Study on Signal Processing Techniques in Protecting Voice Personae Against Speech Synthesis Systems

Nopparut Li, Candy Olivia Mawalim, and Masashi Unoki
Japan Advanced Institute of Science and Technology
E-mail: {nopparut.li, candylim, unoki}@jaist.ac.jp

***Abstract*—Recent advancements in speech synthesis have enabled the generation of natural-sounding speech signals that can closely mimic specific speakers, raising serious concerns about the misuse of voice recordings for impersonation and fraud. Although researchers have extensively studied spoof speech detection, they can only implement these approaches once the spoofed content has been generated. We propose a method based on F_0 component elimination and compare it with conventional filtering and artificial reverberation for the impact on the quality of synthesized speech signals generated using the TorToiSe TTS model, as well as the perceptual quality of the modified speech signals. Results show that the proposed method is able to reduce the identifiability of synthesized speech signals with minimal impact on speech quality, offering a promising direction for voice personae protection against speech synthesis systems.**

I. Introduction

In recent years, speech synthesis systems have made remarkable progress in generating highly natural and intelligible speech signals. These systems can now reproduce the unique vocal characteristics of a specific speaker using only a few reference samples [1], [2]. While these capabilities offer benefits in applications like virtual assistants and automated communication, they also raise serious security concerns. Malicious actors can exploit online voice samples, such as those of a victim's friend or family member, to synthesize impersonated speech signals for fraudulent purposes [3].

To mitigate the emerging threats, many existing works have proposed spoof speech detection (SSD) systems, which aim to distinguish genuine speech signals from synthetic or manipulated signals [4]. Although SSD systems have achieved impressive accuracy, their reactive nature means that they can only detect spoofed audio after it has been generated, making them unsuitable as a standalone preventive measure.

Recent research has explored proactive approaches that attempt to prevent speech synthesis models from accurately replicating a speaker's voice in the first place. Drawing inspiration from adversarial examples in computer vision, where imperceptible perturbations can mislead classifiers [5], studies have adapted the idea to generative models. For example, poisoning face images can prevent misuse in face synthesis tasks [6]. Similar concepts are applied to speech signals: Huang et al. [7] proposed injecting adversarial noise to disrupt white-box voice conversion systems, while Yu et al. [8] introduced model-generalized perturbations using ensemble learning to protect against multiple synthesis models. Although such methods can protect speaker identity, the added noise can remain audible, which limits their practical adoption.

To address this, we explore a set of signal modification methods to find a modification that can affect speech synthesizer performance with minimal impact on the quality of the original speech signals. These include traditional filtering approaches, artificial reverberation, and a proposed method for eliminating the fundamental frequency (F_0) of speech signals. We evaluate these techniques by measuring both their impact on a speech synthesis system and their degradation of voice personae reproduction, which reveals the trade-off between protection strength and audio quality.

In the rest of this paper, Section II describes the details of the targeted signal modification methods. Section III shows details and settings of the experiment conducted. Section IV elaborates on the results of the experiment, and Section V summarizes the study.

II. Signal Modification Methods

Multiple modification methods were considered for evaluating the audio quality degradation and effectiveness against a speech synthesis system. The modification methods are as follows:

A. High-pass, low-pass, and band-pass filtering

A high-pass filter eliminates the frequency components of the audio input that are below the given cutoff frequency, while a low-pass filter eliminates the frequency components that are above the cutoff frequency. A band-pass filter is the combination of the high-pass and low-pass filters, allowing only the signal component within a given range to be kept. The prominent characteristic of these modification methods is the utilization of harmonics to allow human listeners to fill in the missing frequency components.

In this experiment, the cutoff frequencies considered for the high-pass filter are 500 Hz and 1000 Hz, while those for the low-pass filter are 1000 Hz and 4000 Hz. As for the band-pass filter, the band-pass frequencies are 500 to 4000 Hz and 1000 to 4000 Hz. The type of filter used for these modifications is the IIR Butterworth filter [9]. The formula is shown as follows:

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_N z^{-N}}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_N z^{-N}} \quad (1)$$

where N represents the N^{th} order of the filter, and a_N and b_N represent the denominator and numerator coefficients,

979-8-3315-7206-8/25/$31.00

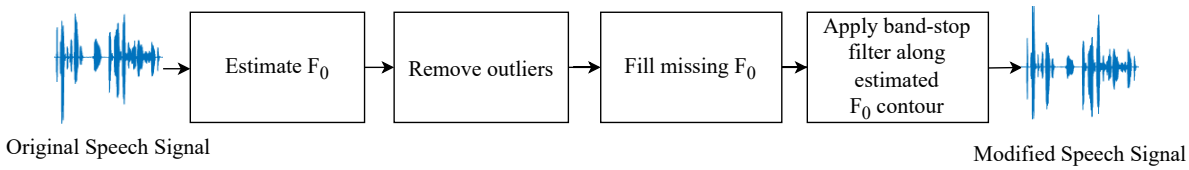

Fig. 1: Process diagram of the proposed F_0-elimination method. The pipeline estimates the f_0 contour using DIO, removes outliers, interpolates gaps, and applies a time-varying IIR notch filter that tracks f_0. This sequence ensures consistent suppression of the fundamental component while preserving harmonic structure.

respectively. The coefficients are derived by designing an analog filter in accordance with the filter type, order, cutoff frequency, and sampling frequency. Frequency transformation is also applied in cases of high-pass and band-pass filters. Then a bilinear transformation is used to convert the filter into the digital domain and get the coefficients.

B. Artificial reverberation

Reverberation is included as one of the signal modification methods in this study, based on the assumption that low levels of environmental reverberation are common in natural listening conditions. As such, minor reverberation is likely to be perceptually negligible to human listeners, helping to preserve speech quality while introducing subtle signal alteration to speech signals.

In this experiment, artificial reverberation is generated by filtering the original utterances with a Schroeder room impulse response (RIR) [10]. The impulse response is constructed by modulating white Gaussian noise with an exponentially decaying envelope:

$$h[n] = e_h[n] \cdot c[n] = a \cdot \exp\left(-6.9[n]/\mathrm{TR}\right) c[n],$$

where $h[n]$ is the RIR, $e_h[n]$ is the temporal amplitude envelope of RIR, $c[n]$ is a white Gaussian noise carrier, 6.9 ensures -60 dB decay over the target reverberation time $\mathrm{TR} = 0.1$ s, and a is a normalization factor to maintain unit energy. The resulting impulse response is applied using FIR filtering to simulate a mild reverberant effect.

C. F_0 component elimination filtering

Based on the results of high-pass filtering, we speculate that the removal of the F_0 component of human speech signals can have an important role in disrupting a speech synthesis system with less impact on speech quality. As such, we proposed a method of eliminating F_0 from speech signals to fully utilize the effect of missing fundamental.

The process diagram of the method is shown in Fig. 1. The pseudocode for the process with variable bandwidth IIR filter based on [12] is as shown in Algorithm 1. Example spectrograms of the original and modified audio are as shown in Fig. 2.

Algorithm 1 F_0 Component Elimination Filtering

Input: Speech signal $x[n]$, sampling rate f_s
Estimate F_0 contour $F_0[n]$ using DIO [11]
Remove outliers beyond $\pm 2\sigma$ from $F_0[n]$ to avoid pitch doubling
Fill missing values in $F_0[n]$:
- Extend first and last valid values to boundaries
- Linearly interpolate internal gaps

for each time frame n **do**
  Set center frequency: $F_c[n] \leftarrow F_0[n]$
  Set bandwidth: $BW[n] \leftarrow 1.4 \cdot F_c[n]$
  Compute normalized values:

$$\omega_c[n] \leftarrow \frac{2\pi F_c[n]}{f_s}, \Delta[n] \quad \leftarrow \frac{2\pi BW[n]}{f_s}, r[n] \leftarrow 1 - \frac{\Delta[n]}{2}$$

  Construct second-order IIR band-stop filter:

$$H(z,n) = \frac{1 - 2\cos(\omega_c[n])z^{-1} + z^{-2}}{1 - 2r[n]\cos(\omega_c[n])z^{-1} + r[n]^2 z^{-2}}$$

  Apply $H(z,n)$ to $x[n]$
end for
Output: Filtered speech signal $\hat{x}[n]$

III. Experimental Setups

To evaluate the effectiveness in protecting voice personae, each signal processing technique is applied to original utterances individually. The modified utterances are then used as references for a speech synthesis system to synthesize impersonating speech signals. The synthesized speech signals are then used to evaluate the protection effectiveness of the corresponding modification method, while the modified speech signals are used to evaluate the audio quality degradation of the modification. The process is summarized in Fig. 3.

A. Baseline method (noise addition)

For the purpose of ensuring that the effectiveness of any particular modification method is not solely dependent on the distortion intensity, white noise and pink noise addition to the original speech signals are used as simple comparisons for each modification method.

During this experiment, the signal-to-noise ratio (SNR) in comparison to the distortion strength of the previously men-

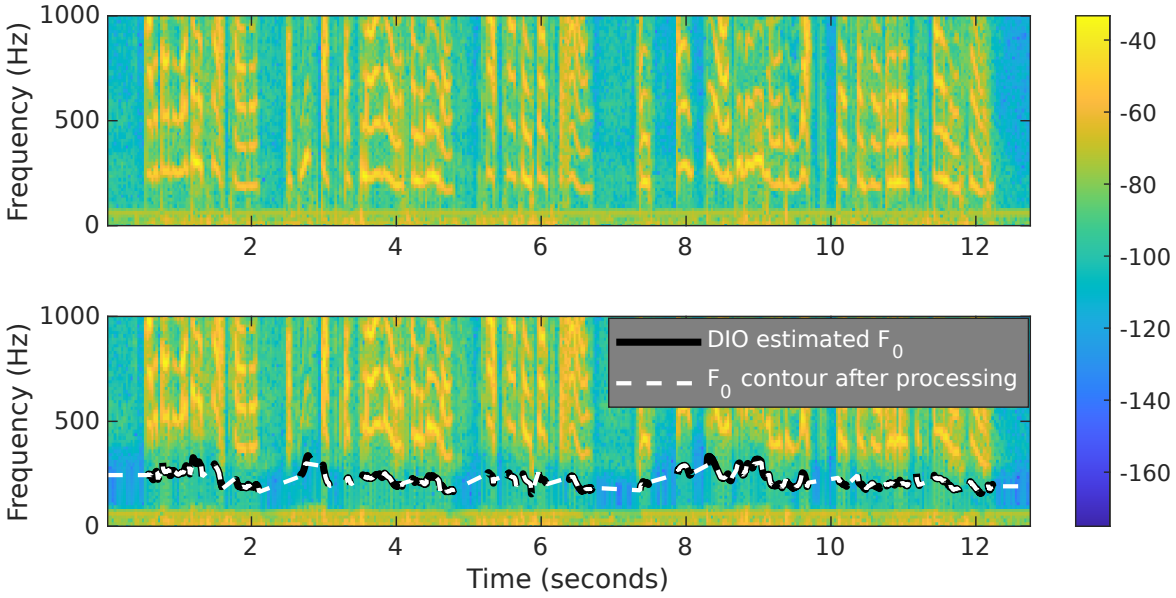

Fig. 2: Spectrogram comparison of an original utterance (top) and its F_0-eliminated version (bottom). Note that the fundamental frequency band is suppressed in the modified signal, while higher harmonics remain largely intact. This illustrates the missing-fundamental effect: listeners perceive pitch continuity even without the explicit F_0.

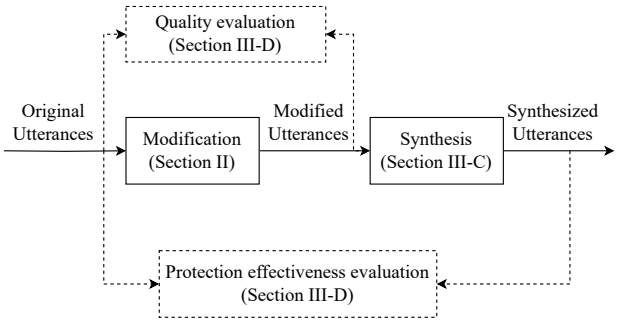

Fig. 3: Block diagram of the experimental process. Dashed line indicates that the data is used for evaluations.

tioned modification methods is considered, which are SNRs of 30 dB, 7 dB, 0 dB, and -5 dB.

B. Dataset

In this experiment, the LibriSpeech corpus [13] is chosen because of its low-noise characteristics, as they are created for developing automatic speech recognition systems. The low noise characteristic of the corpus represents the high-quality speech signals that are becoming common online, which we aim to protect the voice personae.

A total of 10 speakers are selected from this corpus, which are five males and five females speakers. 100 utterances were then selected from each speaker to be applied with modification techniques and used as references for the speech synthesis system. In total, 1000 utterances are used as reference speech signals in this study.

C. Speech synthesis system

The speech synthesis system used for this study is the TorToiSe Text-to-Speech (TTS) proposed by Betker [1], which is based on the improvement in image generation studies. This TTS model is selected for its leading performance in the speech synthesis field. In this study, we provide the TTS model with reference audio samples based on each speaker and modification method. The model will have 17 voice reference groups per speaker (8 modifications, 8 baseline white and pink noises, and 1 original utterances), resulting in a total of 170 voice reference groups for the model to synthesize. For each reference speaker, the TTS system synthesized speech using the exact transcript of that speaker's original utterances. As such, all of the original utterances will have their own modified, synthesized, and synthesized-from-modified counterparts. In total, there are 17 voice reference groups $\times$ 10 speakers $\times$ 100 utterances $= 17,000$ synthesized speech signals to be used for evaluation.

D. Evaluation Metrics

1) Speech Quality Evaluation Metrics: To evaluate the impact of each modification method on speech quality and intelligibility, we adopt four objective metrics: DNSMOS [14], Non-Intrusive Speech Quality Assessment (NISQA) [15], Short-Time Objective Intelligibility (STOI) [16], and Perceptual Evaluation of Speech Quality (PESQ) [17].

TABLE I: Mean Value of Objective Evaluation of Each Signal Modification Method. Modified speech signals are evaluated on the impact of the corresponding modification method on the speech quality. Synthesized speech signals are evaluated on the contribution of the corresponding modification method in voice personae protection against the speech synthesis system.

Modification Method	**Modified speech signals**						**Synthesized speech signals**		
	PESQ↑	DNSMOS↑	NISQA↑	STOI↑	Cos. Sim.↑	ASV-EER↓	NISQA↓	Cos. Sim.↓	ASV-EER↑
Original unmodified utterances	4.549	3.888	3.426	1.000	–	–	3.325	0.686	7.69
White noise addition (−5 dB SNR)	1.232	1.588	0.912	0.590	0.671	0.336	2.930	0.078	40.00
White noise addition (0 dB SNR)	1.328	1.778	0.894	0.687	0.780	0.030	2.724	0.111	33.55
White noise addition (7 dB SNR)	1.607	2.205	0.967	0.806	0.850	0.000	2.548	0.190	20.97
White noise addition (30 dB SNR)	3.837	3.451	2.922	0.982	0.919	0.000	2.986	0.633	8.13
Pink noise addition (−5 dB SNR)	1.236	1.641	0.828	0.560	0.665	0.040	2.846	0.121	39.95
Pink noise addition (0 dB SNR)	1.366	1.790	0.847	0.675	0.792	0.000	2.570	0.128	34.75
Pink noise addition (7 dB SNR)	1.721	2.467	1.046	0.815	0.861	0.000	2.296	0.211	22.36
Pink noise addition (30 dB SNR)	3.967	3.596	3.326	0.981	0.927	0.000	3.016	0.643	8.22
High-pass filtering (500 Hz cutoff)	4.108	2.939	2.287	0.877	0.534	**0.000**	2.683	0.088	24.52
High-pass filtering (1000 Hz cutoff)	3.618	2.480	1.980	0.777	0.415	0.306	2.396	0.052	38.46
Low-pass filtering (1000 Hz cutoff)	3.441	3.431	1.456	0.793	0.492	0.484	2.973	0.136	37.78
Low-pass filtering (4000 Hz cutoff)	**4.548**	**3.884**	**3.297**	**0.998**	**0.755**	**0.000**	2.723	0.342	8.93
Band-pass filtering (500–4000 Hz)	3.871	2.731	2.316	0.810	0.372	0.899	**2.024**	0.017	40.14
Band-pass filtering (1000–4000 Hz)	3.189	2.338	1.314	0.678	0.072	27.072	2.255	**0.000**	**47.41**
Artificial reverberation (0.1 s)	3.030	3.383	2.871	0.849	0.723	**0.000**	3.282	0.592	8.03
F_0 component elimination	3.854	3.458	3.218	0.877	0.677	**0.000**	2.959	0.271	17.02

↑ indicates the evaluation within the column desires a higher evaluation value, while ↓ desires a lower value.
Bolded and underlined numbers indicate the best and second-best evaluation results within the modification group.
The proposed method is highlighted with a light gray background color.

DNSMOS and NISQA are non-intrusive metrics that estimate the Mean Opinion Score (MOS) of speech signals without requiring a clean reference. In this study, they are used to evaluate the perceived quality of modified speech signals, with NISQA also applied to synthesized signals. Both metrics output a MOS score from 1 (poor) to 5 (excellent). For modified speech signals, a higher MOS indicates minimal degradation, while for synthesized speech signals, a lower score may suggest that the modification disrupts voice quality, potentially reducing speaker similarity.

STOI assesses the intelligibility of modified speech signals. As an intrusive metric, it requires a clean reference and produces scores from 0 (unintelligible) to 1 (fully intelligible). Higher scores indicate better preservation of speech content. STOI is not applicable to synthesized speech signals due to the lack of a reference signal.

PESQ is an intrusive metric originally developed to evaluate perceptual speech quality in telecommunication systems, producing scores from 1.0 to 5.0, with higher values indicating better quality. As with STOI, it is used only for modified speech signals in our study, since synthesized signals lack a clean reference. High PESQ scores suggest the modification preserves quality close to the original.

2) Voice Personae Protection Evaluation Metrics: To evaluate the effectiveness of each method in protecting voice identity, we use speaker embedding cosine similarity and the Equal Error Rate (EER) from an Automatic Speaker Verification (ASV) system.

Speaker embeddings are extracted using NVIDIA's TitaNet [18], specifically the TitaNet-Large which is pretrained with the same LibriSpeech data used in our experiment to ensure the speaker characteristic is able to be extracted by the model. Cosine similarity compares embeddings of original and modified or synthesized speech signals, ranging from -1 (opposite) to 1 (identical). Lower similarity in synthesized speech signals indicates effective voice personae protection, while higher similarity in modified speech signals suggests identity preservation.

ASV-EER is calculated by comparing embedding pairs from the same and different speakers. It reflects the point where false acceptance and rejection rates are equal, ranging from 0% (perfect identification) to 50% (random guessing). Low EER indicates identity preservation, while high EER implies reduction in speaker identifiability.

These metrics collectively assess both perceptual quality and identity masking performance across all modification methods studied.

IV. Evaluations and Results

Table I summarizes all objective metrics, and Figs. 4 and 5 visualize quality and identity outcomes. Below, we highlight the main findings and point to the relevant evidence rather than listing full numeric ranges in-text.

F_0 elimination: quality preserved, identifiability reduced: F_0 elimination preserves perceived quality close to the unmodified condition while lowering synthesizer identifiability. It maintains high DNSMOS/NISQA/PESQ and raises ASV-EER on synthesized speech relative to the original baseline (Table I; see also Fig. 4 for quality and Fig. 5 for identity). In short, suppressing the pitch component offers a favorable protection–fidelity trade-off.

Noise baselines: stronger protection at the cost of heavy degradation: Adding white/pink noise can yield higher ASV-EERs than F_0 elimination, but only by imposing substantial quality loss on the modified audio (Table I). As Fig. 4 shows, MOS-type scores drop sharply as SNR decreases. Thus, while

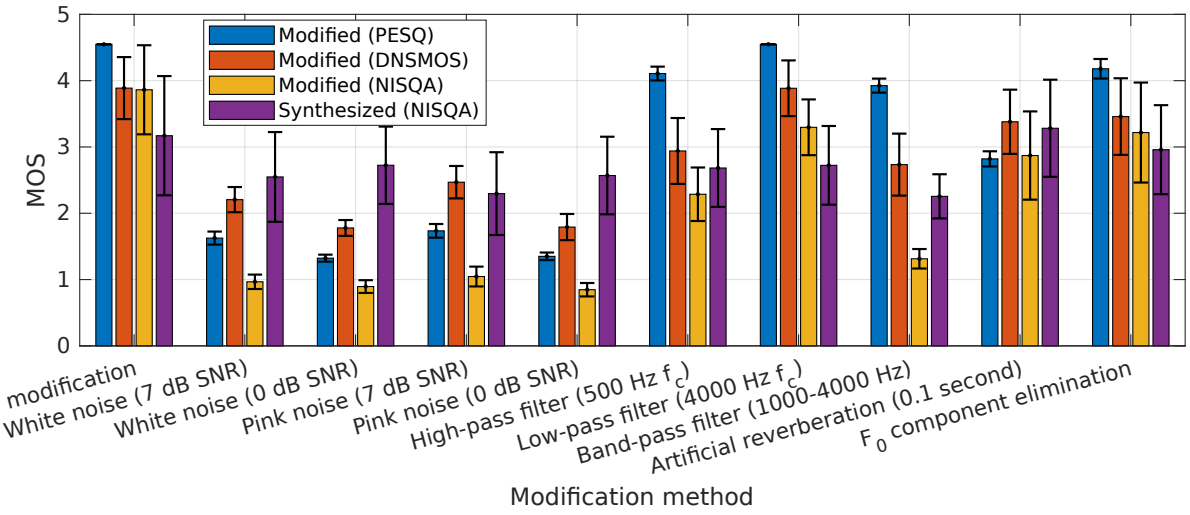

Fig. 4: Mean and standard deviation of MOS-type quality metrics (PESQ, DNSMOS, NISQA) for modified speech. Higher values indicate better perceptual quality. Observe that F_0 elimination retains scores close to the unmodified condition, whereas noise addition and narrow band-pass filtering sharply reduce quality.

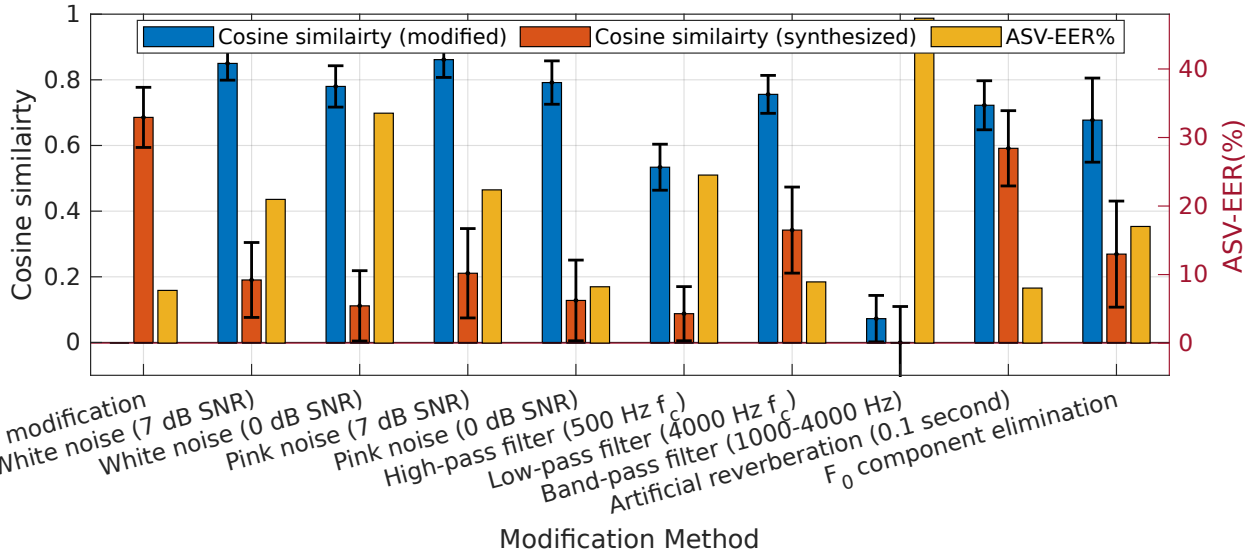

Fig. 5: Speaker-embedding cosine similarity and ASV-EER for modified and synthesized speech. For modified speech, high cosine similarity confirms identity preservation. For synthesized speech, F_0 elimination reduces cosine similarity and increases ASV-EER, indicating effective protection against impersonation while minimally affecting original identity.

effective as a stress test, noise is a poor practical baseline for user-facing protection.

Filter families: extremes trade quality for protection: Simple spectral shaping exhibits the expected pattern. A wide low-pass (e.g., cutoff at 4 kHz) largely preserves quality yet yields little protection (Table I, Fig. 5). Narrow band-pass settings, by contrast, can drive ASV-EER higher but at a pronounced cost to quality (Fig. 4). These outcomes bracket the operating region where F_0 elimination sits—closer to the high-quality end while still reducing identifiability.

High-pass as a secondary compromise: A moderate high-pass (e.g., 500 Hz) offers a middle ground: protection stronger than low-pass with quality that remains acceptable for many use cases (Table I). However, compared with F_0 elimination, it generally shows lower non-intrusive quality scores at comparable protection levels (Figs. 4–5).

Artificial reverberation: quality holds, protection limited: Mild artificial reverberation maintains reasonable quality but provides limited gains in ASV-EER (Table I). This suggests reverberation alone is not an effective primary protection mechanism at perceptually acceptable strengths.

Takeaway: Across methods, targeting pitch via F_0 suppression is the most reliable way to attenuate synthesizer identifiability without paying a large quality penalty. Noise

can push protection further but at impractical degradations, with narrow band-pass behaves similarly. High-pass is a workable secondary compromise. Overall, F_0 elimination delivers the best balance between protection and quality preservation among the evaluated approaches.

V. Conclusion

This study evaluated the effectiveness of various signal processing techniques for protecting voice personae from unauthorized speech synthesis. We examined traditional methods, including high-pass, low-pass, and band-pass filtering, as well as artificial reverberation, and introduced a technique, F_0 component elimination filtering. Among all methods tested, the introduced method demonstrated a minimal impact on the quality of the original speech signals while also being able to hinder the performance of the speech synthesis system. The objective evaluations indicate that the F_0 component elimination achieves high speech quality while effectively reducing speaker identifiability of synthesized speech signals. These results suggest that targeting and suppressing the F_0 contour is a viable strategy for proactive voice personae protection. In future work, human listening tests are to be conducted. In addition, this technique could be combined with adversarial methods to enhance protection further while minimizing audible artifacts.

Acknowledgment

This work was supported by JSPS KAKENHI (23K18491, 25K21245, and 25H01139). This work was also supported by JST Program for co-creating startup ecosystem (JPMJSF2318).

References

[1] J. Betker, "Better speech synthesis through scaling," *arXiv preprint arXiv:2305.07243*, 2023.

[2] J.-C. Chou, C.-C. Yeh, and H.-Y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *arXiv preprint arXiv:1904.05742*, 2019.

[3] A. Giuffrida, *AI phone scam targets Italian business leaders including Giorgio Armani*, The Guardian, [Online], Feb. 2025. [Online]. Available: https://www.theguardian.com/world/2025/feb/10/ai-phone-scam-targets-italian-business-leaders-including-giorgio-armani.

[4] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: Challenges and future directions," *Algorithms*, vol. 15, no. 5, p. 155, 2022.

[5] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2015.

[6] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *29th USENIX security symposium (USENIX Security 20)*, 2020, pp. 1589–1604.

[7] C.-Y. Huang, Y.-C. Lin, H.-Y. Lee, and L.-S. Lee, "Defending your voice: Adversarial attack on voice conversion," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 552–559.

[8] Z. Yu, S. Zhai, and N. Zhang, "AntiFake: Using adversarial audio to prevent unauthorized speech synthesis," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 460–474. DOI: 10.1145/3576915.3623209.

[9] S. Butterworth, "On the theory of filter amplifiers," *Wireless Engineer*, vol. 7, pp. 536–541, 1930.

[10] M. R. Schroeder, "Modulation transfer function: Definition and measurement," *Acustica*, vol. 49, pp. 179–182, 1981.

[11] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, Audio Engineering Society, 2009.

[12] A. G. Constantinides, "Spectral transformations for digital filters," in *Proceedings of the Institution of Electrical Engineers*, IET, vol. 117, 1970, pp. 1585–1590.

[13] 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)*, South Brisbane, QLD, Australia, 2015, pp. 5206–5210.

[14] C. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.

[15] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv:2104.09494*, 2021.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[17] *ITU-T recommendation p.862. perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunication Union. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.

[18] N. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8102–8106.