Towards Generalizable Deep Speech Anonymization

Broukhim, Aaron
Department of Computer Science
UC San Diego La Jolla, CA 92093
`aabroukh@ucsd.edu`

Novack, Zachary
Department of Computer Science
UC San Diego La Jolla, CA 92093
`znovack@ucsd.edu`

Abstract

Speech Anonymization (SA) has arisen as a key domain of interest within the audio privacy world, as the proliferation of voice assistants like Alexa and Siri poses an increased risk of compromising the privacy of everyday users. SA itself, as an offshoot of traditional speech-to-text, poses the difficult problem of protecting user privacy while simultaneously maintaining standard speech-to-text ability. Despite recent interest from the machine learning community, the current state-of-the-art SA systems are language-specific and rely on anonymization methods that could be exploited by malicious actors. In order to combat these issues, we propose a series of ablations that a) leverage a Generative Adversarial Network (GAN) [Goodfellow et al., 2014] to learn a distribution over language-agnostic speaker embeddings and b) use a modified loss to bias the generative model towards an anonymized distribution. We find that though our method displays subpar anonymization compared to standard baselines, our model outperforms other language-agnostic methods for anonymization and achieves SOTA performance in speech recognition. Our code is available at `https://github.com/aabroukh/anonima-voce`.

1 Introduction

Problem Definition. Speech anonymization (SA), which refers to the task of removing the identifiable user characteristics of an audio input while preserving semantic and prosodic information, has seen much attention in the past few years. Despite the recent advances and entire challenges dedicated to this problem [Tomashenko et al., 2020, 2022], multiple issues have arisen endemic to the field. Namely, there is tradeoff between performance and generality, as the best performing systems require careful tuning and are language-specific. Additionally, no attention has been paid to learning an anonymization mapping directly, as researchers instead rely upon simple (and often deterministic) operations to anonymize the speaker embedding. Thus, our goal in this project is to leverage the recent advances in deep generative models to create a SA system that is language-agnostic and directly learns the anonymization transformation, thereby creating a general-purpose and robust SA system.

Problem Significance. As work in the Internet-of-Things (IoT) domain encroaches into people's everyday lives, there is a groing concern over the privacy issues that such IoT systems may create [Jin et al., 2022]. Namely, as voice assistants become more prevalent and voice recognition models improve, there is a need for audio data to be anonymized as early in the data pipeline as possible. Working towards a generalizable solution agnostic of language could help address this issue, as IoT companies across the world would be able to standardize audio privacy practice with such a system [Miao et al., 2022]. Additionally, working towards a generative anonymization system reduces the attack risk of such deployed systems, as current SA systems can be quickly compromised if their pool of speaker data is ever leaked.

Preprint. Under review.

Technical Challenge. With regards to data gathering, we anticipate issues may arise from attempting to collect and clean a wide-array of different audio-based datasets. For the engineering and design aspects, our proposed solutions may require clever fine-tuning of both training hyperparameters and overall model architecture, and to make sure that training such models is tractable within the current time frame of the project. For evaluation, running speaker verification and speech recognition on the main benchmarks is an incredibly resource intensive task, requiring constant processing of high-dimensional audio data.

State-of-the-Art. The current state-of-the-art SA systems primarily come from the VoicePrivacy 2020 [Tomashenko et al., 2020] and 2022 [Tomashenko et al., 2022] challenges. Meyer et al. [2022] currently offers the best performance for deep SA frameworks, and uses a Wasserstein GAN to extract speaker-identification features from the audio, achieving nearly perfect anonymization on VoicePrivacy benchmarks. However, Meyer et al. [2022] specifically does not preserve pitch information in its conversion, which may severely limit its effectiveness on tonal languages like Mandarin Chinese. Khamsehashari et al. [2022] and Yao et al. [2022] both also provide SA systems that achieve good results on the english benchmarks, but are language-specific and propose relatively simple anonymization methods (by using a random speaker embedding from some pool of speakers or using an average speaker embedding respectively).

To our knowledge, Miao et al. [2022] represents the state-of-the-art for general purpose language-agnostic SA. However, the accuracy and privacy performance of their proposed method falls far short of the language-specific methods [Meyer et al., 2022, Khamsehashari et al., 2022, Yao et al., 2022]. As a whole, there is a clear tradeoff between flexibility and performance among state-of-the-art systems, and additionally a still present reliance on the “pool of speakers" approach to anonymization, which if compromised could seriously damage the efficacy of the SA system.

Contributions.

- We adapt current work for language-agnostic voice anonymization with the current SOTA anonymization methods in English to propose a novel SA architecture that is langauge-agnostic yet incorporates a more sophisticated anonymization process.
- We propose a novel method to optimize voice anonymization systems end-to-end such that the anonymization method itself is optimized by the internal DGM (GAN) by minimizing similarity between same-speaker embeddings.
- We show that though both of the above systems perform below SOTA in terms of anonymization performance, our methods outperform the SOTA *language-agnostic* baselines, and additionally find the emergent property that GAN-based speaker anonymization systems work remarkably well at *speech enhancement* (i.e. improving speech-to-text recognition), achieving SOTA speech recognition performance compared to other anonymization systems.

2 Related Work

VoicePrivacy Baselines Tomashenko et al. [2020, 2022] Put forth a large-scale competition at INTERSPEECH, with the goal of encouraging research in speaker anonymization work and providing a cohesive set of standard datasets and evaluation metrics. Namely, they focus on two main baselines, to which submitted methods should improve upon. In Baseline 1 (shown in Figure 1), the authors first extract frequency information, bottleneck features, and speaker identity “x-vectors" from the original audio. To anonymize the speech, the current x-vector is replace with some *other*, dissimilar x-vector drawn from a pool of other users. This anonymized vector is then used, along with the other features, to generate mel-spectrogram features and recover the now anonymous audio. Baseline 2 inherits from the signal processing (rather than machine learning) literature, and simply uses the McAdams coefficient method [McAdams, 1984, Patino et al., 2020] to anonymize the speech. In general, baseline 1 outperforms baseline 2 across all benchmarks.

Phonetic transcriptions and anonymized speaker embeddings Meyer et al. Meyer et al. [2022] utilizes GANs to reduce the utility-privacy trade-off voice anonymization techniques. This is achieved by reducing audio to phonetic transcriptions, generating a nonexistent voice using GANs, and generate an anonymous utterance of the original based off the transcription, anonymous speaker embedding, and estimated pitch. Their pipeline is made up of 4 models. A speech recognition model, speaker embedding extractor, anonymization module, and a text to speech system. The speech recognition model is based on a hybrid connectionist temporal classification-attention architecture with a conformer as encoder and transformer decoder. This model outputs phone sequences instead of text. The speaker embedding extraction module uses x-vector and ECAPA-TDNN embeddings to extract speaker identity. The speaker embedding anonymization module trains a Wasserstein GAN. The speech synthesis module uses an encoder and decoder built from the conformer architecture. The inputs are transformed into articulatory feature vectors for synthesis. It has also been trained on the same speech concatenated speaker embedding setup as the anonymization module. This pipeline performs better than the baselines of the VoicePrivacy2022 challenge for all metrics except pitch correlation.

Language agnostic self-supervised learning Miao et al. Miao et al. [2022] utilizes semi-supervised learning to overcome problems of complexity, multiple language applications, and poor performance. Soft content representations are used to overcome mispronuncations in the anonymized speech. Miao also uses ECAPA-TDNN for speaker encoding. The semi-supervised model earns universal representations through unlabeled data which increases its application for languages other than English. The architecture includes two pretrained and fixed encoders: one HuBERT based soft encoder, ECAPA-TDNN encoder, one F0 extractor, and HiFi GAN neural vocoder decoder. The HuBERT based soft content encoder is used to capture soft features by predicting a distribution over discrete speech units. By training in a Self-Supervised fashion, this avoids the usage of specific phoneme extractors and thus allows the ASR system to be language-agnostic. The ECAPA-TDNN is the current SOTA method for speaker identity extraction. Miao et al. [2022] tests their method on English *and* Mandarin datasets, thus displaying the adaptability of their method on very distinct languages.

3 Methodology

Problem Setting. Our project centers around the problem setting of **voice anonymization**, wherein the goal is to preserve the semantic content of audio data while removing characteristics that could compromise the identity of the speakers. The pipeline for this task divides into three main sections: extracting import features from the original audio data, performing anonymization on the extracted features, and finally reconstituting the anonymized extracted features into anonymized speech with a Text-to-Speech (TTS)-like model.

As an example, in Figure 1, we show the structure of the first baseline from VoicePrivacy [Tomashenko et al., 2020]. In the first step, three sorts of feature are extracted from the audio:

- The fundamental frequency (or F0) extractor, which extracts the pitch content from a given audio file (specifically, this baseline uses the YAAPT algorithm [Kasi and Zahorian, 2002])
- An Automatic Speech Recognition Acoustic Model (ASR AM), which works as a feature extractor to generate a lower-dimensional representation (generally called Bottleneck or BN features) of the content of the speech.
- Time Delay Neural Network (TDNN) x-vector extraction, which is trained to extract *speaker identity* information.

After these features are extracted, the speech is then anonymized by sampling a random x-vector from a large pool of candidate x-vectors, thus attempting to decouple the content of the speech from the speaker identity. In the last step, the anonymized x-vector, BN features, and F0 are passed to a Speech Synthesis Acoustic Model (SS AM) to generate a mel-filterbank, which itself is the decomposition of an audio signal into seperate filter bins in the mel frequency scale. Namely for a given frequency f, the corresponding frequency in mel space is $m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$. This mel-filterbank is then passed along with the anonymized x-vector to a Neural Source Filter (NSF) model, which converts the mel-filterbank back into audio data.

Given this pipeline, the output audio can be assessed for both how well it reconstructs the semantic meaning of the original speech, and *also* how well speaker identification models can de-anonymize the anonymized speech.

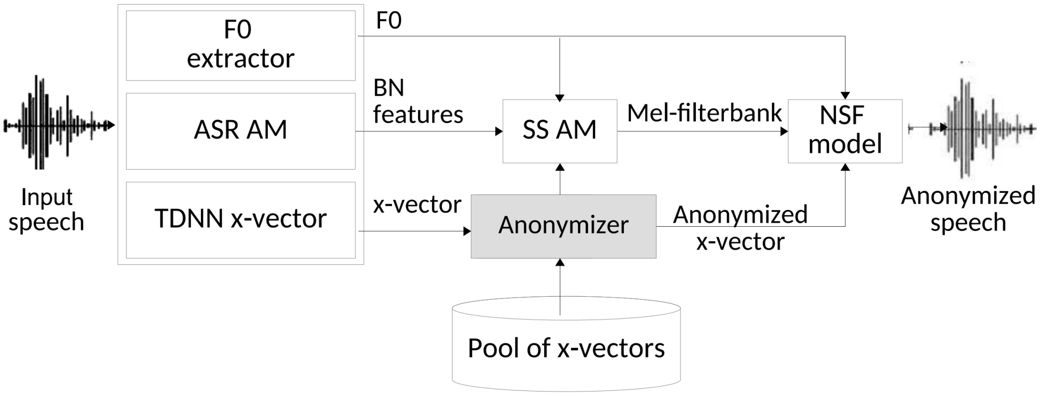

Figure 1: Standard Pipeline for VoicePrivacy baseline B1, taken from Miao et al. [2022].

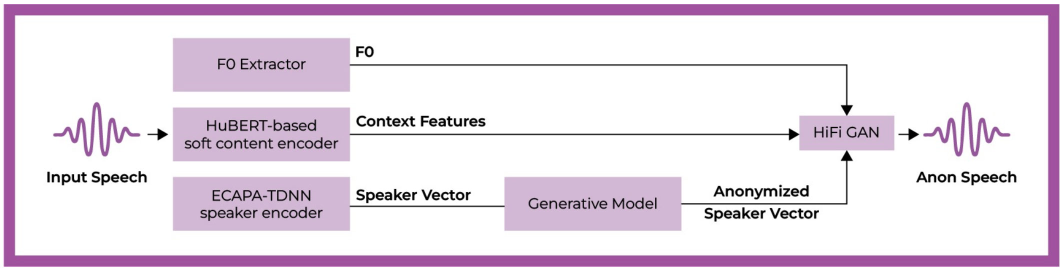

Figure 2: Overview of our proposed method. Notably, our method is a hybrid of Miao et al. [2022] and Meyer et al. [2022], where we replace the x-vector pool anonymizer in Miao et al. [2022]'s setup with a Generative Adversarial Network, inspied by Meyer et al. [2022].

Idea Summary. The current state of the art for voice anonymization is limited by a major tradeoff: systems are either useful but langauge-specific, or language-agnostic but well below other langauge-specific systems in performance. Our solution here is to marry the current state of the art models from both language specific and agnostic camps. Namely, in the current SOTA for language-specific (English) systems, Meyer et al. [2022] forgoes the pitch extraction component and replaces the x-vector anonymization approach with a GAN-based reconstruction of the speaker embedding, where to anonymize speakers samples are drawn from the embedding distribution with low cosine similarity. In the SOTA for language-*agnostic* models, Miao et al. [2022] replace the ASR model with a HuBERT-based soft content encoder (which is language-agnostic), and uses a HiFi-GAN neural vocoder for audio reconstruction.

Given these two models, our solution is to use the HuBERT-based feature extractor and HiFi-GAN from Miao et al. [2022] to preserve the language-agnosticity of the overall system, but borrowing the GAN-based *anonymization* mechanism from Meyer et al. [2022] to leverage the superior performance.

Our second idea, which has not been tried to the best of our knowledge, is to bias the generative anonymization DGM towards learning an *anonymized* embedding distribution, rather than the true distribution. This is accomplished by training the generative model with a regularizer that incorporates whether each sample was later deanonymized, thus penalizing the anonymization module for recreating the true embedding distribution too faithfully.

If our second idea does not bear fruit, our other possible solution is to replace the GAN in the anonymization module with a Diffusion model [Song et al., 2020], in hopes that such a model may train more stably and exhibit less of the vulnerabilities that GANs typically have.

Description. Formally, let the speaker identification extractor be denoted as some function $S : \mathbb{R}^m \to \mathbb{R}^d$ that extracts a d-dimensional vector $\mathbf{x}_{sp}$ (i.e. $S(\mathbf{x}) = \mathbf{x}_{sp}$) for every audio input that encapsulates the speaker's *identity*. Thus, our DGM is tasked with learning the distribution $p(\mathbf{x}_{sp})$

of speaker identification embeddings. In our first experiment, we utilized a WGAN with Quadratic Transport Cost as our DGM. We refer to Liu et al. [2019] for the formulation of the loss function for the discriminator. For the Generator G_θ and discriminator D_ϕ, the generator loss is

$$\min_\theta L(\theta) = -\frac{1}{n} \sum_{\mathbf{z}_i} D_\phi(G_\theta(\mathbf{z}_i)), \tag{1}$$

Where $\mathbf{z}_i \sim p_g(\mathbf{z}_i)$ is sampled from a prior distribution. In this first case, to anonymize a user we sample from G_θ until a given sample has cosine similarity with our original embedding < 0.7

In our second experiment, we propose to add a regularizer to the generator loss. Because we do not care about how well we actually recreate the speaker identity distribution, we can bias our model to perform *worse* (in a traditional sense) and sample from an anonymized distribution at inference.

In order to do this, we want to penalize the success of an automatic speaker verification (ASV) model ASV. The current SOTA ASV systems [Tomashenko et al., 2020, 2022, Miao et al., 2022, Meyer et al., 2022] utilize a speaker identity extractor $S' : \mathbb{R}^m \to \mathbb{R}^d$ that is similar (or could be identical) to S, and then compare the similarity of the embedded features through cosine similarity (i.e. $ASV(\mathbf{x_1}, \mathbf{x_2}) = \frac{S'(\mathbf{x_1}) \cdot S'(\mathbf{x_2})}{\|S'(\mathbf{x_1})\|_2 \|S'(\mathbf{x_2})\|_2}$). In practice, one of the inputs is treated as "enrollment" data, where another recording of the user is used as a reference. Thus, for every generator, step, we can use the modified loss:

$$\min_\theta L(\theta) = -\frac{1}{n} \sum_{\mathbf{z}_i} D_\phi(G_\theta(\mathbf{z}_i)) + \lambda \sum_{\mathbf{z}_i} \frac{G_\theta(\mathbf{z}_i) \cdot \mathbf{x}'_{sp,i}}{\|G_\theta(\mathbf{z}_i)\|_2 \|\mathbf{x}'_{sp,i}\|_2}, \tag{2}$$

where $\mathbf{x}'_{sp,i}$ is another extracted speaker identity vector with the same identity as the ith speaker, but not present in the training set. In practice, for this held out dataset we In this way, we can directly penalize the generator from generating embeddings that, while possibly close to $p(\mathbf{x}_{sp})$, may be *too* close to the identity of the original speaker. In this scenario, instead of resampling for dissimilar samples from the generator, we would be able to sample once, as the generator now represents an anonymized embedding distribution.

Implementation. We are using PyTorch as our main implementation backend. Our models are implemented based on the codebases of VoicePrivacy [Tomashenko et al., 2020, 2022], Miao et al. [2022], SpeechBrain [Ravanelli et al., 2021], and the official WGAN-QC implementation [Liu et al., 2019]. The VoicePrivacy codebase is used for baseline results, SpeechBrain is used for evaluation metrics, and we build on Miao et al. [2022]'s codebase by incorporating the WGAN-QC into the anonymization pipeline. Our architecture for the GAN anonymizer is inspired by Meyer et al. [2022] and Touvron et al. [2022]: we use a reduced size ResMLP for both the Generator and the Discriminator.

4 Experiments

Datasets and Tools. We focus on three major datasets for this present work, which are used to match the evaluation laid out by Tomashenko et al. [2020] and Tomashenko et al. [2022], and the training procedure from Miao et al. [2022]:

1. **The LibriSpeech ASR Corpus**[1]: Contains roughly 1,000 hours of audiobook read speech in English. This dataset, split into development and test sets, is used for evaluation of all metrics.
2. **The LibriTTS dataset**[2]: A multi-speaker English corpus of approximately 585 hours of read English speech, derived from the LibriSpeech corpus. Like Miao et al. [2022], we use this dataset as to train our GAN-based SA pipeline.
3. **The CSTR VCTK Corpus**[3]: Dataset containing data from 110 different speakers reading out 400 sentences with varying accents in English (e.g. Scottish, Irish). This dataset, split into development and test sets, is used for evaluation of all metrics.

[1] http://www.openslr.org/12/

[2] https://research.google/tools/datasets/libri-tts/

[3] https://datashare.ed.ac.uk/handle/10283/3443

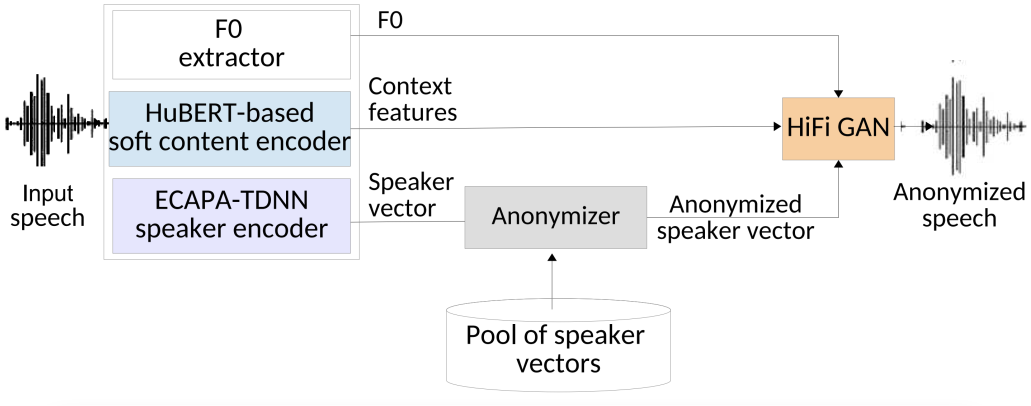

Figure 3: Miao et al. [2022] Anonymization pipeline. Our proposed work in Figure 2 is quite similar for most of the pipeline, the main difference being that anonymization is performed using an x-vector pool of existing speaker embeddings.

Baselines.

- Baseline 1 [Tomashenko et al., 2020, 2022]: This is the first baseline system outline in the VoicePrivacy challenges, which is diagrammed in Figure 1. In this setting, anonymization is performed by matching each speaker's x-vector (i.e. speaker identity embedding) with a *pseudo-speaker* x-vector that is sampled from a given pool of other extracted speaker embeddings such that the pseudo-speaker x-vector is dissimilar to the original speaker. The model then uses an SS AM to generate mel-spectrogram features which are input into an NSF model to recover anonymized audio.
- Baseline 2 [Tomashenko et al., 2020, 2022]: Here, no training is involved, as anonymization is purely performed using signal processing techniques. Namely, the McAdams coefficient [McAdams, 1984, Patino et al., 2020] process is used to deterministically anonymize the audio source data directly.
- Baseline 3 [Miao et al., 2022]: This is the SOTA language-agnostic system that forms the basis of much of the design of our method (shown in Figure 3). Namely, the only difference between our proposed method is that anonymization is performed using the standard pool of x-vectors approach.

We include the first two baselines in order to situate the present work within the wider body of works submitted to the VoicePrivacy challenges, as all submitted works are compared against these two baselines. Additionally, we include the third baselines as it is a natural ablation over whether our novel method improves upon the most similar architecture for SA or not.

Evaluation Metrics. We base our method on two main evaluation metrics:

- **Equal Error Rate (EER)**: This metric quantifies how well the model does in terms of **speaker anonymization**. Specifically, the EER is the rate at which the False Negative Rate (Type 1 error) is equal to the False Positive Rate (Type 2 error) along the ROC curve. While in typical biometric systems a low EER is desired, in SA an *EER close to 50%* is optimal, as this means the verification system cannot determine the speaker identity.
- **Word Error Rate (WER)**: This metric quantifies how well the model does in terms of **speech recognition**. Namely, we have that:

  $$WER = \frac{(S + D + I)}{N} \tag{3}$$

  Where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the reference.

Quantitative Results. We first note that we test our models in three different scenarios:

1. Unprotected (OO): No anonymization used, with the attacker (ASV) having access to the true audio.
2. Ignorant attacker (OA): Audio is anonymized, attacker uses non-anonymized audio for enrollment but is tested on the anonymized audio.
3. Lazy-informed (AA): Audio is anonymized, attacker has access to anonymized audio for both enrollment and the testing.

Notably, AA represents a more realistic and higher-risk scenario than OA, as the attacker now knows something about the anonymization process. To calculate EER metrics, each dataset comes with a set of enrollment audio files, which we embed using a pretrained ECAPA-TDNN and average together per user. These enrollment embeddings are then compared with each anonymized embedding to calculate the EER. In Table 1, we display the EER values as the absolute difference from 50%, as an EER of 50% denotes perfect anonymization. We see that our methods perform slightly worse than the Baseline 1 in terms of anonymization, and note that in general our methods perform better (relative to Baseline 1) in the AA scenario. We too see that there is not much benefit to including the regularization penalty at all, which we surmise is due to instability in the GAN training process. That being said, we do see that our method does consistently outperform the model from Miao et al. [2022], showing that in terms of *language-agnostic* systems, our method is a strict improvement.

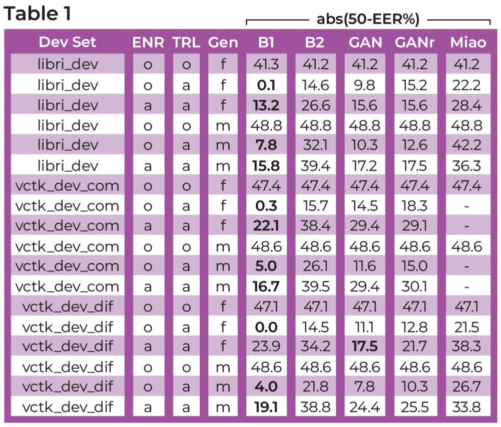

Table 1

				abs(50-EER%)				
Dev Set	ENR	TRL	Gen	B1	B2	GAN	GANr	Miao
libri_dev	o	o	f	41.3	41.2	41.2	41.2	41.2
libri_dev	o	a	f	**0.1**	14.6	9.8	15.2	22.2
libri_dev	a	a	f	**13.2**	26.6	15.6	15.6	28.4
libri_dev	o	o	m	48.8	48.8	48.8	48.8	48.8
libri_dev	o	a	m	**7.8**	32.1	10.3	12.6	42.2
libri_dev	a	a	m	**15.8**	39.4	17.2	17.5	36.3
vctk_dev_com	o	o	f	47.4	47.4	47.4	47.4	47.4
vctk_dev_com	o	a	f	**0.3**	15.7	14.5	18.3	-
vctk_dev_com	a	a	f	**22.1**	38.4	29.4	29.1	-
vctk_dev_com	o	o	m	48.6	48.6	48.6	48.6	48.6
vctk_dev_com	o	a	m	**5.0**	26.1	11.6	15.0	-
vctk_dev_com	a	a	m	**16.7**	39.5	29.4	30.1	-
vctk_dev_dif	o	o	f	47.1	47.1	47.1	47.1	47.1
vctk_dev_dif	o	a	f	**0.0**	14.5	11.1	12.8	21.5
vctk_dev_dif	a	a	f	23.9	34.2	**17.5**	21.7	38.3
vctk_dev_dif	o	o	m	48.6	48.6	48.6	48.6	48.6
vctk_dev_dif	o	a	m	**4.0**	21.8	7.8	10.3	26.7
vctk_dev_dif	a	a	m	**19.1**	38.8	24.4	25.5	33.8

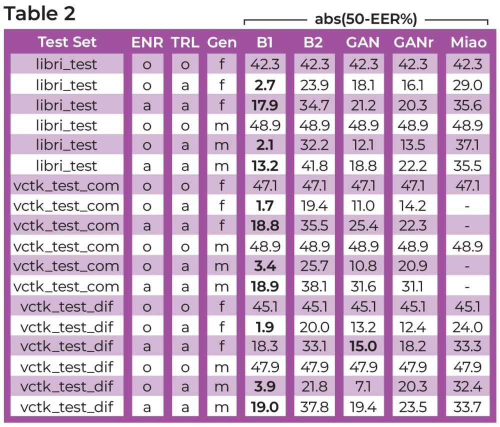

Table 2

				abs(50-EER%)				
Test Set	ENR	TRL	Gen	B1	B2	GAN	GANr	Miao
libri_test	o	o	f	42.3	42.3	42.3	42.3	42.3
libri_test	o	a	f	**2.7**	23.9	18.1	16.1	29.0
libri_test	a	a	f	**17.9**	34.7	21.2	20.3	35.6
libri_test	o	o	m	48.9	48.9	48.9	48.9	48.9
libri_test	o	a	m	**2.1**	32.2	12.1	13.5	37.1
libri_test	a	a	m	**13.2**	41.8	18.8	22.2	35.5
vctk_test_com	o	o	f	47.1	47.1	47.1	47.1	47.1
vctk_test_com	o	a	f	**1.7**	19.4	11.0	14.2	-
vctk_test_com	a	a	f	**18.8**	35.5	25.4	22.3	-
vctk_test_com	o	o	m	48.9	48.9	48.9	48.9	48.9
vctk_test_com	o	a	m	**3.4**	25.7	10.8	20.9	-
vctk_test_com	a	a	m	**18.9**	38.1	31.6	31.1	-
vctk_test_dif	o	o	f	45.1	45.1	45.1	45.1	45.1
vctk_test_dif	o	a	f	**1.9**	20.0	13.2	12.4	24.0
vctk_test_dif	a	a	f	18.3	33.1	**15.0**	18.2	33.3
vctk_test_dif	o	o	m	47.9	47.9	47.9	47.9	47.9
vctk_test_dif	o	a	m	**3.9**	21.8	7.1	20.3	32.4
vctk_test_dif	a	a	m	**19.0**	37.8	19.4	23.5	33.7

Table 1: $|50 - EER|$ results, where lower values are better. Namely, our models performs noticeably worse than Baseline 1 in OA, but only slightly worse (and sometimes better) in AA setting. Additionally, our method generally outperforms the model from Miao et al. [2022].

In Table 2, we see that both GAN and GANr outperform all the baselines across all datasets by significant margins. This displays a key strength of our GAN-based anonymization method, as the x-vector or signal processing based methods seem to suffer from introducing audio artifacts and negatively effect intelligibility.

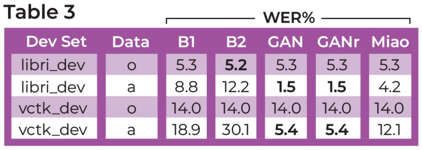

Table 3

		WER%				
Dev Set	Data	B1	B2	GAN	GANr	Miao
libri_dev	o	5.3	**5.2**	5.3	5.3	5.3
libri_dev	a	8.8	12.2	**1.5**	**1.5**	4.2
vctk_dev	o	14.0	14.0	14.0	14.0	14.0
vctk_dev	a	18.9	30.1	**5.4**	**5.4**	12.1

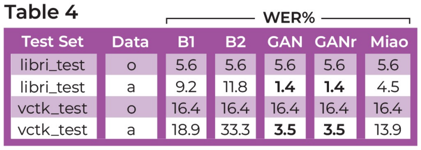

Table 4

		WER%				
Test Set	Data	B1	B2	GAN	GANr	Miao
libri_test	o	5.6	5.6	5.6	5.6	5.6
libri_test	a	9.2	11.8	**1.4**	**1.4**	4.5
vctk_test	o	16.4	16.4	16.4	16.4	16.4
vctk_test	a	18.9	33.3	**3.5**	**3.5**	13.9

Table 2: Word Error Rate Results. Our method significantly out performs all baselines and achieves SOTA values of speech recognition, thus showing that GAN-based speaker anonymization may be quite good at preserving realistic audio quality and intelligibility.

Table 3: Average Cosine Similarity of main models on training set (LibriTTS). Cosine Similarity is close to 0 in both cases, thus alluding to the idea that low cosine similarity in the speaker embedding space may not actually equate to good anonymization, and that penalizing the cosine similarity mostly just makes training dynamics more difficult and is not needed.

GAN Average Cosine Similarity	GANr Average Cosine Similarity
-0.01	0.03

Qualitative Results. We include the link to our webpage showcasing audio samples from our anonymization methods and the first baseline here. Notably, we find that compared to the baseline, our method preserves audio quality significantly better, as the baseline anonymization introduces artifacts into the audio file that add a robotic quality to the output. Additionally, we find that GANr introduces somewhat unrealistic pitch modulations as the audio evolves.

Ablative Studies. As we saw in Table 1, when we remove the regularizer term from the training procedure (i.e. moving from GANr back to GAN), we see that our model does *better* in terms of anonymization performance. In order to get a better idea as to why this behavior occurs, we look at the average cosine similarity of each method during the training process. In Table 3, we see that the cosine similarities are close to 0 *in both regimes*, even though it is only explicitly penalized in GANr. Thus, we see that penalizing cosine similarity in the embedding space may not be the best way to encourage the generative model to anonymize the data implicitly, as the cosine similarity is already reasonably optimal (i.e. close to 0) without the penalization. Additionally, the subpar anonymization performance of both methods with respect to baseline 1 calls in to question whether looking at cosine similarity is even a useful way to think about anonymizing speaker data.

5 Conclusion and Discussion

As a whole, the current work presents the first step into designing speech anonymization systems that balance the generalizability of language-agnostic approaches with the recent advances in generative modeling to improve the anonymization process. Though anonymization performance was subpar, we showed that our pipeline achieves remarkable performance in terms of improving speech recognition, and achieves better performance than the best performing language-agnostic comparisons. The high-level “language-agnostic + DGM anonymization" framework proposed provided is quite flexible, thus allowing future work to explore other generative models as part of the anonymization pipeline. If given more time, the authors would have moved towards utilizing a different class of generative models (such as diffusion models) for the anonymization function, as instability and troubles with GAN training limited the quantitative success of our method. The authors learned a considerable amount about the space of speaker anonymization research, as well as all the benefits and drawbacks that come with training GANs from scratch.

References

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. 2014.

Natalia Tomashenko, Brij ML Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, et al. The voiceprivacy 2020 challenge. *VoicePrivacy, Feb*, 2020.

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.

Haojian Jin, Gram Liu, David Hwang, Swarun Kumar, Yuvraj Agarwal, and Jason I Hong. Peekaboo: A hub-based approach to enable transparency in data processing within smart homes. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 303–320. IEEE, 2022.

Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko. Language-independent speaker anonymization approach using self-supervised pre-trained models. *arXiv preprint arXiv:2202.13097*, 2022.

Sarina Meyer, Florian Lux, Pavel Denisov, Julia Koch, Pascal Tilli, and Ngoc Thang Vu. Speaker anonymization with phonetic intermediate representations. *arXiv preprint arXiv:2207.04834*, 2022.

Razieh Khamsehashari, Yamini Sinha, Jan Hintz, Suhita Ghosh, Tim Polzehl, Carlos Franzreb, Sebastian Stober, and Ingo Siegert. Voice privacy - leveraging multi-scale blocks with ecapa-tdnn se-res2next extension for speaker anonymization, 2022.

Jixun Yao, Qing Wang, Li Zhang, Pengcheng Guo, Yuhao Liang, and Lei Xie. Nwpu-aslp system for the voiceprivacy 2022 challenge. *arXiv preprint arXiv:2209.11969*, 2022.

Stephen Edward McAdams. *Spectral fusion, spectral parsing and the formation of auditory images*. Stanford university, 1984.

Jose Patino, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans. Speaker anonymisation using the mcadams coefficient. *arXiv preprint arXiv:2011.01130*, 2020.

Kavita Kasi and Stephen A. Zahorian. Yet another algorithm for pitch tracking. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:I–361–I–364, 2002.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Huidong Liu, Xianfeng Gu, and Dimitris Samaras. Wasserstein gan with quadratic transport cost. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4832–4841, 2019.

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.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.