Deep Learning-based F0 Synthesis for Speaker Anonymization

Ünal Ege Gaznepoglu, Nils Peters
International Audio Laboratories, Friedrich-Alexander-Universität, Am Wolfsmantel 33, 91058 Erlangen, Germany
{ege.gaznepoglu, nils.peters}@audiolabs-erlangen.de

***Abstract*—Voice conversion for speaker anonymization is an emerging concept for privacy protection. In a deep learning setting, this is achieved by extracting multiple features from speech, altering the speaker identity, and waveform synthesis. However, many existing systems do not modify fundamental frequency (F0) trajectories, which convey prosody information and can reveal speaker identity. Moreover, mismatch between F0 and other features can degrade speech quality and intelligibility. In this paper, we formally introduce a method that synthesizes F0 trajectories from other speech features and evaluate its reconstructional capabilities. Then we test our approach within a speaker anonymization framework, comparing it to a baseline and a state-of-the-art F0 modification that utilizes speaker information. The results show that our method improves both speaker anonymity, measured by the equal error rate, and utility, measured by the word error rate.**

***Index Terms*—speaker anonymization, X-vector, bottleneck features (BNs), F0, deep neural network (DNN)**

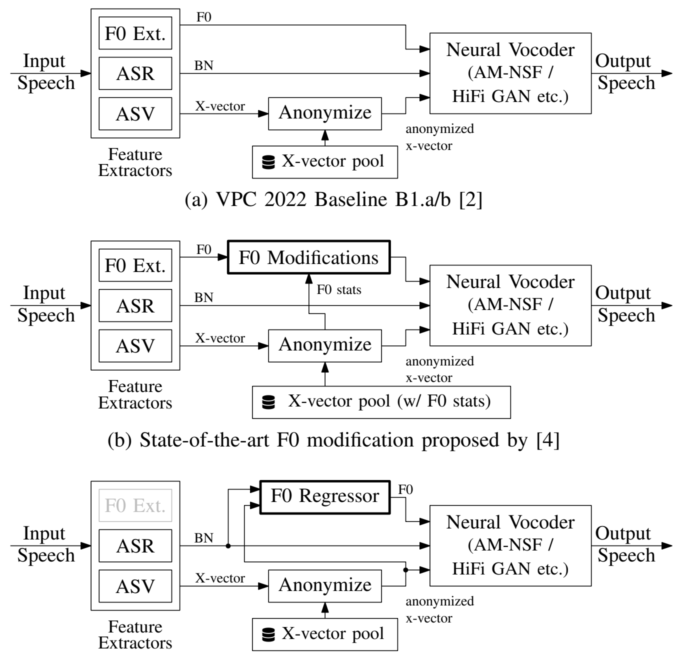
I. INTRODUCTION

Speaker anonymization is a feasible solution to prevent personal information leakage during cloud-enabled speech processing tasks, such as voice assistant usage [1]. The VoicePrivacy Initiative organizes a VoicePrivacy Challenge (VPC) to facilitate further studies [2]. Many researchers adopt their datasets, baselines and evaluation methodology [3]–[6].

The majority of the studies build upon a system based on X-vectors and neural waveform models [7], citing its superior anonymization performance as well as lower intelligibility degradation, in comparison to the DSP-based anonymization systems [8]. However, an important caveat of [7] is that the fundamental frequency (F0) is not altered before synthesis (see Fig. 1a), exposing the original F0, and introducing synthesis artifacts due to the incompatibility between F0 and the anonymized X-vector. Subsequent works [4], [5], [9] investigated different F0 manipulations, however, a joint consideration of the extracted features has not been investigated.

In our VPC 2022 contribution [10], we prototyped a novel approach to address the aforementioned issues simultaneously. Our system, attaining the best naturalness scores in a subjective listening test, was found to be an effective contender [11], but an evaluation beyond the VPC paradigm remained out of scope. In this work, we formally test our approach by evaluating its reconstructional capabilities and comparing it with a state-of-the-art F0 modification for anonymization.

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS.



(a) VPC 2022 Baseline B1.a/b [2]

(b) State-of-the-art F0 modification proposed by [4]

(c) Our F0 synthesis approach

Fig. 1: The three speaker anonymization systems under test.

II. RELATED WORKS

A. Speaker anonymization techniques

1) The VPC 2022 B1.b Baseline: The baseline, which our contribution builds upon, is depicted in Fig. 1a [7]. It consists of three feature extractors, an anonymization block and a neural vocoder. Extractors decompose speech into individual components, namely F0, bottleneck features (BNs), and X-vectors. Table I summarizes their purposes and extraction details. The anonymization block changes the X-vector, and the neural vocoder (B1.b uses a neural source-filter (NSF) trained with HiFi-GAN discriminators) synthesizes a speech signal from the new set of features. The anonymization block works in the following way:

(i) In the X-vector pool, find a set $\mathcal{P}$ of N X-vectors of a certain gender (same as original or opposite) that are the furthest away (per PLDA) from the original X-vector.
(ii) Select K random vectors from $\mathcal{P}$ and average them.

2) Proposed F0 modifications: Preserving the unmodified F0s has been identified as a shortcoming, and several improvements were proposed. For instance, a study compared various

TABLE I: Extracted features per utterance. W: window size, H: hop size, N: number of frames of an utterance

Feature (purpose)	Extractor	Properties
F0 (Prosody)	YAAPT [12]	size: (Nx1), W: 35 ms, H: 10 ms
BN (Verbal content)	TDNN-F [2]	size: (Nx256), H: 10 ms
X-vector (Identity)	TDNN [2]	size: (1x512)

DSP-based modifications [5]. Authors of [6] experimented with speaking rate change and variable F0 shifting. Notably, Champion et al. [4] proposed creating an F0 statistics dictionary (mean, std) for each speaker in the pool, then assumed the same K-subset averages of (mean, std) as the anonymized X-vector statistics. This approach improved the synthesized audio quality, especially for cross-gender anonymization. However, it needs an F0 dictionary, so it is incompatible with many anonymizers such as [3] that have a statistical model instead of a speaker pool. We consider [4], the only system in the literature that focuses on performing feature-aware F0 modification and reports VPC metrics, as the state-of-the-art, and integrated it into our comparison framework (see Fig. 1b).

B. Deep learning-based F0 estimation

In recent years, data-driven F0 extractors were developed that outperform the statistics-based ones. CREPE [13] and FCN [14] perform a binned classification to estimate F0. Tran et al. [15] uses a joint classification (voiced-unvoiced decision) and regression (F0 values) formulation. Mentioned systems use convolutional architectures to process the waveforms and are trained on perfectly annotated data. Typical detection metrics such as accuracy, precision and recall are used to evaluate the voiced-unvoiced decision. Fine and gross pitch errors (FPE, GPE [16]) are popular metrics to evaluate the F0 estimation quality. The definitions that we utilized are given below:

$$\text{GPE: } \frac{\text{num. of frames whose error} > 20\%}{\text{num. of correctly identified voiced frames}} \tag{1}$$

$$\text{FPE: } \frac{\text{num. of frames whose error} > 5\%}{\text{num. of frames whose error} < 20\%} \tag{2}$$

C. Speaker anonymization datasets and evaluation

In our work, we adopt the VPC framework, consisting of datasets, attack models, and metrics and systems for evaluation. The datasets consist of LibriSpeech and VCTK subsets [2]. Table II outlines our dataset utilization and, which conforms to the VPC guidelines. Further information is available in the VPC 2020 [17] and 2022 [2] evaluation plans.

Anonymization performance is measured using automated speaker verification-equal error rate (ASV-EER). Higher EERs correspond to better anonymization as the synthesized speech is less linkable to the original speaker. Some systems introduced in 2020 achieved sufficient anonymization (50% EER) upon comparison to the original speech via a pretrained ASV evaluator [8]. Thus in 2022, a stronger attack model is introduced, where the attackers are able to train ASV systems using anonymized data, i.e., a semi-informed attack model [2]. In our work we use the latter type of EER computation.

TABLE II: Dataset compositions. #F, #M: number of female/male speakers. #Utt: number of utterances

Used for	Subset Name	#F	#M	#Utt	#Frames
training	libri-dev-trials	20	20	1978	1411330
	vctk-dev-trials	15	15	11372	3792243
	libritts-train-clean-100	123	124	33236	19297310
validation	libri-dev-enrolls	14	15	343	227416
	vctk-dev-enrolls	15	15	600	192510
testing	libri-test-{enrolls,trials}	15	15	1934	1563092
	vctk-test-{enrolls,trials}	15	15	12048	3846010

Moreover, the word error rate (WER) in an automated speech recognition (ASR) scenario is measured. Lower WERs are desired, meaning the anonymizer did not compromise the utility. The 2022 challenge trains the ASR evaluation system using anonymized data, however the training process is cumbersome, limiting the breadth of our evaluation. Thus we used the 2020 version of ASR evaluation which uses a pretrained model. In 2022, two auxiliary metrics are introduced: pitch correlation (ρ^{F_0}) and gain of voice distinctiveness (G_{VD}) [18]. To ensure that the emotions and the speaking pace is largely preserved, $\rho^{F_0} > 0.3$ is required. A gain above or below 0 correspond to an increase or decrease in voice distinctiveness, and $G_{\text{VD}} = 0$ would be the optimal value, attained when the voice distinctiveness is preserved [2].

Some works in the speaker anonymization domain utilize the so-called contrastive systems, that feature minor deviations from the proposed idea, to assess the relative contributions of individual design choices. For instance, in [19], speech files with different feature combinations were synthesized and evaluated, to gain insights on how the acoustic model and the waveform model contribute to the anonymization.

III. METHODOLOGY

Rather than modifying extracted F0 trajectories as in previous works, we propose a synthesis by regression approach (see Fig. 1c). This approach eliminates the concern of leaking original F0 to the output signal and can potentially improve the quality of the subsequent neural waveform synthesis as the F0 trajectory would be coherent with the anonymized X-vector.

A. The framewise F0 synthesizer

We propose a 4-hidden-layer DNN (see Fig. 2) to framewise infer F0 from the X-vector and BNs. Similar to [15], it yields two outputs: $\hat{F}_0[n]$, the predicted log-F0, and $g[n]$, the logits of a frame being voiced. Then, a mask is constructed, zeroing $\hat{F}_0[n]$ if the frame is classified as unvoiced (i.e., $g < 0$).

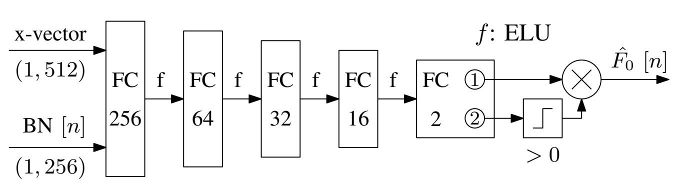


Fig. 2: Proposed architecture. 'FC' is a fully connected layer, with number of neurons indicated below. The output layer neurons are denoted with ①, ②.

The model is trained on a composite loss function as in [15].

$$\mathcal{L}(F_0, \hat{F_0}, g, v) = \mathrm{L1}(F_0, \hat{F_0}) + \alpha \mathrm{BCE}(g, v) \quad (3)$$

A factor α balances the regression and classification tasks. Here, $v = 1$ denotes a voiced frame whereas $v = 0$ is an unvoiced one. The functions $\mathrm{L1}(\cdot)$ and $\mathrm{BCE}(\cdot)$ denote the L1 loss and binary cross-entropy with logits. Different to [15], we do not have access to the perfect F0 annotations, so we assume YAAPT extractions as the ground truth. To diminish the effect of imperfect labels, we use L1 loss instead of MSE.

TABLE III: Hyperparameter search intervals and adopted values. All parameters are searched in the log space.

Search Space	Purpose	Adopted Value
$10^{-3} < \alpha < 500$	loss factor	28.112
$10^{-4} < \delta < 0.5$	dropout probability	0
N/A	learning rate	0.0003
N/A	batch size	262144

B. Training strategies and hyperparameter tuning

The model is implemented with PyTorch [20]. The training logic is provided by PyTorch Ignite [21]. Training utterances (see Table II) are concatenated into a single tall matrix and shuffled. This way, voiced and unvoiced frames from different utterances are present in each batch. We used Nesterov-Adam (NAdam) as our optimizer, with default parameters. Early stopping after 10 epochs, and learning rate reduction by a factor of 0.2 after 5 epochs of non-increasing validation metric are used to combat overfitting. We used the total percentage of accurately processed frames, i.e., correctly classified unvoiced frames and frames that do not have gross pitch errors as the validation metric. Table III outlines the hyperparameters and their tuning procedure with OpTuna [22]. The procedure yielded a $\delta < 0.001$, so we disabled dropout by setting $\delta = 0$.

C. Evaluation

Figures 3a and 3b compare the predictions by our model with the assumed ground truth, for a male and a female utterance from the validation set. Fig. 3c shows a cross-gender (female to male) F0 conversion. The predicted F0 (blue) has a mean that is significantly less than the original F0 (orange) while preserving the global trend, hinting that our model is aware of the effects of gender on F0. Now we utilize a two step evaluation procedure to assess the capabilities of our model.

1) Reconstruction of known F0s: To measure the similarity of the synthesized F0 values to the assumed ground truth, we compute GPE, FPE and voiced-unvoiced classification metrics (Accuracy, Precision, Recall).

2) Utilizing the model in a speaker anonymization system: We integrate our model into the Baseline B1.b and evaluate according to the VPC framework (see Section II-C). We also introduce contrastive systems in Table IV, which use altered neural vocoder inputs to gain insights. The anonymization block behaves the same across trials for a fair comparison.

TABLE IV: Contrastive systems and their feature descriptions.

ID	Input F0 to NSF	Input X-vector to NSF
Ours	synthesized with anonymized X-vector	anonymized
C1	synthesized with original X-vector	original
C2	synthesized with anonymized X-vector	original
C3	synthesized with original X-vector	anonymized

IV. Results and Discussion

A. Reconstruction error of known F0s

Table V shows the F0 reconstruction performance of our model. Despite its simplicity, in particular the narrow temporal scope, it attains a voiced-unvoiced decision test accuracy around 94%, similar to the YAAPT's performance reported in [16]. The visual inspection of the synthesized F0 suggests that most decision errors occur at the edges of voiced segments, instead of erratic switchings at random instances.

The regression performance, measured by GPE and FPE, indicates that our system is only able to provide a crude approximation of the ground truth. End-to-end systems that operate on the full waveform report better GPE and FPE (e.g., [15]). A histogram of the voiced frame prediction errors is provided in Fig. 4. They resemble a Gaussian distribution with a high variance, centered near zero. We conclude a single frame contains partial information to model the F0 trajectories, and we expect the reconstruction performance to improve if the model used the temporal context, e.g., via a recurrent or convolutional architecture. The lack of perfect F0 annotations also bounds the achievable performance, possibly fixable by a self-supervised training scheme such as SPICE [23].

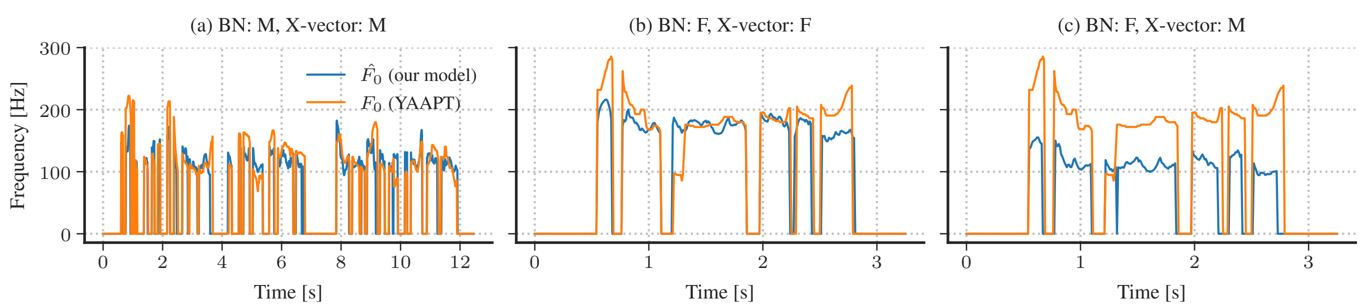


Fig. 3: Model behavior with samples from the validation set. Figures (a) and (b) illustrate reconstructions for those utterances and (c) shows synthesized F0 for the utterance said in (b), if it were said by the speaker in (a).

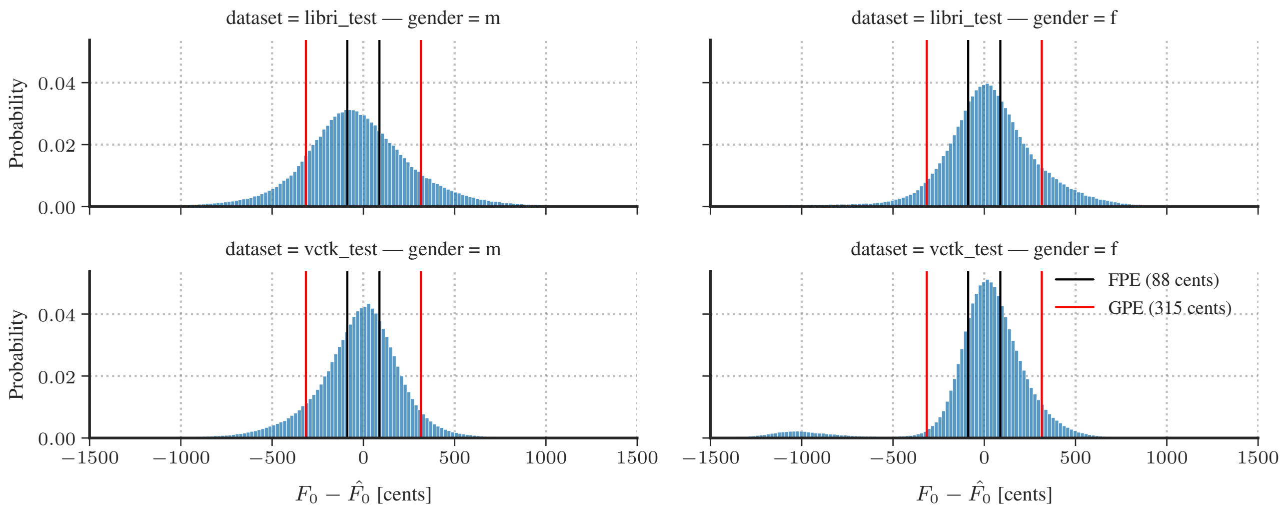


Fig. 4: Test set F0 prediction errors. Red and black bars denote the GPE (315 cents / 20%) and FPE (88 cents / 5%) thresholds.

TABLE V: Model F0 reconstruction performance, all reported in percentages. Acc: Accuracy, Prec: Precision, Rec: Recall

Dataset	**Sex**	**GPE(↓)**	**FPE(↓)**	**Acc.(↑)**	**Prec.(↑)**	**Rec.(↑)**
libri-test	F	31.6	66.9	93.0	94.6	93.3
	M	41.8	71.8	92.5	93.0	93.0
vctk-test	F	24.6	63.9	95.1	94.1	93.5
	M	38.8	69.9	94.6	93.5	92.5

B. Utilizing the synthesizer in a speaker anonymization system

Evaluation of the mentioned anonymizers is presented in Table VI. The new, stronger attack model caused some decrease in EER for the shift-and-scale approach [4] yet it still outperforms the baseline. The different vocoder resulted in better WERs compared to the original publication, however the conclusions are the same: cross-gender synthesis became more intelligible and same-gender synthesis is comparable to the baseline. Our methodology alters the F0 trajectory altogether, thus on average performs significantly better than the other systems in terms of both metrics, in same gender and cross-gender anonymization. Combined with the observations from the reconstruction performance, in particular the voiced-unvoiced decision differences between YAAPT and our method, we think that our system is possibly able to correct some of the mistakes made by YAAPT thanks to the BNs and this would explain the WER improvement.

Evaluation of the contrastive systems provide additional intuition on understanding how F0 modification helps. Usage of the original X-vectors together with the F0 modification (systems C1 and C2) do not yield any significant anonymization and cause an unexpected WER increase (+1%). We plan to further investigate the reasons for this increase. Supplying the synthesized F0 using the original speaker identity but using the anonymized X-vector for synthesis (system C3) yields an insignificant EER improvement and causes no WER change

TABLE VI: VPC framework results for the baseline (B1.b), our implementation of the state-of-the-art [4], our proposal, and contrastive systems (C1-C3). Cross gender conversion is only possible for systems utilizing the anonymized X-vector at least once. The weights for the average are introduced by the VPC 2022.

Dataset	Weight	Gender	EER [2] [%] (↑)						WER [17] [%] (↓)					
		(From → To)	B1.b	[4]	Ours	C1	C2	C3	B1.b	[4]	Ours	C1	C2	C3
libri-test	0.25	F → F	11.13	**12.96**	12.04	14.60	13.14	8.57	5.60	5.58	**5.48**	6.39	6.40	5.59
		M → M	7.35	8.69	**10.02**	1.34	1.34	8.46						
vctk-test-diff	0.20	F → F	12.04	13.73	**16.20**	5.45	8.39	12.81	14.66	14.76	**14.57**	16.28	16.10	14.66
		M → M	8.78	9.93	**10.79**	1.61	2.01	8.84						
vctk-test-com	0.05	F → F	11.56	15.03	**18.50**	2.31	4.04	15.32						
		M → M	9.04	12.71	**14.12**	1.41	0.84	11.30						
weighted average / same gender			9.81	11.53	**12.54**	5.58	5.94	9.92	10.13	10.17	**10.03**	11.34	11.25	10.13
libri-test	0.25	F → M	14.23	**23.18**	22.99	N/A	13.50	12.77	5.99	5.82	**5.66**	N/A	6.79	5.87
		M → F	8.46	15.81	**19.38**	N/A	1.34	9.8						
vctk-test-diff	0.20	F → M	16.67	26.75	**27.83**	N/A	5.14	17.80	15.37	14.98	**14.80**	N/A	17.04	15.13
		M → F	14.24	22.62	**29.97**	N/A	2.53	22.90						
vctk-test-com	0.05	F → M	21.39	36.99	**38.15**	N/A	4.36	26.88						
		M → F	12.99	27.97	**33.05**	N/A	1.70	18.36						
weighted average / cross gender			13.57	22.87	**25.71**	N/A	5.55	16.04	10.68	10.4	**10.23**	N/A	11.92	10.5

with respect to the baseline. Hence, it could conceivably be hypothesised that the performance increase yielded by our system is due to the learned characteristics of the speakers and not due to the artifacts our system introduces.

Not mentioned in Table VI, our system satisfies the VPC requirement $\rho_{F_0} > 0.3$ on all subsets. The G_{VD} values are comparable across the primary systems and indicate a common loss of voice distinctiveness. Previous studies have already shown that the anonymization block is the culprit, because it yields unnaturally similar anonymized X-vectors [3].

Besides the improved metrics, we observe an improved run time for F0 computation and thus speaker anonymization: For all datasets utilized except 'libritts-train-clean-100', it takes only two minutes to synthesize F0 values using our approach, whereas YAAPT extractions take 35 times longer.

V. Conclusion

In this work, we formally evaluated a DNN-based F0 synthesis approach for speaker anonymization. Notwithstanding the architectural simplicity and the lack of perfect F0 annotations for training, the proposed approach managed to improve the EER and WER metrics over the state-of-the-art speaker-dependent F0 modification in the literature. The evidence we present suggests that the F0 provided by our model is sufficient to generate intelligible and natural sounding utterances, when paired with the utilized neural vocoder. Our findings indicate it is worthwhile to perform a follow-up study, to improve the temporal behavior of F0, e.g., with a different architecture. Also, a self-supervised training scheme may tackle the issue of not having perfect F0 annotations for system training.

References

[1] N. Tomashenko *et al.*, "Introducing the VoicePrivacy initiative," in *Proc. Interspeech Conf.*, 2020.

[2] N. Tomashenko *et al.* "2nd VoicePrivacy challenge evaluation plan." (2022), [Online]. Available: https://arxiv.org/abs/2203.12468.

[3] H. Turner, G. Lovisotto, and I. Martinovic, "Speaker anonymization with distribution-preserving x-vector generation for the VoicePrivacy challenge 2020," in *VoicePrivacy Challenge Submission*, 2020.

[4] P. Champion, D. Jouvet, and A. Larcher, "A study of f0 modification for x-vector based speech pseudonymization across gender," in *2nd AAAI Workshop on Privacy-Preserving AI*, 2021.

[5] U. E. Gaznepoglu and N. Peters, "Exploring the importance of f0 trajectories for speaker anonymization using x-vectors and neural waveform models," in *Workshop on MLSLP*, 2021.

[6] C. O. Mawalim, K. Galajit, J. Karnjana, S. Kidani, and M. Unoki, "Speaker anonymization by modifying fundamental frequency and x-vector singular value," *Computer Speech & Language*, vol. 73, 2022.

[7] F. Fang *et al.*, "Speaker anonymization using x-vector and neural waveform models," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019.

[8] N. Tomashenko *et al.*, "The VoicePrivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, 2022.

[9] L. Tavi, T. Kinnunen, and R. González Hautamäki, "Improving speaker de-identification with functional data analysis of f0 trajectories," *Speech Communication*, vol. 140, 2022.

[10] U. E. Gaznepoglu, A. Leschanowsky, and N. Peters, "Voiceprivacy 2022 system description: Speaker anonymization with feature-matched F0 trajectories," in *VoicePrivacy Challenge Submission*, 2022.

[11] N. Tomashenko *et al.* "The VoicePrivacy 2022 challenge results." (2022), [Online]. Available: https://www.voiceprivacychallenge.org/results-2022/docs/VoicePrivacy_2022_Challenge___Natalia_Tomashenko.pdf.

[12] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal Acoust. Soc. of America*, vol. 123, no. 6, 2008.

[13] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[14] L. Ardaillon and A. Roebel, "Fully-convolutional network for pitch estimation of speech signals," in *Proc. Interspeech Conf.*, 2019.

[15] D. N. Tran, U. Batricevic, and K. Koishida, "Robust pitch regression with voiced/unvoiced classification in nonstationary noise environments," in *Proc. Interspeech Conf.*, 2020.

[16] R. Vaysse, C. Astésano, and J. Farinas, "Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech," *The Journal Acoust. Soc. of America*, vol. 152, no. 5, 2022.

[17] N. Tomashenko *et al.* "1st VoicePrivacy challenge evaluation plan." (2020), [Online]. Available: https://arxiv.org/abs/2205.07123.

[18] P.-G. Noé *et al.*, "Towards a unified assessment framework of speech pseudonymisation," *Computer Speech & Language*, vol. 72, 2022.

[19] P. Champion, D. Jouvet, and A. Larcher, "Speaker information modification in VoicePrivacy 2020 toolchain," in *VoicePrivacy Challenge Submission*, 2020.

[20] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. of 33rd Int. Conf. on Neural Inf. Proc. Sys. (NeurIPS)*, 2019.

[21] V. Fomin, J. Anmol, S. Desroziers, J. Kriss, and A. Tejani, *High-level library to help with training neural networks in PyTorch*, 2020.

[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. ACM SIGKDD*, 2019.

[23] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirović, "SPICE: Self-supervised pitch estimation," *IEEE/ACM TASLP*, vol. 28, 2020.