

Investigating Voice Conversion Architecture with Different Bottleneck Features for the Voice Privacy Challenge 2024

Seymanur Akti^{1,*}, Tuan Nam Nguyen^{1,*}, Yining Liu¹, Alex Waibel^{1,2}

¹Karlsruhe Institute of Technology (KIT), Germany

²Carnegie Mellon University (CMU), USA

seymanur.akti@kit.edu, tuan.nguyen@kit.edu, yining.liu9@kit.edu,
alexander.waibel@kit.edu

Abstract

This paper outlines KIT’s submission to the Voice Privacy 2024 Challenge. We present a system designed for both cascaded and end-to-end conditions. In the end-to-end condition, we explore a conditional variational autoencoder enhanced with normalizing flow and adversarial training, proposing methods for extracting clean content information. We disentangle content by imposing an information bottleneck on self-supervised pretrained representations (such as WavLM and HuBERT), ASR bottleneck features, and ASR output in logits form. By incorporating pitch information, we aim to anonymize the original audio while preserving its prosody. In the cascaded condition, we first obtain the transcript using the reliable Whisper speech recognition system and then input this transcript into a TTS model that preserves prosody to anonymize the original audio.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Speech data can reveal a lot of personal details when listened to or analyzed by machines, even without a legal privacy definition. These details can include age, gender, ethnicity, location, physical or emotional state, political views, and religious beliefs. Technologies can identify the speaker, and features used in automatic speaker verification (ASV) systems can be linked to personal information. If not protected, publicly available speech samples could be misused for theft or fraud. Therefore, it’s important to find ways to protect against this risk. One solution is to use voice conversion to make speech anonymous.

The Voice Privacy Challenge aims to address several issues related to voice anonymization, which is the first step towards protecting privacy. This involves altering a speaker’s voice to hide their identity while preserving other speech characteristics, such as emotions, states, and content. Although voice anonymization is an appealing solution, the extent of privacy protection it offers is uncertain due to the lack of formal task specifications, attack models, shared datasets, protocols, and measurements. The series of Voice Privacy Challenges seek to resolve these ambiguities and provide clear guidelines and standards.

In this paper, we present our submission to the Voice Privacy 2024 Challenge, exploring two primary approaches. The first approach utilizes a voice conversion model similar to the provided baselines B5 and B6. We aim to improve these baselines by incorporating a conditional variational autoencoder with a pitch encoder and normalizing flow techniques. Ad-

ditionally, while the baselines use the bottleneck output from ASR to separate content and eliminate speaker information from audio, we investigate various representations, such as self-supervised pre-trained representations, as inputs.

The second approach is based on the pipeline of a speech recognition and text-to-speech system, similar to the baseline B3. Using transcripts from a state-of-the-art speech recognition model, we employ multi-speaker text-to-speech synthesis with a prosody controller to generate audio with prosody similar to the original. For both our end-to-end and cascaded approaches, we also explore various strategies for selecting the target speaker for the generated audio, as described in the section above.

The rest of the paper is organized as followed. Section 2 provides the description of both the cascade and the end-to-end system. It is then followed by Section 3 describing the data set used to train and test the system, and experimental results. In the end, we conclude the paper with Section 4.

2. Proposed Voice Anonymization Systems

In this section, we provide a detailed description of the various models employed for voice anonymization. We have experimented with two basic approaches which are adapting voice conversion system for voice anonymization and cascaded approach where the speech is transcribed with a state-of-the-art ASR model and then re-synthesized with a TTS model.

2.1. Voice Conversion Models

For the voice conversion-based voice anonymization task, we employed FreeVC [1], a state-of-the-art zero-shot voice conversion model. FreeVC is a conditional variational autoencoder (VAE) speech-synthesis model inspired by VITS [2]. The overall architecture could be seen in the Figure 1. It utilizes pretrained WavLM [3] as the content encoder, which is a self-supervised large-scale model for speech representations. These content embeddings are used to create the prior variational latent space by passing through the bottleneck extractor which can also be referred as prior encoder.

For speaker embeddings, FreeVC uses a fixed length speaker encoder module to extract speaker features. Assuming that the speaker information is missing in the prior latent space, these speaker embeddings are expected to inject the speaker information from target speaker.

The model then employs an autoencoder architecture for posterior encoding and decoding. The posterior encoder encodes the speech into a variational latent space, capturing both linguistic and acoustic features. A normalizing flow layer maps the posterior latent space to the prior latent space conditioned on speaker features. The posterior decoder synthesizes high-quality audio from the posterior latent space features using

*Equal contribution

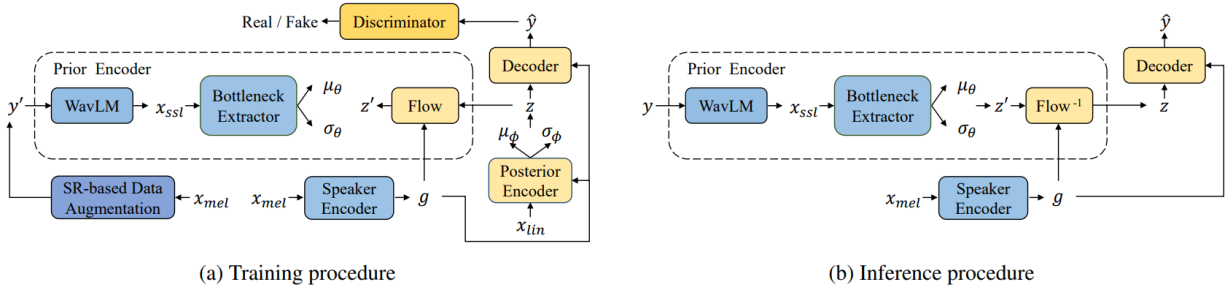


Figure 1: Overall architecture of FreeVC. Image taken from [1].

an adversarial training strategy similar to the state-of-the-art vocoder model Hifi-GAN [4].

In the end, the input encoders disentangle the content and speaker information and posterior decoder aims to reconstruct the audio from these disentangled features by learning the mapping between prior and posterior latent spaces.

The model is trained with the reconstruction objective where no speaker labels were used. During the inference phase, inverse flow is used to estimate the posterior representation from disentangled content and speaker features of different audios. Then, posterior decoder synthesizes the speech.

FreeVC is considered as a good fit for the task due to its high-speed and high-quality speech synthesis for textless one-shot voice conversion setting. Additionally, speaker augmentation applied during the training phase helps the content encoder eliminate speaker-related information from the content embeddings while preserving emotion-related prosodic features to some degree.

For training the FreeVC-based models, we followed the author’s instructions in the official GitHub page¹. Then, we modified the model by changing the content encoder part to investigate the effect of different bottleneck features. All variations of FreeVC-based models are listed below:

- **FreeVC-WavLM:** This variation uses WavLM as the content encoder. WavLM is known for its capability of extracting speech representations while keeping the prosodic information. However, this also means that the WavLM embeddings might have some leakage of source speaker features. In order to normalize the content embeddings, authors proposed using the pitch shift augmentation for speaker anonymization at the input size.
- **FreeVC-Hubert:** Considering the disadvantage of FreeVC-WavLM in potentially leaking speaker information from the source speaker, which could compromise privacy, we explored using HuBERT [5], another pre-trained speech representation model. We extracted the HuBERT units from the 9th layer of HuBERT-Base model and used k-means quantizer for getting discrete units. And trained the model with these features replacing the WavLM features. HuBERT’s training strategy with clustering is designed to map the same content from different speakers to closer positions in the representation space. This creates a narrower bottleneck for content extraction, effectively discarding speaker-related information.

¹<https://github.com/OlaWod/FreeVC>

- **FreeVC-ASR-bottleneck-F0 and FreeVC-ASR-logits-F0**
To further narrow the bottleneck, we trained a Transformer-based speech recognition model similar to [6]. The goal of training a speech recognition model is to minimize the cross-entropy between the predicted logits and the ground truth, with the expectation that speaker-specific information will be removed in the final layer. We then use the ASR bottleneck feature from this layer, as well as the ASR output in logits form, as input features to train the FreeVC model to synthesize audio from these features. Additionally, we train FreeVC with an F0 Encoder (similar to baselines B5 and B6) to enhance the bottleneck features with prosodic information, helping to retain the emotional quality of the original audio. This approach compensates for the loss of prosodic information at the content bottleneck.

After training all models for voice conversion task, in order to adapt the models for voice anonymization, we removed the speaker encoder during the anonymization process and created pseudo-speaker embeddings. Two different strategies were employed for assigning the pseudo-speakers:

- **Random Speaker (rs):** In this setup, the speaker embeddings are randomly generated vectors with the same size of the speaker encoder output (256). Pseudo-speaker is different for each utterance.
- **Single Speaker (ss):** In this setup, pseudo-speaker was chosen among the speakers from the training set which is VCTK [7] for FreeVC-based models. Then, each utterance were converted to the same pseudo-speaker.

2.2. Cascaded System with ASR + TTS

In the cascade condition, the Whisper ASR model [8] receives audio inputs and generates raw transcripts, which will then pass through a TTS model to generate the audio with same content. The output of TTS is the output of system. Here, the pseudo-speaker is again a single speaker for all utterances since single-speaker TTS model was used for speech synthesis. To retain prosodic information, we again use F0 encoder and calculate the F0 for each phoneme input using a framework similar to FastSpeech 2 [9].

3. Experimental Results

Models with different bottleneck features and anonymization strategies were used for anonymizing the datasets given by the challenge organizers, and evaluated based on the given metrics.

Table 1: WER, UAR and EER results for the given dev and test sets.

	WER % ↓			UAR % ↑			EER % ↑				
	dev	test	avg	dev	test	avg	dev-f	dev-m	test-f	test-m	avg
B1	3.07	2.91	2.99	42.71	42.78	42.75	10.94	7.45	7.47	4.68	11.75
B2	10.44	9.95	10.20	55.61	53.49	54.55	12.91	2.04	7.48	1.56	6.00
B3	4.29	4.35	4.32	38.09	37.57	37.83	28.43	22.04	27.92	26.72	26.28
B4	6.15	5.90	6.02	41.97	42.78	42.37	34.37	31.06	29.37	31.16	31.49
B5	4.73	4.37	4.55	38.08	38.17	38.12	35.82	32.92	33.95	34.73	34.35
B6	9.69	9.09	9.39	36.39	36.13	36.26	25.14	20.96	21.15	21.14	22.10
FreeVC-WavLM-rs	5.53	4.97	5.25	40.52	39.54	40.03	25.00	14.31	12.41	8.24	14.99
FreeVC-WavLM-ss	2.76	2.68	2.72	48.27	43.43	45.85	14.92	8.23	11.83	7.79	10.69
FreeVC-HuBERT-Base-rs	8.71	8.12	8.42	35.04	34.22	34.63	33.38	26.23	28.10	28.94	29.16
FreeVC-HuBERT-Base-ss	6.12	5.58	5.85	37.23	35.84	36.54	32.38	50.77	24.27	47.21	38.66
FreeVC-ASR-bottleneck-F0-ss	3.16	3.12	3.14	38.69	39.27	38.98	36.49	32.79	30.84	35.86	34.00
FreeVC-ASR-logits-F0-ss	3.70	3.50	3.60	37.27	36.58	36.93	31.53	22.02	24.24	26.55	26.08
FreeVC-ASR-bottleneck-F0-rs	3.22	3.10	3.16	39.45	37.18	38.32	28.11	16.46	17.33	19.37	20.32
Cascaded ASR + TTS	6.19	6.55	6.37	31.08	30.02	30.55	36.07	22.51	26.30	24.54	27.36

Three metrics were set by the organizers as:

- **WER:** Word error rate of the anonymized audio transcriptions of LibriSpeech dataset [10] for assessing the linguistic information preserving ability of the anonymization system.
- **UAR:** Unweighted average recall for the emotion recognition task on IEMOCAP dataset [11] to assess the prosody preserving ability of the anonymization system.
- **EER:** Equal error rate from the speaker verification model which is trained on the LibriSpeech anonymized data. Higher EER score corresponds to higher privacy, meaning that the speaker verification model is not able to successfully retrieve the original speaker identity.

The comparative results were given at Table 1 with all evaluated systems. The results point that the random speaker anonymization hurts the linguistic and acoustic information transfer with higher WER and lower UAR compared to their counterparts with single speaker anonymization. This behaviour is due to random speaker embeddings being unstable and often cause artifacts in the output since those speaker embeddings are not from a real speaker. In the other hand, single speaker anonymization systems provide a better voice conversion performance using a well-defined speaker embedding from a real speaker which model has already seen during the training.

When comparing voice anonymization systems based on different bottleneck modules for the content embeddings, the results show that as we narrow the bottleneck (FreeVC-WavLM to FreeVC-HuBERT and FreeVC-ASR), the privacy score improves due to reduced leakage of acoustic information from the

source speaker. However, this also means that prosodic information related to expressions might be also eliminated, which explains the lower UAR score. Additionally, WER scores indicate that FreeVC-ASR variations effectively preserve linguistic information surpassing the FreeVC-HuBERT-ss and showing minimal difference compared to FreeVC-WavLM-ss. This highlights the contribution of the speech recognition module in maintaining linguistic integrity. The cascade of ASR and TTS is worse than voice conversion based models across all metrics.

Based on these conclusions, 2 systems were submitted for the challenge. System names and module descriptions were given at Table 2.

4. Conclusion

For the Voice Privacy Challenge 2024, we investigated the voice anonymization capability of the one-shot voice conversion model FreeVC [1]. To enhance the original approach, we addressed the issue of source speaker information leakage, which is critical for high-privacy voice anonymization. By utilizing different bottleneck modules for content extraction, we reduced the effect of this problem, achieving a higher privacy score. Furthermore, to minimize the loss of linguistic and prosodic information, we integrated additional components. Our preliminary evaluation results indicate that these enhanced voice conversion models outperform the baselines on several metrics, demonstrating the effectiveness of our improvements.

Table 2: Module descriptions and training data for the submitted voice anonymization systems.

System Name	Pre-trained Models	Modules	Training Data
freevc_hubert_base_ss	HuBERT Base [5] L9 km500 quantizer	Speaker Encoder: 3-layer LSTM Prior & Posterior Encoders: WaveNet [12] Posterior Decoder: Hifi-GAN [4]	VCTK [7]
freevc-asr-bottleneck-f0		CTC-ASR model: 12 layer Transformer Speaker Encoder: 3-layer LSTM Prior & Posterior Encoders: WaveNet [12] Posterior Decoder: Hifi-GAN [4]	LibriSpeech for ASR [10] VCTK [7]

5. References

- [1] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [2] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.
- [3] 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.
- [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] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [6] N.-Q. Pham, T.-L. Ha, T.-N. Nguyen, T.-S. Nguyen, E. Salesky, S. Stueker, J. Niehues, and A. Waibel, "Relative positional encoding for speech recognition and direct translation," 2020.
- [7] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, vol. 6, p. 15, 2017.
- [8] 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.
- [9] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," 2022.
- [10] 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.
- [11] 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.
- [12] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, vol. 12, 2016.