VoicePrivacy 2022 System Description: Speaker Anonymization with Feature-matched F0 Trajectories

Ünal Ege Gaznepoglu[1,2]*, Anna Leschanowsky*[1,2]*, Nils Peters*[1]

[1] Friedrich-Alexander-Universität, International Audio Laboratories Erlangen, Germany
[2] Fraunhofer IIS, Erlangen, Germany

{uenal.ege.gaznepoglu, anna.leschanowsky}@iis.fraunhofer.de,
nils.peters@audiolabs-erlangen.de

Abstract

We introduce a novel method to improve the performance of the VoicePrivacy Challenge 2022 baseline B1 variants. Among the known deficiencies of x-vector-based anonymization systems is the insufficient disentangling of the input features. In particular, the fundamental frequency (F0) trajectories, which are used for voice synthesis without any modifications. Especially in cross-gender conversion, this situation causes unnatural sounding voices, increases word error rates (WERs), and personal information leakage. Our submission overcomes this problem by synthesizing an F0 trajectory, which better harmonizes with the anonymized x-vector. We utilized a low-complexity deep neural network to estimate an appropriate F0 value per frame, using the linguistic content from the bottleneck features (BN) and the anonymized x-vector. Our approach results in a significantly improved anonymization system and increased naturalness of the synthesized voice. Consequently, our results suggest that F0 extraction is not required for voice anonymization.

Index Terms: neural networks, fundamental frequency (F0), bottleneck features (BN), x-vectors

1. Introduction

Introduction of the VoicePrivacy Challenge has stirred a multinational interest in design of voice anonymization systems. The introduced framework consists of baselines, evaluation metrics and attack models and has been utilized by researchers to improve voice anonymization. Figure 1 depicts baseline B1 (referred to as B1.a in the current edition) [1]. Previous submissions mostly focused on changes to the individual blocks of the baselines. However, regardless of the individual modifications to this baseline by different groups, the obtained audio recordings are considered 'unnatural' [2].

To improve anonymization performance as well as intelligibility, F0 modifications have been explored in the previous edition of the VoicePrivacy Challenge and subsequent works utilizing the challenge framework. Among the techniques investigated are creating a dictionary of F0 statistics (mean and variance) per identity and utilizing these for shifting and scaling the F0 trajectories [3], applying low-complexity DSP modifications [4] and applying functional principal component analysis (PCA) to get speaker-dependent parts [5]. Their results show that F0 trajectories contribute to anonymization and modifications are promising to improve the performance of the system.

Along the same lines, we hinted in a previous work that disentangling the features can increase the performance [4]. In particular, F0s are a complex combination of the identity of the speaker, the linguistic meaning, and the prosody, which also includes situational aspects such as emotions and speech rate [6]. Many speech synthesizers, notably the neural source-filters (NSFs), incorporate F0 trajectories as a parameter to control the initial excitation, mimicking the voice cords [7]. Thus, data-driven parts of the architectures have relatively little control over shaping the excitation. This motivated us to investigate ways to apply a correction to the F0 trajectories before the synthesis such that they match the BNs and x-vectors. Figure 1 shows how our proposal integrates into the baseline B1.

2. Our Contributions

2.1. A regression DNN for F0 trajectories

We utilized a 3-hidden-layer deep neural network (DNN) (see Fig. 2) to frame-wise predict F0 trajectories from the utterance level x-vectors and the BNs. Internals of the so-called fully connected (FC) layer is depicted in Figure 3. F0 trajectories are predicted in logarithmic scale with a global mean-variance normalization. Two output neurons in the last layer signify the predicted pitch value $\hat{F}_0[n]$ (no activation function) and the probability of the frame signifying a voiced sound $p_v[n]$ (sigmoid activation function). According to this probability, the F0 value for the frame is either passed as is (if the probability is greater than 0.5), or zeroed out (otherwise). The loss function $\mathcal{L}$ for a batched input is given in Equation 1 where 'MSE$(\cdot)$' and 'BCE$(\cdot)$' denote the 'mean-squared error' and 'binary cross entropy with logits' as implemented by PyTorch. The variable v denotes the voiced/unvoiced label of the frame and α denotes a trade-off parameter balancing the classification and regression tasks.

$$\mathcal{L}(F_0, \hat{F_0}) = \text{MSE}(F_0 - \hat{F_0})^2 + \alpha\text{BCE}(p_v, v), \qquad (1)$$

2.1.1. Training strategies and hyperparameter optimization

The DNN is implemented using PyTorch [8], and trained using PyTorch Ignite [9]. All files in the `libri-dev-*` and `vctk-dev-*` subsets are concatenated into a single tall matrix, then a random $(90\%, 10\%)$ train-validation split is performed, allowing frames from different utterances to be present in a single batch. We use early stopping after 10 epochs without improvement and learning rate reduction (multiplication by 0.1 after 5 epochs without improvement in validation loss).

OpTuna [10] tunes the learning rate `lr`, the trade-off parameter α and the dropout probability p. Optimal values obtained after 50 trials are listed in Table 1. We found the system to perform better without dropout, thus $p = 0$.

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

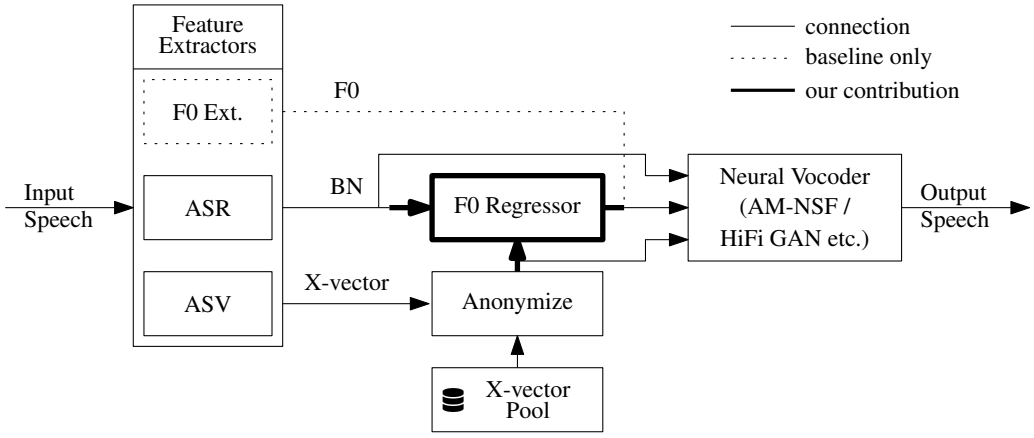

Figure 1: *Signal flow diagrams of the baselines B1.a (if neural vocoder is an AM-NSF), B1.b (if neural vocoder is NSF with GAN) and* `joint-hifigan` *(if neural vocoder is the original HiFi-GAN) together how our contribution "F0 regressor" is integrated.*

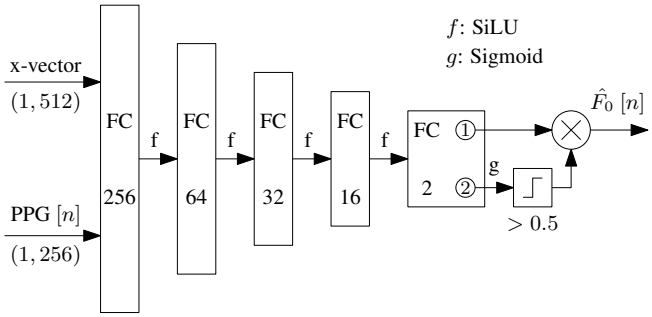

Figure 2: *(a) Architecture of our proposed neural network. The numbers below the expression "FC" denote the number of neurons in each layer. "FC" denotes a fully connected layer. The circles with numbers* $1, 2$ *in the last layer denote the output of* n*th neuron in that layer (after dropout if applicable).*

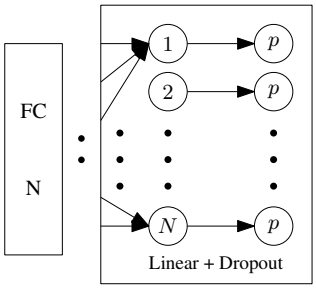

Figure 3: *Internals of a fully connected layer. It comprises of a linear layer followed by a dropout layer, where the dropout probability is* p*. The circles with numbers* $1, 2, \ldots, N$ *denote the number of the neuron in that layer.*

Parameter	Value
α	0.00022
`lr`	0.0007
p	0.0

Table 1: *Hyperparameter values obtained using OpTuna*

3. Evaluation

3.1. Analysis of the generated F0 trajectories

We verified the performance of our F0 regressor by visualizing the reconstructions for matched x-vectors and cross-gender x-vectors. The latter allows to evaluate the generalization capabilities. In Figure 4, the F0 estimates for unaltered target and source speakers (subplots 1 and 2) as well as a cross-gender F0 conversion is given (subplot 3) for the linguistic features from the female speaker and the x-vector from the male speaker. Resulting estimated F0 trajectory has a mean shift of roughly 60 Hz and correctly identifies voiced and unvoiced frames.

While we acknowledge the necessity of thorough objective and subjective evaluation of our methodology, due to the time and space constraints we believe it is better suited as part of a different publication.

3.2. Evaluation via challenge framework

We executed the evaluation scripts provided by the challenge organizers. The anonymization procedure for the evaluation system training utilized the same model we obtained by training on the development subsets. As our system does not include a tunable parameter that governs the trade-off between the equal error rate (EER) and WER, we submit a single set of results. As shown in Table 2, our system significantly outperforms the Baseline B1.b variant joint-hifigan in terms of EER. The loss in WER is negligible. Furthermore, our EER is also significantly better than any other baseline system (c.f. [12]). For every data subset the pitch correlation ρ^{F_0} resides in the accepted interval $[0.3, 1]$. The voice distinctiveness G_{VD} suffered some losses.

4. Conclusion

In this technical report we described our VoicePrivacy Challenge 2022 submission. Rather than extracting the F0 feature from the original speech, we proposed a novel low-complexity DNN-based F0 synthesis method which uses the linguistic content from the BNs and the anonymized x-vector as input features. The evaluation results indicated that our method mostly preserved the WER, the pitch correlation score ρ^{F_0}, some re-

Dataset	Sex	EER [%]			WER [%]			ρ^{F_0}			G_{VD} [dB]		
		B1.b	Submitted	Fixed	B1.b	Submitted	Fixed	B1.b	Submitted	Fixed	B1.b	Submitted	Fixed
libri-dev	F	16.62	23.86	**24.15**	**3.98**	4.12	4.13	0.84	0.82	0.81	-5.86	-6.87	-8.95
	M	5.44	16.15	**16.61**				0.73	0.68	0.65	-5.44	-5.27	-6.35
vctk-dev	F	7.08	19.71	**16.34**	**10.56**	10.36	10.62	0.85	0.86	0.85	-6.57	-6.90	-10.70
	M	6.55	17.42	**21.69**				0.74	0.69	0.70	-8.80	-9.27	-10.94
vctk-dev-com	F	8.43	16.57	**11.63**				0.83	0.83	0.82	-5.34	-4.84	-8.17
	M	9.69	17.38	**22.22**				0.70	0.62	0.59	-6.21	-6.51	-7.81
Ø (dev)		9.15	19.17	**19.49**	**7.27**	7.24	7.38	0.78	0.75	0.75	-6.48	-6.84	-8.95
libri-test	F	8.39	22.63	**22.99**	**4.28**	4.43	4.45	0.83	0.81	0.82	-5.58	-6.16	-7.14
	M	6.46	19.38	**21.83**				0.68	0.60	0.59	-5.52	-5.54	-5.68
vctk-test	F	9.00	22.99	**22.99**	**10.44**	10.32	10.52	0.86	0.85	0.85	-8.21	-8.87	-12.42
	M	8.15	17.51	**17.51**				0.75	0.70	0.70	-8.18	-8.81	-10.42
vctk-test-com	F	11.56	19.65	**19.65**				0.84	0.82	0.82	-6.68	-7.34	-10.66
	M	7.63	12.99	**12.99**				0.69	0.63	0.64	-6.11	-6.14	-7.43
Ø (test)		8.10	20.24	**22.07**	**7.36**	7.38	7.49	0.78	0.74	0.74	-6.69	-7.14	-8.68

Table 2: *Results from Baseline B1.b variant* `joint-hifigan` *taken from [13] compared with the variant including our modifications. Better performing entries (either 'Fixed' or baseline) are highlighted for the primary metrics EER and WER. The column 'Submitted' indicates the results we have shared before the submission deadline. The column 'Fixed' indicates the results we obtained after fixing a bug within our system, counting as 'late submission'. Weighted average per challenge guidelines is denoted with* **Ø**.

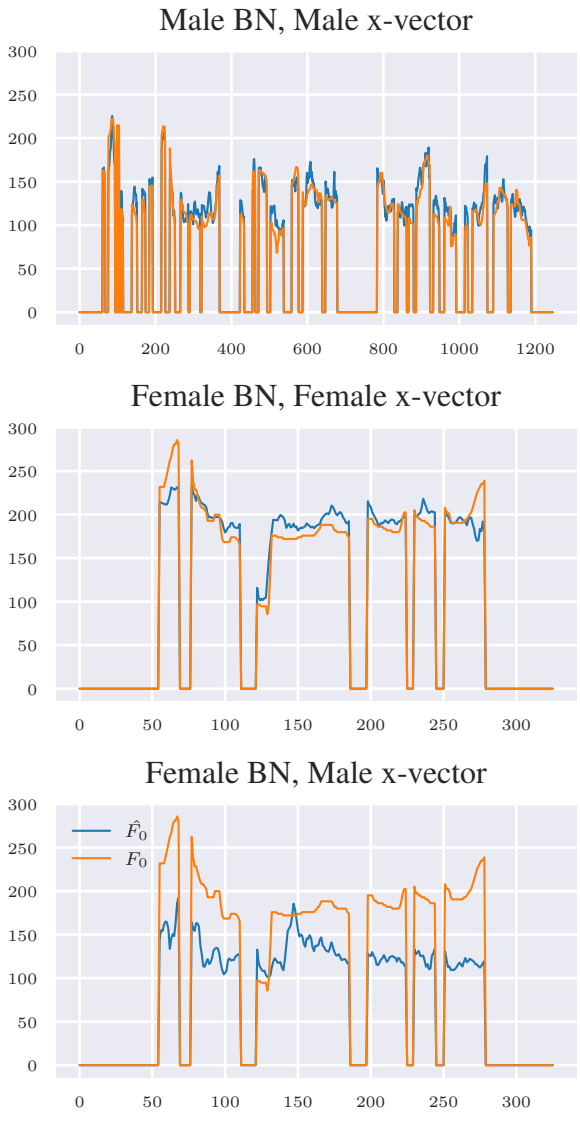

Figure 4: *Ground truth F0 estimates (orange) for the input signal, obtained by YAAPT [11] (F0 extractor of the B1 baselines) together with the F0 estimates obtained by our system.*

duction in the voice distinctiveness G_{VD} of the baseline system, but improved the EER anonymization metric by 2.7 times. Furthermore, we observed a more natural sounding voice synthesis, especially in conditions of cross-gender voice conversion.

5. References

[1] 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:1905.13561 [cs, eess, stat]*, May 2019, 00061 arXiv: 1905.13561. [Online]. Available: http://arxiv.org/abs/1905.13561

[2] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The VoicePrivacy 2020 Challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, Jul. 2022, 00015 arXiv:2109.00648 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2109.00648

[3] P. Champion, D. Jouvet, and A. Larcher, "A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender," *arXiv:2101.08478 [cs, eess]*, Jan. 2021, 00009 arXiv: 2101.08478. [Online]. Available: http://arxiv.org/abs/2101.08478

[4] U. E. Gaznepoglu and N. Peters, "Exploring the Importance of F0 Trajectories for Speaker Anonymization using X-vectors and Neural Waveform Models," *Workshop on Machine Learning*

in Speech and Language Processing 2021, Sep. 2021, 00002. [Online]. Available: https://arxiv.org/abs/2110.06887v1

[5] L. Tavi, T. Kinnunen, and R. G. Hautamäki, "Improving speaker de-identification with functional data analysis of f0 trajectories," *Speech Communication*, vol. 140, pp. 1–10, May 2022, 00000 arXiv:2203.16738 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2203.16738

[6] S. Johar, "Psychology of Voice," in *Emotion, Affect and Personality in Speech: The Bias of Language and Paralanguage*, ser. SpringerBriefs in Electrical and Computer Engineering, S. Johar, Ed. Cham: Springer International Publishing, 2016, pp. 9–15, 00014. [Online]. Available: https://doi.org/10.1007/978-3-319-28047-9_2

[7] X. Wang and J. Yamagishi, "Neural Harmonic-plus-Noise Waveform Model with Trainable Maximum Voice Frequency for Text-to-Speech Synthesis," *arXiv:1908.10256 [cs, eess]*, Aug. 2019, 00027 arXiv: 1908.10256. [Online]. Available: http://arxiv.org/abs/1908.10256

[8] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in neural information processing systems (NeurIPS)*, Vancouver, Canada, 2019, p. 12, 17858.

[9] V. Fomin, J. Anmol, S. Desroziers, J. Kriss, and A. Tejani, "High-level library to help with training neural networks in PyTorch," 2020, 00014 Publication Title: GitHub repository. [Online]. Available: https://github.com/pytorch/ignite

[10] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," Jul. 2019, 01169 Number: arXiv:1907.10902 arXiv:1907.10902 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1907.10902

[11] K. C. Ho and M. Sun, "An Accurate Algebraic Closed-Form Solution for Energy-Based Source Localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2542–2550, Nov. 2007, 00111 Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[12] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J. F. Bonastre, "The voiceprivacy 2022 challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[13] Baseline results for `joint-hifigan`. (last accessed 2022-07-31). [Online]. Available: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2022/blob/master/baseline/results/RESULTS_summary_tts_joint_hifigan