]. Another direction recently investigated by [17] and [18], demonstrated that a cascade of signal-based anonymization modules results in higher anonymization scores compared to single-stage processes specifically in the case of decryption attacks [18]. Based on these findings, we investigate the anonymization capabilities of a fully-neural pipeline based on voice conversion (VC) targeting higher privacy enforcement in the context of voice privacy protection. The system combines, in a two-step procedure, a zero-shot voice conversion (ZS-VC) block and the baseline system of the Voice Privacy Challenge [10]. Privacy and utility scores have been measured as the equal error rate (EER) of an automatic speaker verification (ASV) system and the word error rate (WER) of an ASR model respectively. Furthermore, we computed a minimal secondary evaluation metric, specifically a lower-bound for the pitch correlation between original and synthesised utterances. Both the anonymization and scoring pipelines follow the framework of [19] in terms of datasets allowed for training and testing the anonymization models.

## 2. Proposed Model

In [18], the authors show a combination of multiple speech modifications enhances privacy. Based on those results, we propose a cascade of two deep learning-based voice conversion systems which target specific situations with different anonymazation requirements. Figure 1 provides a schematic overview of the components we used in the two-stage system. The


10.21437/Interspeech.2023-1341

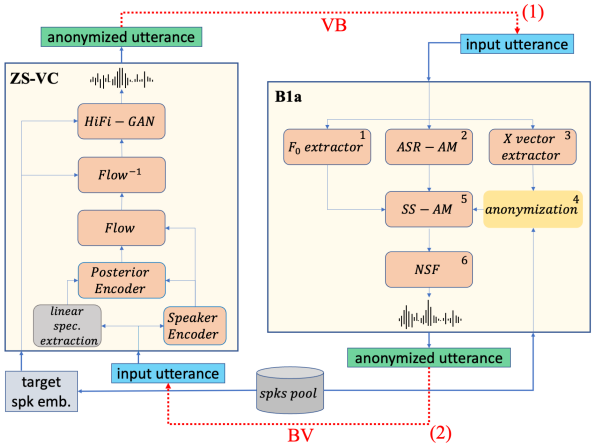


Figure 1: *Overview of voice anonymization components: zero-shot voice conversion (*ZS-VC*) and baseline system (*B1a*). The red dotted lines* (1) *and* (2)*, are mutually exclusive and refer, in the Experiment section, to system VB and BV respectively.*

first component ZS-VC is a zero-shot any-to-any voice conversion system [20] whereas the second is the baseline B1a from [10]. Speech anonymization is achieved by altering the original speaker identity. To do this, both components rely on a set of external speakers (*spks pool*, Fig. 1) utterances. Speakers in *spks pool* are not included in the set to be anonymized: they are used as targets to alter the original speaker identity. Target speaker selection is handled differently by the two components and this is described in detail in the next sections. The red dotted lines in Fig 1 are mutually exclusive: when (1) is active, ZS-VC is applied first; when (2) is operating, B1a is the first stage of the two-stage anonymization process.

### 2.1. Zero-Shot Voice Conversion

This stage exploits YourTTS [20] a multi-speaker TTS architecture with multilingual capabilities based on variational inference and adversarial learning [21]. The model has five main components: text encoder, posterior encoder, alignment stage, an invertible flow-based decoder and a vocoder [22]. At training time, the posterior encoder receives as input the linear spectrogram and outputs a latent representation $Z$. This is used by the vocoder and by the flow-based decoder whose output $Z_{\mathrm{p}}$ is then aligned with the text encoder representation with monotonic alignment search (MAS) [23]. At inference time, the text encoder generates the alignment $Z_{\mathrm{p}}$ which is then used by the inverted flow decoder to produce $Z$ which, when input into the vocoder, generates the audio. Due to the fact that all five main components are conditioned on the speaker embedding, this same architecture can be used for voice conversion. In this case, given that at inference time we have an utterance instead of text, $Z_{\mathrm{p}}$ is obtained from the output of the chain posterior encoder and direct flow, which are conditioned on the original speaker embedding. At this point, by conditioning the inverted flow and the vocoder on the target speaker embedding, we can achieve effective VC. Here, the speaker encoder was implemented with the H/ASP architecture [24] based on ResNet-34. We explored two modalities for target speaker selection addressing different privacy-to-utility ratios. We hypothesize random target selection with gender preservation might allow ZS-VC to generate higher quality speech as observed in other voice conversion approaches [25]. This could lead to fewer ASR errors and therefore lower WER. Contrarily, we argue that relaxing the gender constrain might result in higher EERs due to the greater mismatch between target and original utterance which might heavily affect the ASV embedding model.

### 2.2. X-vector Speaker Anonymization

This step involves state-of-the-art techniques for extracting speaker identity, linguistic and prosodic information from the original speech signal and re-synthesizing a new utterance after modifying the speaker content. Blocks 1-6 in Fig. 1 come from [10]. They consist of a first stage for speaker and acoustics information extraction (blocks 1-3) and a second re-synthesis stage (blocks 4-6), with the idea of combining original speaker acoustics with a new speaker embedding for efficient voice anonymization. The first stage comprises a fundamental frequency ($\mathrm{F}_0$) extractor (block 1) based on [26] a time-delay neural network (TDNN) ASR to extract bottle neck (BN) features modeling speech acoustics (block 2) and a speaker encoder (block 3) computing x-vectors [27] with a second factorized-TDNN architecture. Anonymized speech is synthesised in two steps. First, a speech synthesis acoustic model (SS AM - Block 5) computes Mel-filterbank features given $\mathrm{F}_0$, BN features and the anonymized x-vector. Second, a neural source-filter (NSF - block 6) generates a waveform from $\mathrm{F}_0$, anonymized x-vector and Mel-filterbank. In this case, the anonymized x-vector has been obtained using the following distance-based rule. First, a set of *N* closest speaker embeddings is selected for each original speaker representation. Then, a subset *M* is randomly sampled from *N* and averaged to produce the target speaker embedding (*N*=200, *M*=20).

## 3. Experiments

### 3.1. Data

In the experiments we used the same training and testing subsets listed in [19] with the same constraints. Specifically, ZS-VC was trained first on *LibriTTS-100-clean* [28] for $7 \times 10^5$ steps with batch size 52 and then fine-tuned on *LibriTTS-500-others* for further $3 \times 10^5$ epochs. Both *LibriTTS* subsets were resampled at 16kHz and RMS normalized with target level -27dB [20]. Furthermore, [29] was used to remove silences. The speakers pool in Fig. 1 coincides with the *LibriTTS-train-other-500* subset and it was used to extract target speaker representations. First, a speaker embedding was computed for each utterance in the pool. Second, the target speaker representation was calculated by averaging all the speaker embeddings for each speaker. This process was computed for both H/ASP in ZS-VC and the TDNN x-vector extractor in B1a. Differently from [10], we employed pre-training for both the ASV and ASR models. Specifically, instead of training the models from scratch only on the anonymized *LibriSpeech-360-clean* subset [30] ($LS^{360}_{anon}$), we first pre-trained the ASV speaker embedding model on VoxCeleb 1,2 [31] subsets following the recommendations of [32] and then fine-tuned on $LS^{360}_{anon}$ for 20 more epochs. The ASR model was pre-trained on *LibriTTS-100-clean* and *LibriTTS-500-others* for 60 epochs and fine-tuned $LS^{360}_{anon}$ for 30 more epochs. We used Adam optimizer with initial learning rate (*lr*) 0.001, *lr*-scheduler from [33] and batch size of 128. Pre-trainig and fine tuning on $LS^{360}_{anon}$ was conducted separately for each anonymizaton condition for both ASR and ASV. Finally, the anonymizaton results were tested on *clean* development and test sets of *LibriSpeech* and on a subset of VCTK [34] obtained following the same procedures as in [19].

Table 1: *Primary anonymization metric (*EER*, higher is better) results for development (light gray) and test (dark gray) sets.* Orig*: Non-anonymized original data.* B1a*: baseline from [10].* $V_R$*:* ZS-VC *with random target speaker selection.* $V_{GP}$*:* ZS-VC *with random target speaker selection and gender preservation.* $BV$*:* B1a *followed by* ZS-VC*.* $VB$*:* ZS-VC *followed by* B1a*.* VCTK *subsets comprise a set of utterances with same (*comm.*) and different (*diff.*) linguistic content across the speakers.*

| **Set** | **Gender** | **Weight** | **EER[%]** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Orig | B1a | $V_R$ | $V_{GP}$ | $BV$ | $VB$ |
| Libri-dev | female | 0.25 | 4.12 | 14.6 | 21.6 | 22.3 | 32.95 | 52.6 |
| | male | 0.25 | 0.93 | 10.2 | 16.2 | 18.3 | 34.6 | 43.9 |
| Vctk-dev | female | 0.20 | 0.84 | 9.1 | 26.9 | 30.4 | 35.1 | 53.3 |
| diff. | male | 0.20 | 0.64 | 8.1 | 27.9 | 30.0 | 19.52 | 41.4 |
| Vctk-dev | female | 0.05 | 0.87 | 10.2 | 26.8 | 28.7 | 25.6 | 42.4 |
| comm. | male | 0.05 | 0.58 | 9.7 | 30.5 | 29.1 | 32.48 | 51.3 |
| $\boldsymbol{Avg^W dev}$ | | | 1.63 | 10.64 | 23.3 | 25.1 | 30.7 | 47.8 |
| Libri-test | female | 0.25 | 2.55 | 12.7 | 18.8 | 14.8 | 31.8 | 46.0 |
| | male | 0.25 | 0.43 | 10.5 | 11.4 | 11.8 | 30.0 | 48.5 |
| Vctk-test | female | 0.20 | 1.59 | 14.7 | 30.4 | 32.3 | 29.8 | 35.2 |
| diff. | male | 0.20 | 0.97 | 12.2 | 27.1 | 28.3 | 32.9 | 39.0 |
| Vctk-test | female | 0.05 | 0.34 | 13.8 | 29.5 | 30.9 | 27.8 | 41.6 |
| comm. | male | 0.05 | 0.28 | 7.1 | 27.4 | 29.4 | 31.3 | 48.9 |
| $\boldsymbol{Avg^W test}$ | | | 1.29 | 12.2 | 21.9 | 21.8 | 31.0 | 43.0 |

### 3.2. Metrics

The two primary scores to evaluate an anonimization system are the EER and the WER [19]. The first assesses the privacy protection capability of the anonymization pipeline whereas the second measures the utility of the anonymized speech to perform downstream tasks. Given a generic biometric authentication system G, $R^{\mathrm{G}}_{\mathrm{fa}}(\theta)$ and $R^{\mathrm{G}}_{\mathrm{fr}}(\theta)$ are the false acceptance and false rejection rates at a given decision threshold $\theta$. The EER corresponds to the rate at which $R^{\mathrm{G}}_{\mathrm{fa}}(\theta) = R^{\mathrm{G}}_{\mathrm{fr}}(\theta)$. The WER is calculated from the ASR output transcription as

$$\mathrm{WER} = \frac{N_{\mathrm{sub}} + N_{\mathrm{ins}} + N_{\mathrm{del}}}{N_{\mathrm{tok}}}$$

where $N_{\mathrm{sub}}, N_{\mathrm{ins}}, N_{\mathrm{del}}$ are the number of substitutions, insertions and deletions in the ASR output and $N_{\mathrm{tok}}$ is the number of tokens in the reference transcript. Moreover, it has been shown that when manipulating speech data to enhance privacy (higher EER), the utility of the anonymized signals drops (higher WER) [10]. Therefore, the choice of the most appropriate anonymization system heavily depends on the specific application: when privacy is paramount, large WER degradation might be tolerated in favour of privacy. However, for other applications when even small WER variations heavily impact the performance, an anonymization system with a limited WER deterioration might be the best choice despite its lower privacy protection capability. Here, we propose the privacy-to-utility trade off ($\mathrm{PU}_{\mathrm{tr}}$), a compressed metric combining the primary anonymization metric and primary utility score, designed to evaluate the anonymization system at different operating points. Given $\mathrm{WER}_{\mathrm{i}}, \mathrm{EER}_{\mathrm{i}} \in (0, 1], i \in \{0, 1\}$ denotes the metrics calculated on original ($i = 0$) or anonymized ($i = 1$) utterances. We define

$$\mathrm{PU}_{\mathrm{tr}} = \lambda \frac{\log\left(1 + \frac{\mathrm{WER}_1}{\mathrm{WER}_0}\right)}{\log\left(1 + \frac{1}{\mathrm{WER}_0}\right)} - (1 - \lambda) \frac{\log\left(1 + \frac{\mathrm{EER}_1}{\mathrm{EER}_0}\right)}{\log\left(1 + \frac{1}{\mathrm{EER}_0}\right)}$$

where $\lambda \in [0, 1]$ controls the trade off between WER and EER. $\mathrm{PU}_{\mathrm{tr}} \in [-1, 1]$ and a lower value indicates a more favorable trade-off at a specific operational point $\lambda$. Furthermore, we calculated $\rho^{F_0}$, also called pitch correlation metric, to assesses intonation preservation of the anonymization process. First, $\mathrm{F}_0$ was extracted from each utterance using [26]. Then, $\rho^{\mathrm{F}_0}$ was computed as the Pearson correlation between the $\mathrm{F}_0$ of original and anonymized speech.

### 3.3. Scoring systems

In our experiments, the attacker is defined as *Semi-Informed* [10]. This means it has full knowledge of the anonymization system, but can not access the mapping between original and anonymized speakers. In this condition, the attacker anonymizes the training set *LibriSpeech-360-clean* by selecting a different target speaker for each utterance in the dataset and fine-tunes the ASV model on these data. The speaker verification model we used is an x-vector TDNN-based speaker encoder coupled with a probabilistic linear discriminant analysis (PLDA) classifier [35]. The system uses the same verification files in Kaldi format as [19] in which enrollment and trial utterances have been anonymized with different target speakers. We employed an ASR system based on a transformer acoustic model encoder and a joint transformer decoder with connectionist temporal classification (CTC) [36], with decoding stage integrating also CTC probabilities. The ASR and ASV models were implemented with [32].

## 4. Results and Discussion

### 4.1. Primary Anonymization Metric

Privacy protection capabilities of each model have been evaluated with the EER metric. Results for each system have been reported in Table 1. Here, ZS-VC on its own provides higher protection compared to the baseline system scoring a 2-fold increment in the EER. Moreover, the concatenation of the two anonymization systems greatly enhances privacy protection of stand-alone pipelines with $VB$ scoring close to perfect anonymization (EER=50%) for many test and development sets. Finally, preserving the gender information in the original-to-target mapping for ZS-VC appears to lead to greater

anonymization results ($V_{GP}$ column, Table 1) when compared to random target selection ($V_R$ column, Table 1).

### 4.2. Primary Utility Metric

We assessed linguistic information preservation by calculating the WER with the ASR model fine-tuned separately for each anonymization condition. Table 2 summarises the scores for each anonymization stage. As expected, all anonymization processes degrade ASR performance. However, in the case of ZS-VC, preserving the gender of the original speakers after the conversion, produces a 14.7% WER reduction (on average between development and test sets) when compared to random gender mapping. Moreover, two-stage processing ($BV$ and $VB$ in Table 2), although degrading WER with respect to single stage processes, can achieve better utility scores by employing ZS-VC as the second stage of the anonymization pipeline ($BV$).

Table 2: *Primary utility metric (WER, lower is better).*

| **Set** | **WER[%]** | | | | | |
|---|---|---|---|---|---|---|
| | Orig. | $B_{1a}$ | $V_R$ | $V_{GP}$ | $BV$ | $VB$ |
| Libri-dev | 2.33 | 2.77 | 4.22 | 4.04 | 4.72 | 6.51 |
| Vctk-dev | 8.21 | 9.59 | 14.58 | 13.95 | 16.24 | 19.39 |
| ***Avg dev*** | 5.27 | 6.18 | 9.4 | 9.0 | 10.5 | 12.95 |
| Libri-test | 2.47 | 2.85 | 3.84 | 3.78 | 4.17 | 7.19 |
| Vctk-test | 7.63 | 9.39 | 13.65 | 9.86 | 9.63 | 18.12 |
| ***Avg test*** | 5.1 | 6.12 | 8.75 | 6.82 | 6.92 | 12.7 |

### 4.3. Privacy Utility Trade-Off

One of the outstanding problems of privacy evaluation is that two interplaying quantities need to be evaluated at the same time. Specifically, when privacy improves (higher EER), it comes at the cost of WER degradation (Table 1,2). Here, we suggest to combine these two measures into a compressed metric $\mathrm{PU_{tr}}$. Figure 2 shows $\mathrm{PU_{tr}}$ results for each anomymization system and for different values of $\lambda$. In privacy applications where a low WER is crucial ($\lambda = 0.9$) B1a results to be the best choice. However, for applications in which users require stronger privacy and can tolerate higher WER increments (lower values of $\lambda$) ZS-VC and two-stage processes become a better choice in terms of $\mathrm{PU_{tr}}$.

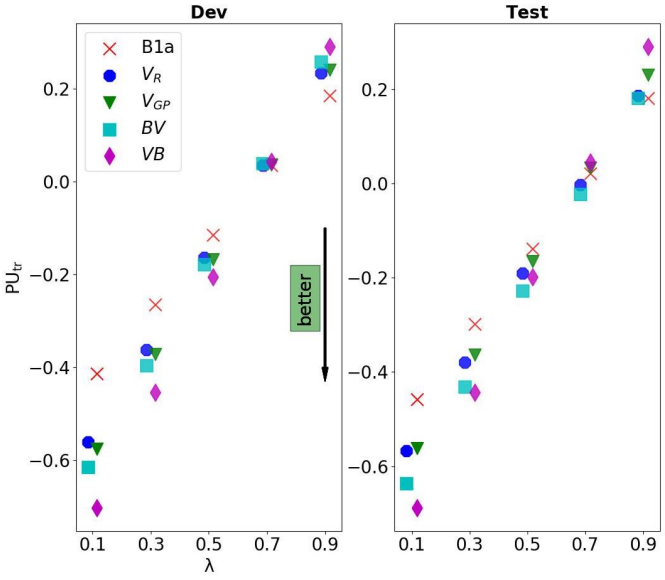


Figure 2: *Privacy-to-utility trade-off for development (***Dev***) and test (***Test***) sets evaluated for* $\lambda \in [0.1, 0.3, 0.5, 0.7, 0.9]$.

### 4.4. Intonation Preservation

We tested intonation preservation by calculating $\rho^{\mathrm{F_0}}$ from original and anonymized utterances. Although $\mathrm{F_0}$ has been proven to incorporate speaker information [37], all ZS-VC models display high $\rho^{\mathrm{F_0}}$ while improving the EER, showing that it is possible to achieve strong anonymization while maintaining prosody information (Table 3). Specifically, two-stage methods can achieve EERs comparable with state-of-the-art ASR+TTS anonymization systems [14] while greatly improving $\rho^{\mathrm{F_0}}$. This is particularly important when ASR models trained on anonymized data need to be applied to real speech [15], [16].

Table 3: *Pitch correlation ($\rho^{\mathrm{F_0}}$, higher is better). Average results computed with the same weights as in Table 1.*

| **Set** | **Gnd** | $\rho^{\mathrm{F_0}}$ | | | | |
|---|---|---|---|---|---|---|
| | | $B_{1a}$ | $V_R$ | $V_{GP}$ | $BV$ | $VB$ |
| Libri-dev | F | 0.77 | 0.8 | 0.81 | 0.8 | 0.75 |
| | M | 0.73 | 0.78 | 0.75 | 0.77 | 0.75 |
| Vctk-dev | F | 0.84 | 0.81 | 0.85 | 0.82 | 0.79 |
| comm. | M | 0.78 | 0.78 | 0.77 | 0.78 | 0.74 |
| Vctk-dev | F | 0.79 | 0.77 | 0.81 | 0.79 | 0.76 |
| diff. | M | 0.72 | 0.75 | 0.74 | 0.75 | 0.71 |
| **Avg dev** | | 0.78 | 0.79 | 0.79 | 0.79 | 0.76 |
| Libri-test | F | 0.77 | 0.8 | 0.85 | 0.81 | 0.76 |
| | M | 0.69 | 0.71 | 0.72 | 0.73 | 0.67 |
| Vctk-test | F | 0.84 | 0.83 | 0.86 | 0.82 | 0.81 |
| comm. | M | 0.79 | 0.78 | 0.78 | 0.78 | 0.76 |
| Vctk-test | F | 0.79 | 0.8 | 0.83 | 0.79 | 0.77 |
| diff. | M | 0.70 | 0.74 | 0.72 | 0.73 | 0.71 |
| **Avg test** | | 0.77 | 0.78 | 0.80 | 0.78 | 0.75 |

## 5. Conclusions

In this paper we present a novel anonymization pipeline cascading two fully-neural anonymization systems. This was achieved using a combination of zero-shot voice conversion and a state-of-the-art anonymization model. Results show that two-stage processes can preserve prosodic information while concealing speaker identity with EER scores comparable with ASR+TTS methods. We introduce $\mathrm{PU_{tr}}$, a compressed metric to evaluate the anonymization models for privacy applications with different WER and EER constrains and we showed that with this new score, the choice of the best anonymization technique can be tuned with the $\lambda$ parameter according to specific anonymization requirements in terms of WER and EER.

## 6. Acknowledgements

This project was funded by the European Union's Horizon 2020 program under the Marie Skłodowska-Curie grant No 956369.

# 7. References


[1] M. Katore and M. R. Bachute, "Speech based human machine interaction system for home automation," in *IBSS*, 2015, pp. 1–6.

[2] M. Vacher, D. Istrate, F. Portet, T. Joubert, T. Chevalier, S. Smidtas, B. Meillon, B. Lecouteux, M. Sehili, P. Chahuara, and S. Méniard, "The sweet-home project: Audio technology in smart homes to improve well-being and reliance," in *Annual Int. Conf. of the IEEE Eng. in Medicine and Biology Society*, 2011, pp. 5291–5294.

[3] A. Arabo, I. Brown, and F. El-Moussa, "Privacy in the age of mobility and smart devices in smart homes," in *2012 Int. Conf. on Privacy, Security, Risk and Trust and 2012 Int. Conf. on Social Computing*, 2012, pp. 819–826.

[4] J. L. Kröger, O. H.-M. Lutz, and P. Raschke, *Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference*. Springer International Publishing, 2020, pp. 242–258. [Online]. Available: https://doi.org/10.1007/978-3-030-42504-3_16

[5] A. Nelus, J. Ebbers, R. Haeb-Umbach, and R. Martin, "Privacy-Preserving Variational Information Feature Extraction for Domestic Activity Monitoring versus Speaker Identification," in *Proc. Interspeech*, 2019, pp. 3710–3714.

[6] P. Thaine and G. Penn, "Extracting mel-frequency and bark-frequency cepstral coefficients from encrypted signals," in *Proc. Interspeech*, 2019.

[7] 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, pp. 3700–3704.

[8] M. Hao, H. Li, G. Xu, S. Liu, and H. Yang, "Towards efficient and privacy-preserving federated deep learning," in *ICC 2019 - 2019 IEEE Int. Conf. on Communications*, 2019, pp. 1–6.

[9] Z. Zhao, M. Luo, and W. Ding, "Deep leakage from model in federated learning," *arXiv preprint arXiv:2206.04887*, 2022.

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

[11] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices for x-vector based speaker anonymization," 2020. [Online]. Available: https://arxiv.org/abs/2005.08601

[12] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, "Differentially private speaker anonymization," 2022. [Online]. Available: https://arxiv.org/abs/2202.11823

[13] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," 2019. [Online]. Available: https://arxiv.org/abs/1905.13561

[14] S. Meyer, F. Lux, P. Denisov, J. Koch, P. Tilli, and N. T. Vu, "Speaker anonymization with phonetic intermediate representations," 2022. [Online]. Available: https://arxiv.org/abs/2207.04834

[15] Z. Chen, A. Rosenberg, Y. Zhang, G. Wang, B. Ramabhadran, and P. J. Moreno, "Improving speech recognition using gan-based speech synthesis and contrastive unspoken text selection," in *Proc. Interspeech*, 2020.

[16] J. Li, R. Gadde, B. Ginsburg, and V. Lavrukhin, "Training neural speech recognition systems with synthetic speech augmentation," 2018. [Online]. Available: https://arxiv.org/abs/1811.00707

[17] H. Kai, S. Takamichi, S. Shiota, and H. Kiya, "Lightweight voice anonymization based on data-driven optimization of cascaded voice modification modules," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 560–566.

[18] K. Hiroto, T. Shinnosuke, S. Sayaka, and K. Hitoshi, "Robustness of Signal Processing-Based Pseudonymization Method Against Decryption Attack," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 287–293.

[19] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, "The voice privacy 2022 challenge evaluation plan," 2022. [Online]. Available: https://arxiv.org/abs/2203.12468

[20] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *Int. Conf. on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[21] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Int. Conf. on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[22] 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.

[23] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[24] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," 2020. [Online]. Available: https://arxiv.org/abs/2009.14153

[25] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[26] D. J. Hirst, "A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation," in *16th Int. Congress of Phonetic Sciences*, Saarbrücken, Germany, Aug. 2007. [Online]. Available: https://hal.science/hal-03625441

[27] 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.

[28] 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 *Proc. Interspeech*, 2019, pp. 1526–1530.

[29] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2021.

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

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[32] 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.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[34] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[35] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision – ECCV*. Springer Berlin Heidelberg, 2006, pp. 531–542.

[36] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd Int. Conf. on Machine learning*, 2006, pp. 369–376.

[37] C. O. Mawalim, S. Okada, and M. Unoki, "Speaker anonymization by pitch shifting based on time-scale modification," in *to appear in the 2nd Symposium on Security and Privacy in Speech Communication joined with 2nd VoicePrivacy Challenge (SPSC 2022)*, 2022.