Orange SHIVA system description for the Voice Privacy Challenge 2024

Olivier Le Blouch[1]*, Rayane Bakari*[1]*, Nicolas Gengembre*[1]

[1]Orange, France

olivier.leblouch@orange.com, mohamedrayane.bakari@orange.com,
nicolas.gengembre@orange.com

Abstract

This paper introduces our systems submitted to Voice Privacy Challenge 2024. The aim of this work is to analyze potential levers for improving existing voice conversion frameworks used as voice anonymization systems. We concentrated our efforts on an available and modular voice conversion framework named DISSC. Its modularity enables an ablation study focused on improving the three main expected qualities of anonymized speech, namely its intelligibility, the anonymization robustness and the preservation of the original expressiveness. The experiments carried out confirmed a number of assertions, such as the importance of using datasets rich in expressiveness, or of modifying the prosody of the source speaker to improve the quality of anonymization.

Index Terms: voice privacy, voice conversion, anonymization, prosody, emotion embeddings

1. Introduction

In today's digital age, the collection and analysis of voice data have become pervasive, enabling a wide range of applications such as voice assistants, voice recognition systems, and voice-based authentication. However, the increasing reliance on voice data raises concerns about privacy and the potential misuse of personal information. Voice anonymization, a technique aimed at protecting the identity of individuals in voice recordings, has emerged as a crucial solution to address these concerns.

Voice anonymization involves altering the characteristics of a person's voice to prevent their identification while preserving the overall intelligibility and naturalness of the speech. By obfuscating the unique vocal traits that can be used to identify an individual, voice anonymization techniques provide a layer of privacy protection, ensuring that sensitive information remains confidential. Specifically, according to the VoicePrivacy 2024 Challenge [1], the speaker anonymity system needs to satisfy: (i) conceal the speaker identity, (ii) preserve the linguistic content and (iii) preserve paralinguistic attributes.

According to voice privacy challenge (VPC) 2024, the proposed anonymization system should provide a trade-off between privacy protection and utility preservation. Privacy protection assesses the system's ability to prevent automatic speaker verification (ASV) from identifying the speaker after anonymization. Utility encompasses the quality of transcripted anonymized speech using an automatic speech recognition (ASR) and the preservation of emotional content as measured by speech emotion recognition (SER).

So, to quantify these aspects, VPC 2024 has defined a metric dedicated to each of them, respectively (i) the equal error rate (EER) measuring the non-recognition of the source speakers (higher is better, optimal value 50%), (ii) the Word Error Rate (WER) measuring the amount of words correctly transcripted in the anonymized signal (lower is better), and (iii) the Unweighted Averaged Recall (UAR) measuring the recognition of original emotions in anonymized voices (higher is better).

Various approaches to safeguard speaker privacy utilize not only digital signal processing (DSP) but also neural speech conversion and synthesis. The first approach consists in modifying, with signal processing techniques, instantaneous speech characteristics such as pitch, spectral envelope and time scaling. The second approach mainly focuses on fine-grained disentangled latent representation learning which supposes that speech can be decomposed into content, speaker identity and prosodic representation. The latter representation contains some paralinguistic attributes such as intonation, emotion, style, accent, elocution. Recent work has shown that prosodic representation is able to identify speakers even after anonymization [2].

VPC 2024 [1] has introduced six primary baselines for voice anonymization. The baseline B2 is the only one that uses the DSP approach, which uses the McAdams coefficient to randomize the anonymization method by shifting the pole positions derived from linear predictive coding (LPC) analysis of speech signals [3]. All the other baselines are based neural synthesis. The first baseline, denoted B1 proposed in [4], involves transforming three distinct components extracted from a waveform: an F0 representation; a speaker representation encoded by x-vector; a linguistic content encrypted by bottleneck features. B3 as proposed in [5, 6] employs an ASR+TTS based system. It modifies prosody by doing value-wise multiplication of F0 with random values in [0.6, 1.4), uses GAN to create artificial speaker embeddings and a connectionist-temporal-classification/attention hybrid ASR model to obtain phonetic transcriptions. Baselines B3, B4, and B5 take benefit from the observation that speaker information can be minimized through quantization. The approach proposed in [7] leverages the capabilities of neural audio codec (NAC) to develop a speaker anonymization system in their study, which is denoted B4. The authors in [8] introduced an anonymization pipeline based on quantization, which incorporates vector quantization (VQ) to enhance the separation of linguistic and speaker characteristics, building upon a similar concept as B1. Thus, B5 and B6 draw inspiration from this research.

The remainder of this paper is organized as follows. Section 2 describes our proposed methods in detail. Section 3 represents the experiment setup and experiment results. Conclusions and future plans are presented in section 4.

2. Proposed method

This section discusses the proposed methods in which we tune DISSC, an existing voice conversion framework developed in

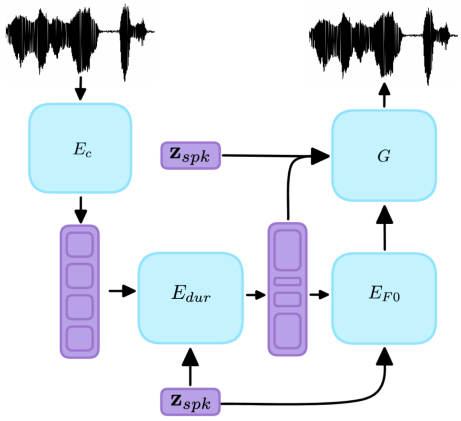

Figure 1: *Original DISSC scheme, taken from [9]*

[9], to build an anonymization system.

Figure 1, from [9], describes the global DISSC scheme which follows the following steps: extraction of a discrete Hubert-based [10] representation E_c from raw audio signal, prediction of new phonemes duration E_{dur} and new pitch E_{F0} conditioned on a target speaker embedding z_{spk}, and then generation of a converted audio signal from this units/pitch/durations/speaker set via a generator G. The modularity of the scheme and the availability of code and pretrained models [1] make that framework particularly interesting to work with.

Considering the VPC2024 objectives, most voice conversion systems can be used as anonymization engines, provided that the target voices are chosen in such a way as to break any possible identification between the source audio and the generated audio. This is what we do in this study, by choosing, for each source sentence, a target voice randomly selected from the pool of speakers in the VCTK dataset, which is used to train the generator G and the predictors E_{dur} and E_{F0} .

The approach that interest us in this paper relies on several assumptions that we seek to validate:

- Anonymization systems should take advantage of expressive datasets to properly convey expressiveness
- Inject pretrained emotional embeddings in anonymization systems can help to improve expressiveness preservation
- Prosody conveys speaker information, so modifying prosody improve anonymization
- Prosody and expressiveness are closely intricated, so modifying prosody degrades expressiveness if not dealt subtely

With these intuitions in mind, the main objectives are to improve expressiveness preservation through the anonymization pipeline and to confirm the role played by the prosody in the full pipeline and use it to build the best anonymization system upon the DISSC framework.

3. Experiments

To discuss about how to improve a voice converter such as DISSC to perform anonymization, 6 experiments are proposed in this section. Each of them tends to confirm step-by-step our aforementioned intuitions.

[1] https://github.com/gallilmaimon/DISSC

3.1. Resources

In addition to paper and code, DISSC authors propose several pretrained models trained on VCTK [11] and ESD [12]. VCTK is known to be a great, high quality dataset for text-to-speech or voice conversion. Nevertheless, it lacks of expressive samples to enable a good diffusion of emotions through the pipeline. ESD (Emotional Speech Dataset), that is designed for emotional speech synthesis, could have been tested in this study via a pretrained provided model but the MSP-Podcast dataset [13] was preferred, as it covers many more speakers and, by design, a wider range of conveyed emotions. This choice has been motivated by the fact that the preservation of expressiveness is the main weakness of the usual anonymization systems. However, it comes with a counterpart, as MSP-Podcast was not designed for speech synthesis and includes samples of lower quality, such as narrow band phone captured signals. It is also unbalanced in speakers, what could imply that under represented speakers are poorly described.

To improve the diffusion of expressiveness, we also propose to take benefit from emotion embeddings produced by a system proposed in [14] and available online[2] which was trained on MSP-podcast to predict the three continuous components of emotions, namely the arousal, valence and dominance. In the remainder of this article, the so-called *emotion embeddings* directly refer to these three values, while the model is denoted as AUD.

We believe that the use of a cleaned, more balanced, version of MSP-Podcast could improve the synthesis quality, but this has been left for future work. Combining expressiveness and audio quality could also be tackled by using the AUD model onto high quality data to select and annotate expressive content.

3.2. Submitted systems

3.2.1. M0 - Original DISSC

First of all, we evaluated how the original DISSC framework, used with its VCTK preset and pretrained models, deals with anonymization when z_{spk} is randomly chosen among VCTK speakers. This model is tagged as M0. Note that in order to use the provided pitch predictor of DISSC with Librispeech (which contains audios with greater durations), the inference script has been modified to process the audio inputs in chunks.

3.2.2. M1 - Prosody preservation

As expressiveness is known to be partially conveyed by pitch, the M1 model is based on a full preservation of the prosody by bypassing the pitch and rhythm predictors in order to estimate the best achievable performance for UAR with initial conditions, i.e. with a model trained on VCTK. The preserved prosody is extracted with pyAAPT as proposed in DISSC.

3.2.3. M2 - A dataset suited for expressiveness

As VCTK is known to be quite monotonous, MSP-Podcast is used to retrain the audio generator G. By still preserving the original prosody, this M2 model is expected to reach the best performance in terms of expressiveness preservation.

[2] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

3.2.4. M3 - Add some randomness to prosody

Inspired by [6], this M3 model is a M2 model in which prosody is slightly altered by multiplying each pitch value by a factor randomly picked in [0.8, 1.2]. This pitch degradation should confirm the importance of fundamental frequency both in expressiveness with a reduction of the UAR, and in anonymization with an increase of the EER.

3.2.5. M4 - Prosody prediction with MSP-Podcast

In M4 model, prosody DISSC predictors are switched on after being retrained on MSP-Podcast. An improvement is expected on anonymization metrics. According to DISSC framework, these predictors will modify both fundamental frequency and phonemes duration so a degradation of intelligibility is also expected.

3.2.6. M5 - Emotion embeddings

In M5 model, 3-dimensions emotion embeddings extracted thanks to the pretrained model AUD are extracted from the source signal and concatenated to z_{spk} in the original scheme. Apart from this modification, M5 is identical to M4. We expect that these features are used by the generator to reconstruct the signal so that this attribute can be controlled in the generation. During the inference process, while z_{spk} is modified to anonymize the speaker, the emotion embeddings are kept unchanged so as to preserve the original expressiveness.

3.3. Results

The results of these experiments are compiled in Table 1. Most of the assumptions are validated.

First of all, as a new kind of baseline, M0, the original implementation of DISSC, achieves quite good EERs with a poor intelligibility around 8% WER and a low 32% UAR on IEMOCAP.

The effect of preserving prosody by bypassing pitch and phonemes duration predictors are shown by model M1: compared to M0, WER g*ains 4 points and is back around 4%, UAR is increased by 4 points and EERs are altered by more than 10 points. These results both confirm the relation between prosody, intelligibility and expressiveness, so as the amount of speaker information conveyed by prosody through an anonymization pipeline. Notably, it emphasizes the tricky nature of anonymizing a voice while preserving its expressiveness : some aspects of the prosody are speaker related and must be hidden, while some others are closely connected to the conveyed expression and must be retained. The capabilities of a system therefore relies on its ability to disentangle these two parts of the prosody.

Comparing M2 to M1 and M4 to M0 provides some indications about the interest of using a more expressive dataset for training, as these two pairs of systems only differ by the training set (VCTK vs. MSP-Podcast). The results show improvements in UAR when more expressive training data (MSP-Podcast) are used : M2 achieves up to 6 points higher than M1, while M4 exceeds M0 by 4 points. With 44.53% UAR on IEMOCAP_dev and 42.09% UAR on IEMOCAP_test, M2 reaches the best performances in our experiments and is among the best baselines (except for McAdams-based B2 which stays far beyond).

With M3 model, we expected to raise a first proof of the need to disturb prosody to boost anonymization. Indeed, EERs are increased by one to five points just by slightly blurring fundamental frequency. Once again, in complement to what was observed with M1, the importance of prosody in speaker characterization is demonstrated. Unfortunately, the increased robustness of anonymization comes with a drop of the intelligibility and emotion detection, which are both altered by approximately 1.5 points.

The randomized prosody involved in M3 induces a slight modification of the original prosody while the predicted prosody involved in M4 leads to a stronger modification. So, from the analysis of the results obtained for M2 (prosody unchanged), M3 and M4, we deduce that the more prosody is altered, the greater the WER and the lower the UAR.

Finally, M5 model, when compared to M4, shows an increase in expressiveness preservation, which demonstrates that predictors conditioned on emotion embeddings allow to partly resynthesize the original expressiveness, even if the prosody is not the original one.

The EER results presented in Table 1 show significant differences between male and female test sets of librispeech. This includes the EER measured on unconverted signals, ie. the classical speaker identification, for which the difference is the most important. This indicates a possible bias in the data that may impact, at least partially, the conclusions drawn here. Moreover, since the EERs of anonymization systems rely on speaker verification models trained on anonymized data [1], their performance can be slightly modified from one run to another, due to randomness. Although a complete computation of confidence intervals could not be managed, it is worth mentioning a drift up to 1-2% for the highest EER, although it does not question the trends described hereabove.

4. Conclusions

This paper presents several anonymization systems based on the DISSC voice conversion scheme and used for the voice privacy challenge 2024. A particular attention has been paid to expressiveness preservation, which is a usual weakness of the current systems. For this purpose, expressive data have been used for training through the MSP-Podcast dataset, and emotional embeddings have been added as an explicit control attribute of the speech generation. Evaluations show that such strategies contribute to a better preservation of expressiveness, with a moderate impact on anonymization capabilities. Nevertheless, with a decrease of about 25 points of the emotion recognition after anonymization, there is still a long way to go before reaching this goal. Indeed, this work also underlines the importance of prosody in both speaker identification (to be remove) and expressiveness (to be preserved) so that the key point of an optimal anonymization system relies on its ability to disentangle these two aspects. Further investigations should be conducted to better control the attributes responsible for the expressiveness and to isolate them from the speaker related features. It includes expressive and high quality data selection from training and new strategies to disentangle expressiveness from speakers characteristics.

5. References

[1] 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," 2024.

[2] N. Gengembre, O. Le Blouch, and C. Gendrot, "Disentangling prosody and timbre embeddings via voice conversion," in *Interspeech 2024*. International Speech Communication Association, 2024.

Table 1: *VPC Results of the 6 proposed systems in terms of WER%, UAR% and EER%. Bottom rows show original, i.e. audio signals without anonymization, as so as baselines published results tagged from B1 to B6. Second column "Trained on dataset" mentions the dataset used for generator and predictors training; "Prosody" column describes how the corresponding system manages prosody and "Emotion embeddings" column indicates that only M5 system uses emotion embeddings.*

	Trained on dataset	Prosody	Emotion embeddings	**WER - libri** dev	test	**UAR - IEMOCAP** dev	test	**EER - libri** dev f	dev m	test f	test m
M0	VCTK	predicted	✗	8.47	8.16	32.30	32.17	39.91	35.40	28.83	29.12
M1	VCTK	preserved	✗	4.17	3.86	37.86	36.98	25.00	16.93	12.96	16.26
M2	MSP	preserved	✗	4.20	3.91	44.53	42.09	24.55	17.24	15.15	15.14
M3	MSP	randomized	✗	5.85	5.51	42.51	40.45	28.27	21.42	18.27	20.71
M4	MSP	predicted	✗	7.78	7.07	36.03	36.30	40.34	**35.86**	28.87	28.29
M5	MSP	predicted	✓	7.80	7.11	41.19	40.79	**40.49**	34.97	28.29	30.01
Orig.	-	-	-	**1.8**	**1.85**	**69.08**	**71.06**	10.51	0.93	8.76	0.42
B1	-	-	-	3.07	2.91	42.71	42.78	10.94	7.45	7.47	4.68
B2	-	-	-	10.44	9.95	55.61	53.49	12.91	2.05	7.48	1.56
B3	-	-	-	4.29	4.35	38.09	37.57	28.43	22.04	27.92	26.72
B4	-	-	-	6.15	5.90	41.97	42.78	34.37	31.06	29.37	31.16
B5	-	-	-	4.73	4.37	38.08	38.17	35.82	32.92	**33.95**	**34.73**
B6	-	-	-	9.69	9.09	36.39	36.13	25.14	20.96	21.15	21.14

[3] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the mcadams coefficient," *arXiv preprint arXiv:2011.01130*, 2020.

[4] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *arXiv preprint arXiv:1905.13561*, 2019.

[5] 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.

[6] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody Is Not Identity: A Speaker Anonymization Approach Using Prosody Cloning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/10096607

[7] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4725–4729.

[8] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," *arXiv preprint arXiv:2308.04455*, 2023.

[9] G. Maimon and Y. Adi, "Speaking style conversion in the waveform domain using discrete self-supervised units," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8048–8061. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.541

[10] 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.

[11] 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.

[12] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 920–924.

[13] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[14] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.