System Description for Voice Privacy Challenge 2024

Arnab Das[1]*, Carlos Franzreb*[1]*, Tim Herzig*[1]*, Philipp Pirlet*[1]*, Tim Polzehl*[1,2]

[1]Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI), Germany [2]Quality and Usability Lab, Technische Universität Berlin, Germany

{arnab.das, carlos.franzreb, tim.herzig, philipp.pirlet, tim.polzehl}@dfki.de

Abstract

The growing use of speech-based cloud devices and services has heightened the risk of identity theft and misuse of personal information. Speech anonymization techniques help exercise our right to privacy and shield us from falling prey to such malpractices. In this paper, we propose three speech anonymization systems to be submitted to the Voice Privacy Challenge 2024 and describe them in detail. Voice anonymization systems often lack utility for downstream applications, resulting in issues like poor emotion preservation or low intelligibility. This has led to research focused on balancing the privacy-utility tradeoff. We propose two methods, that use the KNN-based voice conversion (VC) system as a core anonymization method and show improved intelligibility and emotion preservation. We also propose to employ a vector quantized mutual information-based VC system that learns to distinguish between speaker and content features and alters speaker information during inference time to achieve speaker anonymity. We evaluate these two types of voice conversion systems within the framework of speaker anonymization and analyze the utility-privacy trade-off achieved by each system.

Index Terms: voice anonymization, privacy-utility tradeoff

1. Team Details

The name of our team is DFKI_SLT.

2. System Description

We plan to submit results for three anonymization systems for VPC 2024. As described in the VPC 2024 evaluation plan all of the proposed systems are designed to achieve utterance level anonymization.

2.1. Anonymization using single layer knn-vc: KNNS

The core idea of this system is based on knn-vc as proposed in [1]. The system architecture diagram is depicted in Figure 1. The source utterance is first processed via WavLM Large [2], which outputs speech representation of size $F \times 24 \times 1024$, where F is the number of frames, 1024 is the latent dimension and 24 signifies the output of all transformer layers. WavLM outputs 50 frames from a 1 s long waveform sampled at 16 kHz rate. Then as advised in knn-vc, only the output of the 6^{th} layer is selected, which makes the speech representation of size $F \times 1 \times 1024$, called as *query set*. For anonymization, first, a random target speaker is chosen from all the English speakers present in the emotional speech database (ESD)[3]. Consequently, all the utterances of the chosen target speaker are processed through WavLM, which gives rise to a large feature array of size $N \times F \times 24 \times 1024$, where N is the number of utterances for the chosen random target speaker. Again we chose only the output of the 6^{th}, which makes the feature array of size $N \times F \times 1 \times 1024$. This is called the *matching set*. Then a KNN operation is performed for each frame of the query set to the matching set and the top-4 nearest neighbors are selected, producing a *matched set* of size $F \times 4 \times 1024$. Then it goes through an average pooling operation to average the top 4 neighbors bringing down the matched set size to $F \times 1024$. Finally, a HiFiGAN [4] is applied to reconstruct back the audio signal in the time domain. The HiFiGAN is pre-trained on Librispeech: train-clean-360 [5], ESD, CREMA-D [6] and RAVDESS [7] datasets.

2.2. Anonymization using multi-layer knn-vc: KNND

The working principle of this KNND system is similar to the previously discussed KNNS system in 2.1 and the system diagram is shown in Figure 2. The difference is, that instead of selecting only the output of the 6^{th} layer of WavLM, we select the output of both 6^{th} and 12^{th} layer. In our experiments, we discovered that the 12^{th} layer encodes emotional cues more effectively than the 6^{th} layer. Therefore, we opted to incorporate both layers to enhance emotion preservation. This increases the size of the query set and matching set to $F \times 2 \times 1024$ and $N \times F \times 2 \times 1024$ respectively. Consequently, the matched feature also becomes an array of size $F \times 2 \times 1024$. To accommodate this additional dimension, the HiFiGAN is augmented with a convolutional *PreNet* module, which gets pre-trained jointly with the HiFiGAN before applying to the anonymization pipeline. In both KNNS and KNND, except the HiFiGAN component nothing else is trainable.

2.3. Anonymization by disentangled representation using vqmivc: VMC

The core of the vqmivc-based (in short VMC) anonymization systems is a VC system proposed in [8] which disentangles speech representations using vector quantization and lowering mutual information in an unsupervised way. The anonymization pipeline for the VMC system is depicted in Figure 3. The core VC framework has four major components, an F0 extractor, a speaker encoder, a content encoder, and a decoder.

Step I - Training: During training the F0 extractor receives the raw source waveform and uses the *pyworld-dio* to extract temporal F0 features of share $F \times 1$, where F is the number of frames. The speaker encoder takes input from a Mel spectrogram of the source utterance of share $F \times 80$ and produces a global speaker feature vector of shape 1×256. The content encoder also receives the same Mel spectrogram and produces content feature representation of size $F/2 \times 512$. The content encoder comprises convolution layers and a code book to quantize the content features. The disentanglement of F0 features, speaker features, and content features is achieved by minimizing parameterized mutual information as described in [8]. Afterward, the content feature is upsampled to achieve the shape $F \times 512$ and the global speaker feature is repeated in temporal dimension to achieve the shape of $F \times 256$. Consequently, the three

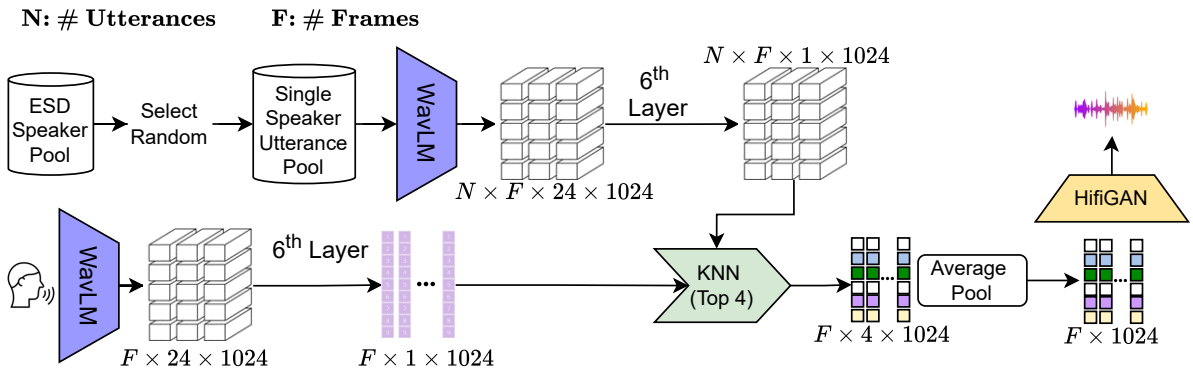

Figure 1: *System architecture diagram for KNNS*

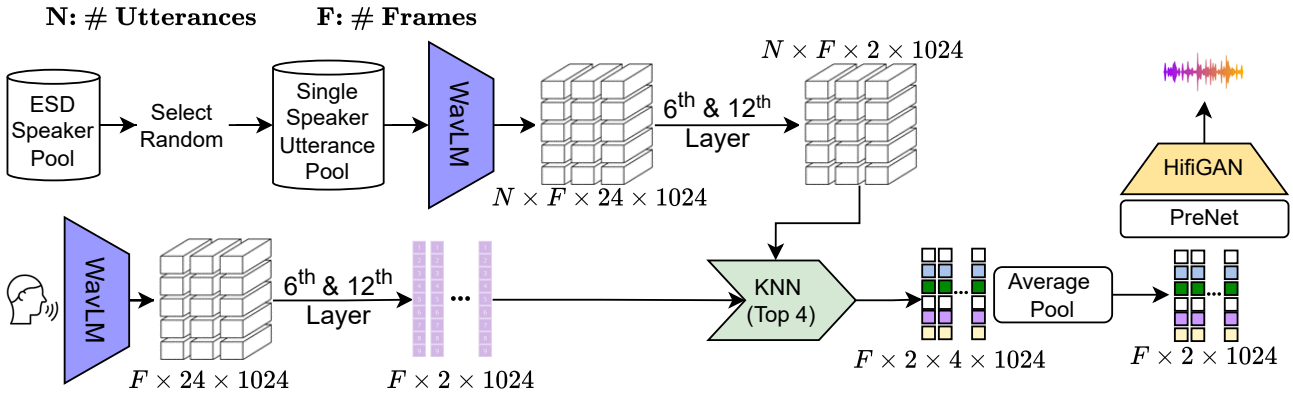

Figure 2: *System architecture diagram for KNND*

Table 1: *Modules and training corpora for the anonymization system VMC*

Module	Description	Output	Training Data
F0 Extractor	Pyworld dio and stonemask Input: Raw waveform	Normalized F0 contour of shape $F \times 1$	-
Speaker Encoder	Convolution Layers + temporal global pool Input: Mel Spectrogram of shape $F \times 80$, with hop length 160 and 80 Mel bands	1×256 speaker feature	LibriSpeech:train-clean-360 ESD RAVDESS CREMA-D
Content Encoder	Convolution Layers + vector quantization Input: Mel Spectrogram of shape $F \times 80$	$F/2 \times 512$ vector quantized content feature	

speech features are concatenated and fed to a decoder which outputs the original Mel spectrogram. The whole VC framework is trained unsupervised by optimizing reconstruction loss. We have kept the architecture of each module the same as recommended in [8]. Implementation details of individual components are outlined in Table 1. Finally, a HiFiGAN is applied to reconstruct back the waveform from the Mel spectrogram. Similar to KNNS and KNND systems, the HiFiGAN is pre-trained on Librispeech: train-clean-360 [5], ESD, CREMA-D [6] and RAVDESS [7] datasets.

Step II - Inference: During the anonymization, a random target utterance is selected from the ESD utterance pool, and the speaker encoder is fed with the Mel spectrogram of the target utterance whereas the other two components F0 extractor and the content encoder receive the source utterance. This produces the output utterance in the target speaker's voice thus effectively archiving anonymization.

3. Experiment Setup

3.1. Datasets

For training the HiFiGAN modules in all three systems and the VQMIVC voice converted for the VMC systems we used in total of four datasets - Librispeech: train-clean-360 [5], ESD [3], CREMA-D [6] and RAVDESS [7]. The train-clean-360 subset contains more

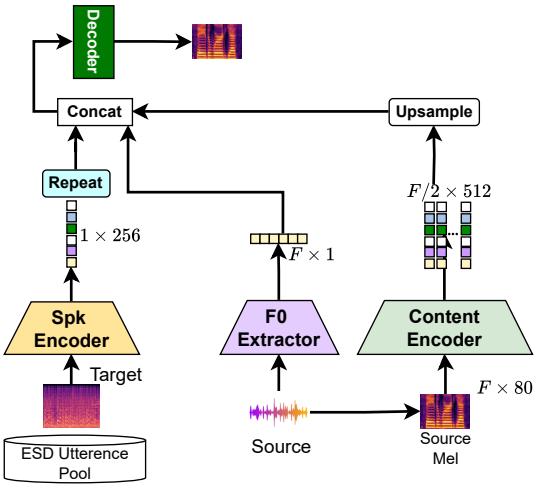

Figure 3: *System architecture diagram for VMC*

than 960 hours of clean speech data from 921 speakers (439 female and 482 male) with an average of 25 minutes of speech per speaker. The ESD database includes parallel utterances conveying 5 emotion categories (neutral, happy, angry, sad, and surprised), spoken by 10 native English (gender-balanced) speakers. CREMA-D dataset contains $>$7000 clips from 91 actors (48 male and 43 female) covering 6 different emotions - Anger, Disgust, Fear, Happy, Neutral, and Sad. In RAVDESS, 24 professional actors (12 females and 12 males) uttered parallel emotional contentment depicting 7 different emotions - calm, happiness, sadness, anger, fear, surprise, and disgust. The evaluation of the proposed systems is performed on the dev-clean and the test-clean subsets from the Librispeech corpus and the IEMOCAP [9] dataset is used as prescribed by the VPC challenge.

3.2. Objective metrics

The anonymization systems are evaluated using the models suggested by the VPC challenge. To objectively measure the privacy benefits, an attack using automatic speaker verification (ASV) is applied and an equal error rate (EER) score is computed. The attack uses both a semi-informed scenario alternative to an ignorant one. In the ignorant case, the ASV model is trained on real utterances and applied to anonymized utterances. Alternatively, in the semi-informed scenario, the ASV model is trained on the anonymized utterances from the training subset and then applied to the utterances from dev-clean and test-clean subsets. Both of these scenarios consider the anonymization system as a black box. Evidence shows that a semi-informed attack is stronger than an ignorant one.

To measure the utility of the anonymized utterances, automatic speech recognition (ASR) and speech emotion recognition (SER) models are applied to judge the intelligibility and amount of emotion preservation respectively. For intelligibility, word error rate (WER), and for emotion preservation unweighted average recall is reported.

4. Results & Discussion

The results for objective evaluation to measure privacy strength are summarized in Table 2. We compare our proposed systems with the signal processing dependent McAdams (system ID - **B2**) baseline and DL model dependent GAN-based anonymization (system ID - **B3**) baseline system as reported in [10]. The attack scenario is semi-informed as the ASV system is trained on anonymized data. $B2$ archives EER scores of 12.91 and 2.05 in the dev subset for female and male trail utterances respectively. On the test subset, the EER scores are even lower, 7.48 and 1.56 respectively for female and male trials. GAN-based $B3$ baseline archives much higher EER scores indicating stronger anonymization. For $B3$, EER scores in the dev subset are 28.43 and 22.04, whereas for the test subset, the scores are 27.92 and 26.72 for female and male trials. Among the 3 proposed systems, VMC achieves the highest EER scores for both the subsets and KNND EER scores are the second best. For the dev subset, VMC archives EER scores of 19.462 and 8.54, closely followed by the KNND system, which archives 17.07 and 8.54 EER scores for female and male trail utterances. For the test subset, EER scores for the VMC system are 16.79 and 11.8 and for the KNND system the scores are 10.77 and 9.36 Comparatively, EER scores for the $KNNS$ system are a bit low. It achieves EER scores of 12.64 and 4.19 on the dev subset and for the test subset, the scores are 9.51 and 4.45. So in terms of anonymization strength, all the proposed systems beat the $B2$ baseline but are somewhat inferior to the $B3$ baseline. Noticeably, for all the systems including the baseline EER score for Female trial utterances are higher than that of males, which indicates it's easier to anonymize female speakers than male speakers.

Table 2: *Objective evaluation results for privacy metric - EER using the semi-informed scenario. Among the proposed systems, the best scores are in bold, and the second-best scores are underlined per gender.*

Systems	**ASV EER ↑**			
	Librispeech			
	Dev		Test	
	Female	Male	Female	Male
McAdams - B2 [10]	12.91	2.05	7.48	1.56
GAN - B3 [10]	28.43	22.04	27.92	26.72
KNNS	12.64	4.19	9.51	4.45
KNND	17.07	**8.54**	10.77	9.36
VMC	**19.462**	**8.54**	**16.79**	**11.8**

Table 3: *Objective evaluation results for utility metrics - WER for ASR and mean UAR (5 folds) for SER. Among the proposed systems, the best scores are in bold, and the second-best scores are underlined.*

Systems	**SER UAR [%] ↑**		**ASR WER [%] ↓**	
	IEMOCAP		Librispeech	
	Dev	Test	Dev	Test
Original [10]	69.08	71.06	1.8	1.85
McAdams - B2 [10]	55.61	53.49	10.44	9.95
GAN - B3 [10]	38.09	37.57	4.29	4.35
KNNS	**46.94**	**47.69**	2.6	2.52
KNND	43.2	45.54	**2.47**	**2.49**
VMC	35.14	33.92	20.75	18.25

The results of the objective utility evaluation are presented in Table 3. In terms of UAR for the SER task, the scores are 69.08% and 71.06% for the dev and test subsets of IEMOCAP data respectively when tested on original utterances. However, the $B2$ baseline archives UAR scores of 55.61 and 53.49 whereas the scores achieved by the $B3$ baseline are way lower, 38.09 and 37.57 for the dev and test subset respectively as reported in [10]. Among the proposed systems, KNNS achieves the highest UAR scores for the SER task closely followed by the KNND system. The KNNS system archives 46.94% and 47.69% on the dev and test subset respectively whereas the KNND system achieves 43.2% and

45.54%. The emotion preservation capability for the VMC system is comparatively lower as it achieves 35.14% and 33.92% on the IEMOCAP dev and test subsets respectively.

In terms of intelligibility evaluation, the original data get WER scores of 1.8% and 1.85% on Librispeech dev and test subsets respectively. For the $B2$ baseline the scores are 10.44% and 9.95% however the $B3$ baseline achieves better intelligibility as the WER scores are 4.29% and 4.35%. This shows that even though the $B3$ baseline is not as good as $B2$ in terms of emotion preservation for intelligibly of anonymized speech $B3$ is much better than $B2$ Among the proposed systems, both KNND and KNNS systems significantly beat both the baseline systems as the WER scores are very low and almost analogous to the original data. The best WER scores are achieved by the KNND system, 2.47% and 2.49% respective on the dev and test subset, whereas KNNS achieves 2.6% and 2.52%. Similar to the SER task, VMC is lacking on the ASR task as well, as it achieves high WER scores of 20.75% and 18.25%.

The privacy and utility results presented in Table 2 and Table 3 clearly show that we get higher privacy at the cost of lower utility of the anonymized speech. For the GAN-based $B2$ baseline system the privacy scores are high but the utility scores are low, especially for the emotion preservation task. Our proposed KNND system achieves the best balance of privacy and utility dimensions, as evidenced by significantly higher WER scores than other systems, higher emotion preservation compared to the $B3$ baseline, and higher privacy compared to the $B2$ baseline. The proposed KNNS is also capable of producing high utility scores for the anonymized data as it beats the $B2$ baseline in both SER and ASR tasks however in terms of speaker privacy the performance is a bit lower than the KNND system. The VMC system, which archives anonymization by disentangling content and speaker features, can archive high privacy for anonymized speech but has poor utility.

5. References

[1] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," *arXiv preprint arXiv:2305.18975*, 2023.

[2] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

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

[4] 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.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[6] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[7] 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.

[8] 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.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[10] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, "The voiceprivacy 2024 challenge evaluation plan," *arXiv preprint arXiv:2404.02677*, 2024.