Speecher: Towards Privacy Ensuring Decoder Only Speech Reconstruction Through Disentanglement for German Speech Anonymization Using Any-to-Many Voice Conversion

Arnab Das[1], *Carlos Franzreb*[1], *Suhita Ghosh*[2], *Tim Polzehl*[1], *Sebastian Möller*[3]

[1]German Research Center for Artificial Intelligence, Germany [2]Artificial Intelligence Lab (AILab), Otto-von-Guericke-University, Germany [3]Quality and Usability Lab, Technical University of Berlin, Germany

{arnab.das, carlos.franzreb, tim.polzehl}@dfki.de, suhita.ghosh@ovgu.de, sebastian.moeller@tu-berlin.de

Abstract

Voice conversion (VC) has emerged as an essential tool for speaker anonymization providing privacy in speech data. Recent reconstruction-based voice conversion (VC) frameworks learn to reconstruct speech by disentangling content, pitch, and speaker representations. Often these methods show poor content and prosody preservation. Furthermore, these models are constrained in their ability to execute cross-lingual voice conversion, where the source and target speech are from different languages due to the inherent coupling of the encoder and decoder components to specific languages within the model architecture. We propose the decoder-only reconstruction-based VC framework Speecher, trained with perceptual losses, and demonstrate that speech features can be extracted from pre-trained networks without additional encoder training. A thorough objective and subjective study using German speech data reveals that our framework improves prosody and content preservation while maintaining anonymization capabilities.

Index Terms: voice conversion, speech anonymization, speech reconstruction, Speech representation disentanglement

1. Introduction

In recent years, the remarkable increase in the general usage of speech data and a proliferation of use of voice-based services and devices can be attributed to rapid advances in human-computer interaction, specifically speech processing research. However, the expanding usage of voice data has brought significant privacy and security concerns, as sensitive information, particularly speaker identification, might be exploited to further malicious intent if not protected adequately. The urgent need to provide security and privacy for spoken data has prompted academics to propose solutions. Speaker anonymization is a solution promising to protect speaker anonymity while speaking to cloud-based voice-enabled services, protecting whistleblowers, providing anonymous statements to legal aspects, or even archiving medical speech data as per guidelines by General Data Protection Regulation (GDPR) guidelines [1]. Speaker anonymization reduces the risks associated with illegal identification and profiling by obscuring a speaker's unique voice features. At its core, speaker anonymization strives to improve privacy by preventing individuals from being identified via their voice data. This is especially important in online speech-based communications, where voice data is vulnerable to illegal access and manipulation. Speaker anonymization technologies contribute to the protection of individual privacy rights by making it more difficult to link voice data to the original speaker. However, excessive distortion can render speech incoherent, thereby making it unusable for downstream applications. Hence a careful consideration between privacy and utility, also known as privacy-utility trade-off is essential based on the use of anonymized speech data.

In speech processing research, voice conversion (VC) is a problem that involves changing the speaker's identity of an utterance while retaining the linguistic content and prosody intact, such that an utterance from one speaker i.e. the source speaker sounds like another i.e. the target speaker [2]. This makes a VC system a suitable and important tool for speaker anonymization. In a pathological use case, the anonymized speech data can also be utilized for further analysis. To execute the task, VC systems require reference utterances from the target speaker. In our research, we concentrate on any-to-many VC systems, which may convert utterances from any arbitrary speaker to a predetermined set of target speakers.

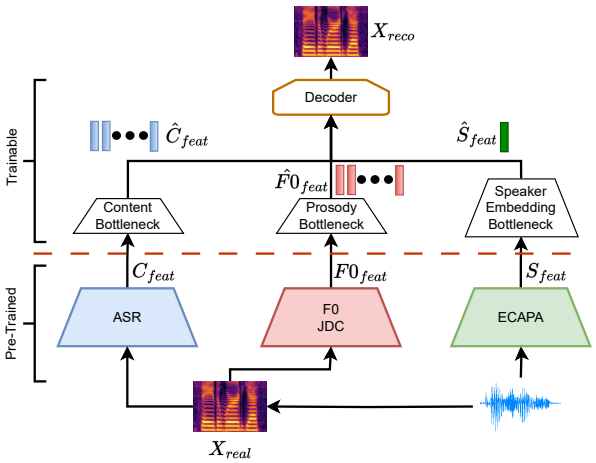

Figure 1: *Schematic diagram of the proposed method.*

The pursuit of high-quality, natural-sounding VC systems is a very active and ongoing research area. Early VC systems [3, 4] relied on parallel training data in which the source and target speakers utter identical linguistic material. Parallel training data was also used in early deep learning-based voice converter systems [5, 6]. Although the introduction of deep learning enhanced the quality of these systems, they remain far from flawless. Not only is gathering parallel training data time-consuming and costly, but it does not reflect the real-world scenario in which these systems might be employed with arbitrary

source speakers [7]. In addition, these systems are noted for having issues with mispronunciation and unreliable training [8]. To alleviate the problem of data acquisition, non-parallel VC systems were proposed. Most recent non-parallel VC systems [9, 10, 11, 12] use different variants of the generative adversarial network (GAN) [13]. The essential premise of GAN-based techniques is to treat any individual speaker and their voice features as a style domain, and to re-frame the task of VC as a domain transfer problem [7]. Though these approaches can yield good quality samples, they are criticized for the harder training process since GAN training has an ill-defined convergence criterion [7]. Furthermore, to produce more natural-sounding samples, more and more loss objectives are incorporated [12], making tuning of each component extremely difficult.

Another paradigm to VC systems is to learn a task of reconstructing speech from disentangled speaker and content representations rather than approaching the task of VC as an explicit domain transfer problem. During inference, the source speaker's representation is altered with the target speaker's representation to accomplish VC or anonymization while retaining linguistic content. Several approaches have been proposed over time to obtain disentangled speech representations while preventing information leakage across them. In line with this concept, [14] proposed a method that uses vector quantized variational auto-encoder (VQ-VAE) to learn disentangled content and timbre features from an utterance. In [7], the authors introduced a trainable temporal and dimensional information bottleneck to disentangle different speech features and proposed an auto-encoder-based model for audio reconstruction. Alternatively, [15] introduces adaptive instance normalization (AdaIN) as a technique for accomplishing feature disentanglement. The techniques proposed in [16, 17], combined mutual information (MI) upper bound minimization between different pairs of speech representations and VQ to achieve disentangled content, speech, and pitch features. In this reconstruction-based paradigm, [16] shows through comprehensive objective and subjective results that their proposed method, VQMIVC, is superior to several other predecessor methods in all aspects, including prosody and content preservation as well as naturalness. All of these reconstruction-based VC frameworks train auto-encoder architectures with two or more encoders. Furthermore, most research on VC frameworks employs English/Chinese datasets to illustrate their results, making it difficult to assess their efficiency in low-resource European languages.

This paper introduces a novel decoder-only, reconstruction-based any-to-many VC framework *"Speecher"* that combines trainable bottleneck layers with AdaIN [18] without training feature-specific encoders. Content, pitch, and speaker embeddings are extracted from disjoint pre-trained task-specific supervised models, which are not re-trained as part of the decoder training procedure. This way, we open up new research perspectives towards cross-lingual voice conversions without requiring training auto-encoder frameworks hard-coupled to a single language which is especially helpful for underrepresented languages. For experimentation, we train our model merely on $\sim$ 39.5 hours of German speech data from seven individuals, reflecting its efficacy in low-resource settings where availability or training data is scarce. We thoroughly compare our proposed method both objectively and subjectively to VQMIVC [16] baseline. Additionally, we propose to use perceptual losses to improve prosody and linguistic content preservation. Experimental results demonstrate that our proposed strategy considerably increases overall quality by $\sim$ 10%, intelligibility by $\sim$ 17%, content preservation by $\sim$ 13%, and pitch correlation by $\sim$ 9% while maintaining similar anonymization capabilities. Our results demonstrate that different speech features obtained from supervised pre-trained networks tailored to specific tasks are adequate to reconstruct speech with high content and preservation of prosody.

2. Method

2.1. Architecture

A schematic architecture diagram of the proposed decoder-only framework is presented in Figure 1.

Decoder: For a source mel-spectrogram X_{real}, the decoder D learns to reconstruct the mel-spectrogram from disentangled content, pitch, and speaker embeddings such that $X_{reco} = D(\hat{C}_{feat}, \hat{F0}_{feat}, \hat{S}_{feat})$, where $\hat{C}_{feat}$, $\hat{F0}_{feat}$ and $\hat{S}_{feat}$ are the outcome of the content bottleneck C_{BN}, pitch bottleneck $F0_{BN}$ and speaker embedding bottleneck S_{BN} layers respectively. Trainable bottleneck layers serve the purpose of dynamic feature selection to achieve disentanglement.

Content feature: Linguistic content features C_{feat} are extracted from a phoneme-based supervised automatic speech recognition (ASR) framework [12] pre-trained on German data. As suggested by [19] the features are obtained from the intermediate layers before the LSTM layers $C_{feat} = ASR_{int}(X_{real})$. These content features are passed through the trainable content bottleneck layer to output $\hat{C}_{feat}$ before passing on to D.

Pitch feature: Pitch features $F0_{feat}$ are obtained from a pre-trained supervised joint detection and classification (JDC) F0 prediction network [20]. Similar to content features, $F0_{feat}$ is the output of the intermediate convolution layer before the recurrent layers $F0_{feat} = JDC_{int}(X_{real})$. Subsequently, the disentangled F0 feature is extracted by passing $F0_{feat}$ through the pitch bottleneck layer $\hat{F0}_{feat} = F0_{BN}(F0_{feat})$. The temporal dimensions for content and pitch features are identical.

Speaker Embedding We use 192-dimensional ECAPA-TDNN [21] global speaker embedding S_{feat} directly retrieved from the speech utterance using SpeechBrain toolbox [22]. Afterward, the dimensionality of the speaker feature is reduced by passing it through S_{BN}, which outputs $\hat{S}_{feat} = S_{BN}(S_{feat})$.
C_{BN} and $F0_{BN}$ layers consists of residual convolution blocks whereas S_{BN} is stacked linear layers with ReLU activations. We use a similar decoder architecture that is proposed in [12] that uses several upsampling blocks comprising AdaIN. During inference, the speaker embedding S_{feat} is extracted from an utterance from the target speaker, while C_{feat} and $F0_{feat}$ are from the source utterance. The generated mel-spectrogram is converted to a waveform by a HiFi-GAN [23] vocoder.

2.2. Training objective

In the training phase, we jointly train the bottleneck layers along with the decoder to learn to generate X_{reco} from $\{C_{feat}, F0_{feat}, S_{feat}\}$. Hence reconstruction loss is an essential part of the overall training objective. In addition, we use

adversarial loss to improve the quality of reconstructed speech, and perceptual pitch and content loss to improve linguistic information and prosody retention.

Reconstruction loss: Given a source mel-spectrogram X_{real} the decoder outputs a reconstruction X_{reco}, then the reconstruction loss is a combination of mean absolute error (MAE) and the mean square error (MSE) as depicted in Eqn. 1.

$$\mathcal{L}_{reco} = \mathbb{E}_X \Big[\big|\big| X_{reco} - X_{real} \big|\big|_1 + \big|\big| X_{reco} - X_{real} \big|\big|_2 \Big] \quad (1)$$

Content loss: In addition to reconstruction loss, we also minimize a perceptual content loss which is an L1 loss between the intermediate ASR feature extracted from the source and reconstructed mel-spectrograms as depicted in Eqn. 2.

$$\mathcal{L}_{asr} = \mathbb{E}_X \Big[\big|\big| ASR_{int}(X_{reco}) - ASR_{int}(X_{real}) \big|\big|_1 \Big] \quad (2)$$

Pitch loss: As suggested by [12], to enhance overall prosody preservation in the reconstructed speech we also penalize the network by L1 loss between the predicted normalized absolute F0 of source and reconstructed speech as shown in Eqn. 3 where $JDC_{norm}(X) = \frac{JDC(X)}{||JDC(X)||_1}$ and $||JDC(X)||_1$ refers to the per instance temporal mean .

$$\mathcal{L}_{F0} = \mathbb{E}_X \Big[\big|\big| JDC_{norm}(X_{reco}) - JDC_{norm}(X_{real}) \big|\big|_1 \Big] \quad (3)$$

Adversarial loss: Together with the decoder, we also train a Discriminator $\mathcal{G}$ that classifies between real and reconstructed mel-spectrograms given a speaker identifier y. We train our reconstruction network by minimizing the adversarial loss as depicted in Eqn. 4.

$$\mathcal{L}_{adv} = \mathbb{E}_{X,y} \Big[\log(\mathcal{G}(X_{real}, y) + \log(1 - \mathcal{G}(X_{reco}, y) \Big] \quad (4)$$

The overall training objective is to minimize the weighted sum of all individual losses as presented in Eqn. 5, where $\alpha, \beta, \gamma, \lambda$ are hyper-parameters.

$$\mathcal{L} = \alpha \mathcal{L}_{reco} + \beta \mathcal{L}_{asr} + \gamma \mathcal{L}_{F0} + \lambda \mathcal{L}_{adv} \quad (5)$$

Superficially, it may seem that our proposed method is similar to existing methods that explore feature disentanglement techniques for VC, such as VQMIVC, but fundamentally, it is not. VQMIVC learns to disentangle features by jointly training feature-specific encoders along with decoders which makes the encoders hard coupled with the decoder and forces the encoders and decoder to be hard coupled. Hence where the training data is scarce it fails to achieve good-quality results. In contrast, our method provides the flexibility of using any pre-trained encoders, as those are not re-trained as part of decoder training.

3. Datasets and experiments

3.1. Dataset

We train the baseline model and our proposed method on approximately 40 hours of German speech data from the multilingual LibriSpeech (MLS) corpus [24] which is derived from read audiobooks. The training set includes approximately 9550 utterances from seven speakers, four female and three male. Gender parity is achieved by including more utterances from male speakers than female speakers. The length of the training utterances ranges from 10 to 20 seconds. The validation set includes another 2660 utterances from the same speakers. For training our proposed method, the utterances are upsampled from 16 to 24 kHz whereas for the baseline training, it remains at 16 kHz. We evaluate both the models on in-domain utterances from MLS and on out-of-domain utterances from HUI-Audio-Corpus-German [25] clean dataset. The HUI dataset also comprises high-quality read audiobooks downloaded from LibriVox[1]. The phoneme-based ASR and the vocoder are pre-trained with the German subset of the MLS dataset. Additionally, the vocoder is also trained with English utterances from VCTK [26], ESD [27], and RAVDESS [28] datasets.

3.2. Training details

We named our framework "Speecher" and consider VQMIVC as the baseline for comparison. We train our model for 100 epochs on an NVIDIA H100 GPU with a batch size of 32, for a total training time of approximately 26 hours. To optimize, we use AdamW [29] with a fixed learning rate of 10^{-4} and weight decay of 10^{-4}. We perform hyper-parameter tuning on the validation set and choose $\alpha = 3.0$, $\beta = 10.0$, $\gamma = 0.75$ and $\lambda = 2.0$. The baseline is trained for 500 epochs. The code can be found online [2].

3.3. Evaluations

We evaluated the proposed framework both subjectively and objectively.

Objective evaluations: To objectively evaluate our proposed any-to-many framework, we randomly selected four source unseen speakers from the MLS dataset and four target speakers from the seen speaker set, yielding 16 conversion pairs. The source and target testing speaker groups are gender balanced. We chose around 108 test utterances per source speaker, leading up to 1732 MLS→MLS test conversions. We calculate the character error rate (CER) to assess content preservation. The converted utterances are automatically transcribed with Whisper [30] ASR. Additionally, as a measure of prosody preservation evaluation we report the pitch correlation coefficient (PCC) [31]. A higher PCC value indicates better prosody preservation whereas for CER a lower value is desirable. Additionally, we compute the speaker distance score (SDS) between the converted and the source utterances using ECAPA-TDNN speaker embeddings, where $SDS = 1 - CosineSimilarity(S_{feat}(Source), S_{feat}(Converted))$. Furthermore, to assess our model on out-of-domain data, we perform objective evaluation on 764 HUI→MLS conversions, randomly selecting one male and one female source speaker from the HUI dataset while keeping the four target speakers the same as before, yielding eight conversion pairs.

Subjective evaluations: We conducted a detailed subjective user survey on the Crowdee[3] platform, where a total of 130 na-

[1] https://librivox.org/

[2] https://github.com/arnabdas8901/speecher.git

[3] https://www.crowdee.com/

Table 1: *Objective evaluation results. Each cell contains mean values and in brackets* 95% *confidence interval. The Group column denotes gender-dependent conversion subgroups for the source and target speakers.*

Source - Target	Group	PCC[$\times 10^2$]↑		CER[%]↓		SDS↑	
		Speecher	VQMIVC	Speecher	VQMIVC	Speecher	VQMIVC
MLS→ MLS	All	**82.51** (0.31)	73.85 (0.4)	**6.53** (0.46)	19.55 (0.65)	0.57 (0.01)	**0.76**(0.01)
	M2M	**85.23** (0.39)	77.16 (0.74)	**5.66** (1.01)	17.54 (1.19)	-	-
	M2F	**86.06** (0.38)	77.69 (0.57)	**6.37** (1.16)	18.01 (1.18)	-	-
	F2M	**78.24** (0.65)	69.94 (0.85)	**7.76** (0.79)	22.32 (1.41)	-	-
	F2F	**80.72** (0.66)	70.82 (0.72)	**6.29** (0.68)	20.23 (1.38)	-	-
HUI→ MLS	All	**80.86** (0.46)	75.66 (0.57)	**5.34** (0.62)	12.87 (0.74)	0.7(0.01)	**0.8**(0.01)
	M2M	**79.47** (1.06)	75.3 (0.97)	**5.49** (1.55)	14.52 (1.75)	-	-
	M2F	**81.04** (0.9)	71.72 (1.26)	**6.37** (1.58)	16.56 (1.75)	-	-
	F2M	**80.85** (0.91)	78.59 (1.01)	**4.97** (0.66)	9.94 (0.88)	-	-
	F2F	**82.14** (0.7)	77.25 (1.02)	**4.46** (0.78)	10.17 (1.01)	-	-

tive German speakers participated. For the subjective study, we randomly choose two unseen source speakers and two seen target speakers, both the source and target speaker set comprise one male and one female speaker, resulting in four conversion pairs. Each conversion pair is evaluated for 15 converted utterances, totaling 60 converted utterances. Each of these is assessed by at least 12 participants. We subjectively measure the overall quality of the stimuli and intelligibility in a continuous scale mean opinion score (MOS) ranging $[1, 5]$ with 1:bad *('Mangelhaft')*, 2:poor (*'Mäßig'*), 3:fair (*'Ordentlich'*), 4:good (*'Gut'*), 5:excellent (*'Ausgezeichnet'*). Alongside, the participants are also asked to mark the source and target speaker similarity of the converted samples in a continuous scale ranging $[1, 5]$, with 1:different (*'Unterschiedlich'*) and 5:same (*'Derselbe'*) speaker. As a quality control measure, we presented hidden trapping questions (not recognizable to the participants) and removed annotations and participants upon failure.

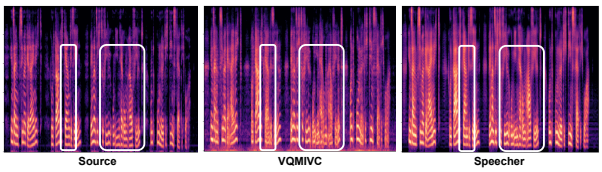

Figure 2: *Spectrogram comparison between source and converted utterances by VQMIVC and Speecher.*

4. Results and discussion

Table 1 summarizes the results from our objective evaluations. Our proposed method Speecher outperforms the VQMIVC baseline in terms of linguistic content and prosody preservation objective metrics for all conversion groups, whether in-domain MLS→MLS or out-of-domain HUI→MLS. The table includes both mean values and a 95% confidence interval (CI). For PCC, Speecher archives a mean value of 82.51 for all MLS→MLS conversions and 80.86 for HUI→MLS conversions. The baseline results range clearly lower, i.e. with average PCC score of 73.85 and 75.66 for MLS→MLS and HUI→MLS conversions respectively, which is significantly ($p < 0.05$ on paired t-test) lower than our proposed method. Improvement in prosody preservation can also be validated by inspecting and comparing the mel-spectrograms of converted utterances to the source mel-spectrogram. Figure 2 illustrates one such example, with white rectangles indicating noteworthy segments. The converted speech by VQMIVC fails to keep harmonic nuances, resulting in over-smoothed intonation, whereas our proposed Speecher framework is capable of preserving more of the harmonic pattern.

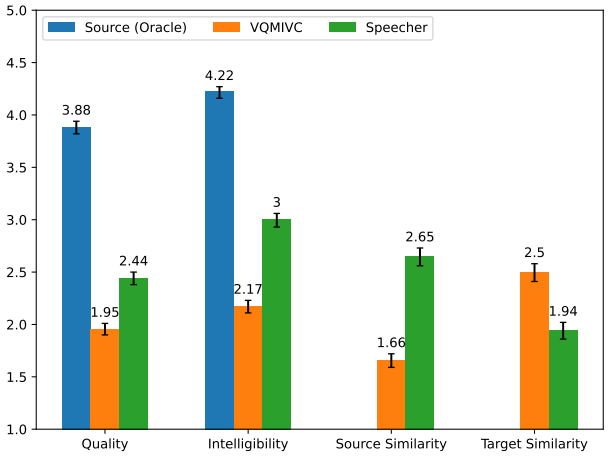

Figure 3: *Subjective evaluation results. Mean and 95% confidence interval values are reported.*

In terms of CER, the baseline archives a score of 19.55% for MLS→MLS, whereas the same for Speecher is significantly ($p < 0.05$) lesser, 6.53%. For HUI→MLS, our method archives a low CER value of 5.34% compared to the baseline's CER of 12.87%. The scores obtained by VQMIVC are comparable to those reported in [16] for the English (VCTK) dataset, with slight degradation. Table 1 also presents PCC and CER results for gender-specific conversion subgroups, and the trend

remains consistent, in every case, the performance of our proposed framework surmounts that of the baseline. Both the models achieve comparatively high PCC scores for MLS→MLS conversions when the source speaker is male, but the trend is not observed for HUI→MLS conversions. The proposed perceptual pitch and content loss help our method achieve better PCC and CER results compared to the baseline without any such losses. To support our claim we have also performed an ablation study and the results are reported in Table 2.

Table 2: *Ablation study results for pitch and content preservation. Mean and 95% confidence interval values (in brackets) are reported.*

Model	**PCC[$\times 10^2$]**$\uparrow$	**CER[%]**$\downarrow$
Speecher	**82.51** (0.31)	**6.53** (0.46)
No Pitch Loss ($\gamma = 0$)	71.85 (0.51)	-
No Content Loss ($\beta = 0$)	-	13.08 (0.54)

In terms of SDS, our method achieves a mean score of 0.57, which is a bit lower than that of the VQMIVC baseline (0.76) for in-domain MLS→MLS conversions. However, the mean SDS scores for HUI→MLS conversions are 0.7 and 0.8 for Speecher and VQMIVC respectively, hence seem to be comparable. A higher average SDS score indicates a further separation between the source and converted utterance's speaker embeddings. However, according to [32], a cosine distance of > 0.3 between the speaker embeddings of the source and target speaker is sufficient for an effective anonymization.

Subjective evaluation results are depicted in Figure 3. The *source* refers to the annotations received by the original stimuli from the dataset without any processing, which is indicative of the upper bound score for the respective evaluation dimensions. For overall quality, the VQMIVC baseline archives a MOS score of 1.95 whereas our proposed method gets 2.44, which is significantly ($p < 0.05$ on a two-sided paired t-test) higher than the baseline. A hearing test, conducted by us, also demonstrates that VQMIVC introduces a persistent low-intensity background noise into converted utterances. In terms of spoken content intelligibility, our method Speecher archives a mean score of 3, significantly outperforming VQMIVC which achieves a mean intelligibility score of 2.17. Proposed additional discriminator loss could be contributing towards the improved overall quality results, which is consistent with the findings by [33].

In terms of similarity to the source speaker, the mean scores of Speecher and VQMIVC are 2.65 and 1.66 respectively. This score is expected to be as low as possible for the speech anonymization use case. This subjective evaluation result corroborates the SDS scores presented as part of the objective examination. In terms of similarity with the target speaker, VQMIVC achieves a mean score of 2.5 in comparison to a mean score of 1.94 by Speecher. A higher score in this aspect is desirable for the use case of voice conversion.

The objective and subjective evaluation results demonstrate the privacy-utility trade-off [31] for reconstruction-based voice converter systems. A larger degree of disentanglement between the speaker, content, and pitch features improves inference time similarity to the target speaker while sacrificing content and prosody preservation. On the other side, better overall speech quality, intelligibility, and prosodic similarity reduce the target speaker similarity of the converted utterance. To support our claim, a separate objective study is performed and the results are reported in Table 3. The results show that when the perceptual pitch loss coefficient is reduced from 0.75 to 0.25, mean PCC goes down and SDS goes up indicating a higher privacy score at the cost of lower utility.

Table 3: *Objective study results demonstrating privacy-utility trade-off. Mean and 95% confidence interval values (in brackets) are reported.*

Model	**PCC[$\times 10^2$]**	**SDS**
Speecher with $\gamma = 0.75$	**82.51** (0.31)	0.57 (0.01)
Speecher with $\gamma = 0.25$	79.58 (0.38)	**0.63** (0.01)

5. Conclusion

In this paper, we propose a decoder-only speech reconstruction-based any-to-many VC framework named Speecher, suitable for performing speech anonymization. Additionally, we propose to use perceptual losses for better prosody and linguistic content preservation. Extensive objective and subjective evaluation results reveal that our proposed method can significantly improve overall quality, intelligibility, and intonation preservation while converting utterances without compromising its anonymization capability for both in-domain and out-of-domain source speakers. Although being superior in many dimensions, our proposed method shows marginally lower target speaker similarity compared to the baseline. In our future work, we will further work on improving the target speaker similarity for our framework. We also work towards extension to other underrepresented European languages. Moreover, we aspire to broaden the scope of our methodology by incorporating cross-language sources and target utterances, facilitating one-shot any-to-any voice conversion use cases.

Speaker anonymization is a crucial tool for protecting speakers' identities and upholding their right to privacy, however, it can also be abused by malicious individuals who may take advantage of this technological advancement. Anonymization technologies may be used to elude law enforcement, trick identity verification systems, or commit fraud. For example, voices that have been anonymized may be exploited in social engineering schemes or to spread disinformation by enjoying anonymity and avoiding accountability. Furthermore, since it becomes increasingly difficult to distinguish between a real speaker and a converted speech, VC systems may have a profound and varied social impact. The widespread deployment of VC systems may damage people's confidence and trust in voice-based communication services. To address these issues, appropriate legal frameworks and ethical guidelines for using anonymization technologies must also be put in place to reduce these risks of misuse. With the introduction of the Digital Services Act (DSA) and AI-Act on European level, the European Union is pioneering these guidelines by imposing concepts of transparency, labeling and marking obligations, and human oversight to AI components suitable to cause harmful social impact. However, as national lawful guidelines are just being conceived and scripted at the moment, concrete restrictions and applied methods can be expected in the upcoming 1-2 years only.

6. Acknowledgements

This research has been partly funded by the Federal Ministry of Education and Research Germany (BMBF 16KISA007, project Medinym) and partly by the Volkswagen Foundation.

7. References

[1] A. Das, S. Ghosh, T. Polzehl, and S. Stober, "Stargan-vc++: Towards emotion preserving voice conversion using deep embeddings," *arXiv preprint arXiv:2309.07592*, 2023.

[2] T. Walczyna and Z. Piotrowski, "Overview of voice conversion methods based on deep learning," *Applied Sciences*, vol. 13, no. 5, p. 3100, 2023.

[3] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 3, pp. 806–817, 2011.

[4] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The nu-naist voice conversion system for the voice conversion challenge 2016." in *Interspeech*, 2016, pp. 1667–1671.

[5] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3893–3896.

[6] H. Ming, D.-Y. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion." in *Interspeech*, 2016, pp. 2453–2457.

[7] 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.

[8] Z. Lian, Z. Wen, X. Zhou, S. Pu, S. Zhang, and J. Tao, "Arvc: An auto-regressive voice conversion system without parallel training data." in *INTERSPEECH*, 2020, pp. 4706–4710.

[9] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[10] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," *arXiv preprint arXiv:1804.02812*, 2018.

[11] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[12] Y. A. Li, A. Zare, and N. Mesgarani, "Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," *arXiv preprint arXiv:2107.10394*, 2021.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[14] S. Ding and R. Gutierrez-Osuna, "Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion." in *Interspeech*, 2019, pp. 724–728.

[15] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-y. Lee, "Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5954–5958.

[16] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," *arXiv preprint arXiv:2106.10132*, 2021.

[17] S. Yang, M. Tantrawenith, H. Zhuang, Z. Wu, A. Sun, J. Wang, N. Cheng, H. Tang, X. Zhao, J. Wang *et al.*, "Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion," *arXiv preprint arXiv:2208.08757*, 2022.

[18] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[19] A. Polyak, L. Wolf, Y. Adi, and Y. Taigman, "Unsupervised cross-domain singing voice conversion," *arXiv preprint arXiv:2008.02830*, 2020.

[20] S. Kum and J. Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Applied Sciences*, vol. 9, no. 7, p. 1324, 2019.

[21] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[22] 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.

[23] 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.

[24] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[25] P. Puchtler, J. Wirth, and R. Peinl, "Hui-audio-corpus-german: A high quality tts dataset," in *KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, September 27–October 1, 2021, Proceedings 44*. Springer, 2021, pp. 204–216.

[26] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.

[27] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[28] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[29] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[31] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, "The voiceprivacy 2022 challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[32] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 912–919.

[33] Y. Ren, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Revisiting oversmoothness in text to speech," *arXiv preprint arXiv:2202.13066*, 2022.