Deep voice conversion and video generation

Dario Avalle
EURECOM
dario.avalle@eurecom.fr

Marco Sorbi
EURECOM
marco.sorbi@eurecom.fr

Gabriele Spina
EURECOM
gabriele.spina@eurecom.fr

***Abstract*—Voice conversion (VC) is a method to transform recordings of one person's voice (source) into another person's voice (target). It has various applications, including dubbing for movies, modifying singing voices, forensics analysis, voice cloning, therapy, and rehabilitation. While VC technology can raise concerns about spoofing voice biometric systems, it can also be used to anonymize and protect individual privacy. This project focuses on modifying a speaker anonymization system, from the Voice Privacy Challenge (VPC) [1], into a voice conversion one. The proposed system utilizes DNN-based encodings for speaker and speech context attributes and the F0 (pitch) of the audio, reconstructing its waveform by means of a HiFi-GAN.**

The results of the voice conversion are then used to assess the vulnerability of automatic speaker verification (ASV) systems, which we proved to be vulnerable to attacks using voice conversion, and also as an integration for deepfake video generation. This last part was achieved in collaboration with the STARS team at INRIA Méditerranée in Sophia Antipolis.

I. Introduction

In recent years, the protection of personal data has become a very important topic, both in the legislature [2][3][4] and in the field of AI. According to the GDPR definition, personal data is defined as "any information relating to an identified or identifiable natural person" [5]. A very important personal data is the voice, but this data has the particularity of being characterized by several factors that are difficult to model: timbre, accent, cadence, mood, and others.

This issue led in 2020 to the birth of the Voice Privacy Challenge [1], from which several algorithms were born, including voice conversion. These algorithms aim to "change the voice" of an audio, i.e. given a certain audio pronounced by a certain speaker (called source), obtain another audio, containing the same speech, but pronounced by a second speaker (called target). These algorithms can be used both to improve privacy (since the source is anonymized) and for speaker identification attacks. For example, they can be used to worsen the performance of Automatic Speaker Verification (ASV) algorithms: the aim is to identify a speaker by checking the similarity between audio and his real voice.

Starting from the architecture proposed by Xiaoxiao Miao et al. in [6] for speaker anonymization, this project aims to modify it to address the voice conversion task, dividing it into two phases: extraction and generation.

II. Dataset

The datasets used for the task are those available in the VPC 2020 [1]. In particular, the models in [6] were trained on the Training set of VPC, while our experiments on ASV assessment were conducted using the Development and Evaluation sets, which contain portions of *LibriSpeech* [7] and *VCTK* [8] corpora. Moreover, for the task of deepfake generation and for the subjective evaluation of the results, *VoxCeleb-2* [9] was employed. Table I shows the composition of the datasets available in VPC 2020.

TABLE I: Number of speakers and utterances in the VoicePrivacy 2020 training, development, and evaluation sets.

Subset			**Female**	**Male**	**Total**	**#Utter.**
Training	VoxCeleb-1,2		2 912	4 451	7 363	1 281 762
	LibriSpeech train-clean-100		125	126	251	28 539
	LibriSpeech train-other-500		564	602	1 166	148 688
	LibriTTS train-clean-100		123	124	247	33 236
	LibriTTS train-other-500		560	600	1 160	205 044
Development	LibriSpeech dev-clean	Enrollment	15	14	29	343
		Trial	20	20	40	1 978
	VCTK-dev	Enrollment	15	15	30	600
		Trial (common)				695
		Trial (different)				10 677
Evaluation	LibriSpeech test-clean	Enrollment	16	13	29	438
		Trial	20	20	40	1 496
	VCTK-test	Enrollment	15	15	30	600
		Trial (common)				700
		Trial (different)				10 748

III. Models

The proposed model consists of two modules (Fig. 1): a feature extractor and a generator. The feature extractor extracts the feature F0, context features, and the speaker vector, while the generator uses these features to reconstruct the audio.

Feature extractor

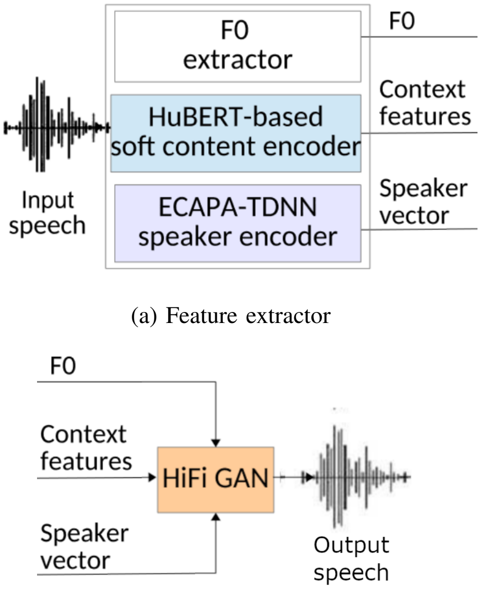
Given an utterance of size $L \times 1$, the feature extractor extracts:

- F0: it is extracted using the YAAPT algorithm [10] and consists of a vector $L \times 1$ whose i-th element represents the fundamental frequency of the speaker at the time instant i. Consequently, it is a sparse vector, since in some instants of time the speaker does not emit vocal sounds.
- Context features: they are extracted by a HuBERT-based soft content encoder as in [6] and are a matrix of size $L/320 \times 200$. The features extracted from the pre-trained HuBERT model [11] are representative of the speech, but as demonstrated in [12], they contain several personal information about the speaker. Consequently, soft content encoding is performed, which keeps more of the content information and increases intelligibility [12].
- Speaker vector: it is extracted using a pre-trained ECAPA-TDNN speaker encoder [13] and it is a vector of size 1×192. This model was developed for speaker verification and has the task of extracting a representative

speaker vector, independent of context. Ideally, it associates a vector uniquely with a given speaker, regardless of utterance.

Generator

The generator is a pre-trained HiFi-GAN [14], a generative adversarial network that has been proven to generate high-fidelity speech synthesis. The two key considerations to achieve this are that a phoneme can be longer than 100 ms and that audio consists of sinusoidal signals with different periods. Consequently, in order to obtain the most faithful audio possible, Kong et al. [14] implemented two discriminators: a multi-scale discriminator, to capture consecutive patterns and long-term dependencies, and a multi-period discriminator (MPD), itself composed of sub-discriminators each handling a portion of periodic signals of input audio.



(a) Feature extractor

(b) Speech generator

Fig. 1: Proposed model

IV. Experiments

The audio conversion has been made in two steps:

- speaker vector: taking inspiration from [6], our first approach to this task was to extract features from the source audio, substitute the speaker vector with the one of the target, and combine them to generate the desired audio. The obtained output was subjectively not very good, intuitively due to a frequency mismatch between the target and the output, but we used this as base of the next step.
- F0: to fix the frequency mismatch, additionally to substitute the speaker vector, we adapted the source's F0 to the target characteristics, in particular, mean and standard deviation in logarithmic scale [15]. Namely,

$$\mu_{audio} = \text{mean}\left(\log\left(f_{0,audio}\right)\right) \quad (1)$$

$$\sigma_{audio} = \text{std}\left(\log\left(f_{0,audio}\right)\right) \quad (2)$$

$$f_{0,new} = \exp\left(\mu_t + \frac{\sigma_t}{\sigma_s}\left(\log\left(f_{0,s}\right) - \mu_s\right)\right) \quad (3)$$

with s source and t target. The 0-values of the F0 were ignored in this operation, as they represent segments of audio where the speaker does not emit vocal sounds.

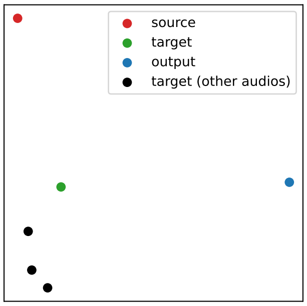


Fig. 2: PCA of speaker vectors of source, target ad output

To evaluate the outputs, we used two approaches

- subjective evaluation: this consists of various tests, spacing from the human listening to drawing plots of some characteristic of target and output
- objective evaluation: an Automatic Speaker Verification system (ASV) was used to evaluate the results on the *LibriSpeech* and *VCTK* datasets using the attack models described in [1], by testing various attacker-victim pairs using the attacker as source and the victim as target. In this case, as we had multiple audios for the same victim, we averaged the speaker vectors and F0s of all the audios of the victim in the *dev* portion of the dataset.

V. Results

Subjectively, listening to the outputs, they seem to be generally closer to the target speaker, which is a sign that the procedure worked well. However, we were able to detect a slightly higher amount of noise and, sometimes, an unnatural pitch. This, together with some inconsistencies in vocal metric patterns (i.e. rhythm, intonation, emphasis) and the lack of transfer of emotional expressiveness, contributed to making some of the generated examples less natural than both the source and the target audios. Some of these problems were expected, as we didn't modify the context features, and there is for sure a margin for improvement and future work to be done in this direction. Fig. 2 is a visualization of the speaker vectors of source, target, and output, it is clear that these results can be improved.

Objectively, we compared the ASV results on the output-target pairs with the ones on the original source-target pairs, using two metrics.

- Equal Error Rate (EER): this equals the false alarm and miss rates of the classifier, for the particular threshold at

TABLE II: EERs in percentage

Trial subset	Gender	Original source	Generated output
LibriSpeech dev-clean	Male	0.313	13.50
	Female	2.840	19.20
LibriSpeech test-clean	Male	0.888	12.40
	Female	0.545	15.40
VCTK-dev (different)	Male	0.154	11.80
	Female	0.507	9.94
VCTK-test (different)	Male	0.113	9.87
	Female	1.170	12.00

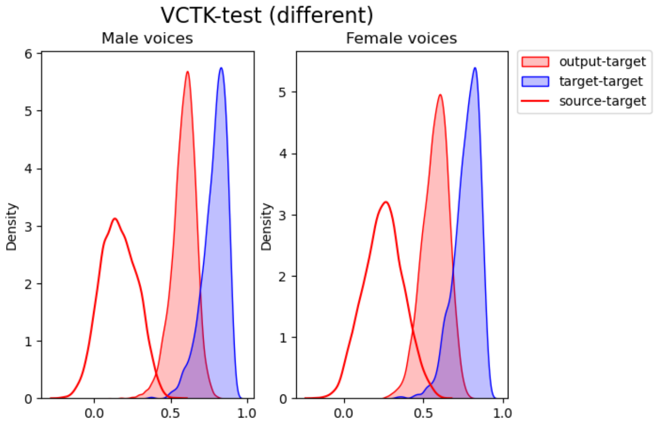


Fig. 3: Results of the ASV performed on VCTK-test

which these two values are the same. Table II shows the EERs for each partition of the datasets.

- Probability distributions: the distributions of the matching score between source-target, output-target, and target-target, for the *VCTK-test* dataset, are plotted in Fig. 3, the distribution of output-target scores gets significantly closer to the target-target one, with respect to the source-target one, which is good because it means that the ASV is more likely to confuse output and target.

We also tried to iterate the conversion process, by substituting the source with the output of the previous iteration, but we found out that this does not converge to the target, and generally it does not improve after the first iteration. A visualization is reported in Fig. 4.

VI. Conclusions

In this project, we started from a voice anonymization pipeline and we focused on modifying the architecture to perform voice conversion, targeting a specific speaker. We then assessed the robustness of an ASV system and we evaluated the results subjectively.

Regarding the ASV assessment with our pipeline, we notice a good improvement from the original source-target scores, both with the EER and with the probability distributions. Anyway, there is still a margin for improvement, considering that we got an EER between 10% and 20%, depending on the dataset, and that the optimal value is 50%. We believe this margin is caused by the output speaker vector not being too close to the target one.

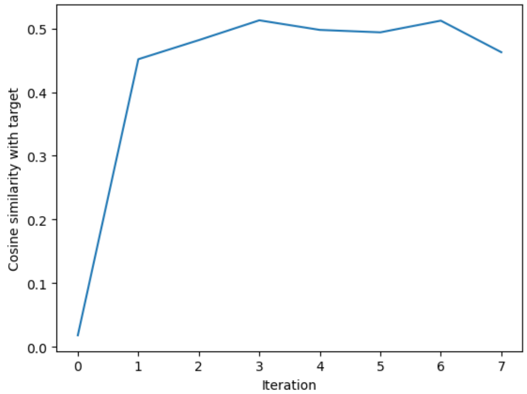


Fig. 4: Results of the iterated conversion

Listening to some generated examples, instead, we were able to always recognize a shift from the voice of the source to the one of the target. This shift, however, is not total and there is still a noticeable mismatch between the expected and the actual results. We believe that the context features extracted by the soft content encoder implicitly contain some information about the pitch and the speaker's identity. Our manipulation only of the F0 and the speaker vector could be not enough to achieve a complete voice conversion, and further research needs to be done in this direction.

VII. Future Works

To make the output speaker vector closer to the target one, we thought about retraining the HiFi-GAN modifying the generator loss, adding a cosine similarity loss term that penalizes their difference. We also thought, alternatively, about adding a module that modifies the speaker vector before feeding it to the HiFi-GAN, but we would prefer the first approach because the HiFi-GAN uses the context features which still contain speaker identity information [12]. We believe that this approach could help in the task of fooling an ASV, while it would be interesting to understand if such methodology could also improve the results of the subjective evaluation.

References

[1] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy initiative," in *Interspeech 2020*, ISCA, oct 2020.

[2] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council."

[3] "Ethics guidelines for trustworthy ai." https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[4] "AI Act." https://artificialintelligenceact.eu/.

[5] "Art. 4 GDPR." https://gdpr-info.eu/art-4-gdpr/.

[6] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Language-independent speaker anonymization approach using self-supervised pre-trained models," 2022.

[7] 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)*, pp. 5206–5210, 2015.

[8] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)." https://datashare.is.ed.ac.uk/handle/10283/3443, 2019.

[9] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech 2017*, ISCA, aug 2017.

[10] K. Kasi and S. A. Zahorian, “Yet another algorithm for pitch tracking,” in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–361–I–364, 2002.

[11] 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,” 2021.

[12] B. van Niekerk, M.-A. Carbonneau, J. Zaidi, M. Baas, H. Seute, and H. Kamper, “A comparison of discrete and soft speech units for improved voice conversion,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, may 2022.

[13] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Interspeech 2020*, ISCA, oct 2020.

[14] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” 2020.

[15] P. Champion, D. Jouvet, and A. Larcher, “A study of f0 modification for x-vector based speech pseudonymization across gender,” 2021.