

Evaluation of the Speech Resynthesis Capabilities of the VoicePrivacy Challenge Baseline B1

Ünal Ege Gaznepoglu, Nils Peters

International Audio Laboratories, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

{ege.gaznepoglu, nils.peters} @fau.de

Abstract

Speaker anonymization systems continue to improve their ability to obfuscate the original speaker characteristics in a speech signal, but often create processing artifacts and unnatural sounding voices as a tradeoff. Many of those systems stem from the VoicePrivacy Challenge (VPC) Baseline B1, using a neural vocoder to synthesize speech from an F0, x-vectors and bottleneck features-based speech representation. Inspired by this, we investigate the reproduction capabilities of the aforementioned baseline, to assess how successful the shared methodology is in synthesizing human-like speech. We use four objective metrics to measure speech quality, waveform similarity, and F0 similarity. Our findings indicate that both the speech representation and the vocoder introduces artifacts, causing an unnatural perception. A MUSHRA-like listening test on 18 subjects corroborate our findings, motivating further research on the analysis and synthesis components of the VPC Baseline B1.

Index Terms: speaker anonymization, x-vector, bottleneck features, F0, neural source-filter (NSF), quality evaluation

1. Introduction

Numerous developments in the speech signal processing domain have rendered the collection of speech data as well as its adversarial utilization simpler [1]. As a result, voice privacy is an emerging issue in today’s world. Many technical applications either require by law, or would benefit from, a preliminary processing to mitigate the risks to user privacy. In this regard, a VoicePrivacy Challenge (VPC) has been organized to promote the development of voice anonymization systems via the introduced baselines, evaluation metrics and attack models, which are widely adopted by the researchers in the field.

Depending on the downstream task, i.e., the purpose the acquired speech signals shall serve, the anonymization procedure may be expected to preserve the prosody and the naturalness. One such use case is a psychiatric support context where the patients want to stay anonymous [2]. However, the results from the VPC 2020 and 2022 point out that none of the published systems up to date can achieve subjective naturalness scores on par with recorded human speech [3], [4]. Furthermore, our previous work utilizing contrastive systems revealed that using original x-vectors during synthesis surprisingly yields worse utility and an increase in the privacy [5]. Therefore, in this work, we evaluate the speech resynthesis capabilities of the VPC Baseline B1, using metrics from other domains, to understand if the speech representation and synthesis block shared across systems of multiple contestants have any improvement potential.

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

2. Related work

2.1. VPC Baseline B1 and its derivatives

The VoicePrivacy Challenge Baseline B1 has been a source of inspiration to many challenge participants [3], [4]. The system [6], which consists of three feature extractors, an anonymization block, and a neural vocoder, is depicted in Figure 1. The feature extractors and their purposes are outlined in the Table 1.

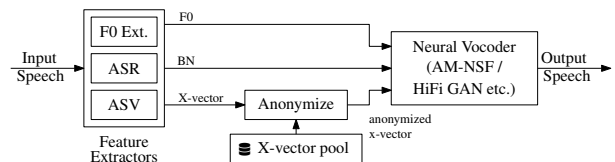


Figure 1: The VPC 2022 Baseline B1.a/b.

Table 1: Extracted features per utterance. The quantity in parentheses indicates the resulting tensor shape. N : number of frames of an utterance. W : window size (ms), H : hop size (ms)

Feature (purpose)	Extractor	Properties
F0 (Prosody)	YAAPT	($N \times 1$), W : 35, H : 10
BN (Verbal content)	TDNN-F	($N \times 256$), H : 10
X-vector (Identity)	TDNN	(1×512)

More than 10 systems are proposed to improve the various aspects of the baseline over the last three years. Majority of these contributions target the anonymization block and keep the speech representation or the vocoder intact. Some however, such as [7], propose alternatives to the bottleneck features. For the speaker embedding, [8] proposes switching to ECAPA and [9] reports increased speaker representation capabilities when both ECAPA and x-vectors are used together.

The neural source-filter (NSF), i.e., the neural vocoder, has also received some attention. The 2022 edition of the challenge included two vocoders (NSF-HiFiGAN and HiFiGAN) that directly predict the waveforms from the speech representation, discarding the acoustic model (AM) that was present in the 2020 baseline. Works such as [9], [10] use the IMS Toucan toolkit that provides a modular neural vocoder and utilize local energy in addition to F0 for further prosody control.

2.2. Evaluation of voice conversion systems

The voice anonymization problem, especially the way VPC framework treats it, has some similarities to the voice conversion problem. An overview on voice conversion mentions intrusive metrics like perceptual evaluation of speech quality

(PESQ) and mel-cepstral distortion (MCD) to evaluate synthesized speech quality [11]. In our study, the availability of the reference signals lets us employ such methods. Recently, torchaudio-SQUIM was proposed to estimate metrics such as PESQ on synthesized speech without needing a reference [12].

2.3. Evaluation of voice anonymization systems

The VPC framework introduced objective and subjective metrics to evaluate different aspects of the anonymized speech signals [13]. The word error rate (WER), whose lower values indicate a better utility, is measured by an automated speech recognition (ASR) system. An automated speaker verification (ASV) system is used to measure the anonymization success, where higher equal error rate (EER) values indicate better anonymization. Prosody retention to a certain extent is ensured by a lower bound on F0 correlation and finally, a gain of voice distinctiveness measures whether the speaker diversity of the input speech datasets are preserved by the anonymization process. However, none of the introduced objective metrics can successfully measure the perceived naturalness thus the challenge organizers have resorted to a subjective evaluation of the utterances [13].

The VPC community uses contrastive systems [5], [7], [14], an idea similar to the ablation studies performed by the machine learning community. A contrastive system is a marginally different configuration of the anonymization system that provides further insights into how different modules thereof contribute to the performance. The cited studies use VPC metrics to assess the privacy versus utility (ASR scenario) tradeoff and also reported that synthesis with original set of features cause an increase to EER as well as to WER, hinting that some artifacts are introduced by the analysis-synthesis pipeline.

To conclude, the existing objective metrics of the VPC do not account for naturalness. The alternative, subjective listening tests, are non-ideal because they are time-consuming and costly. Furthermore, the evaluation methods in the VoicePrivacy literature are not capable of detecting abnormal behaviors in time, which in our opinion is necessary to find what causes unnatural outputs. To go beyond the contrastive system studies with EER and WER, we decided to investigate whether intrusive metrics and their non-intrusive estimates could be exploited.

3. Methodology

3.1. Dataset

We use the VPC datasets `libri-*` and `vctk-*` for our evaluations. A summary of their content is provided in Tab. 2. These datasets are resynthesized using the systems in Table 3. The system `B1b-sp` is the same as the 2022 baseline, except it uses the original speaker-level x-vectors for synthesis. The system `B1b-utt`, using the utterance-level x-vectors for synthesis, imitates the training conditions of the neural vocoders. Both systems were trained using HiFiGAN discriminators. The system `joint-hifigan-sp` denotes the alternative vocoder (HiFiGAN [15]) provided by the VPC organizers. The system `am-nsf-sp` is the baseline used in

Table 2: VPC data subsets [13] utilized in this work. #F, #M: number of unique female/male speakers

Subset Name	#F	#M	#Utterances
libri-test- <code>{enrolls, trials}</code>	15	15	1934
vctk-test- <code>{enrolls, trials}</code>	15	15	12048

Table 3: Systems evaluated in this paper. The x-vectors are not anonymized to assess the resynthesis capability. Vocoders are VPC PyTorch implementations, unless noted in the table.

ID	X-vector	Vocoder
B1b-sp	speaker-level	joint NSF (+HiFiGAN-D)
B1b-utt	utterance-level	joint NSF (+HiFiGAN-D)
joint-hifigan-sp	speaker-level	joint HiFiGAN
am-nsf-sp	speaker-level	(C-based) AM + NSF
mel-nsf-sp	speaker-level	(C-based) NSF
mel-nsf-pt-sp	speaker-level	NSF
mel-nsf-sp-4k	speaker-level	(C-based) NSF

2020, that features an additional autoregressive AM that converts the speech representation into mel-spectrograms. The system `mel-nsf-sp` bypasses the AM and performs synthesis using the mel-spectrograms computed from original utterances, also referred to as *copy-synthesis* [16]. We feature both the PyTorch variant (denoted with a suffix `-pt`) and the C-based implementation utilized in VPC 2020. We also included an anchor equivalent `mel-nsf-sp-4k` that sets the mel-spectrogram values for frequency bands with $f_c > 4\text{kHz}$ to zero.

A number of pre-processing steps are performed before the evaluation. Systems we evaluate introduce different amounts of delay, so we align the outputs with the references using cross-correlation. Many of the utterances contain silence, as well as some pauses, hence we ran Silero voice activity detection [17] on the references and computed the metrics on the segments with voice. Also, a number of utterances were visually inspected to ensure that the synthesis procedure preserves the loudness, which could bias the evaluation scores [18].

3.2. Objective evaluation metrics

We adopt four different objective metrics to evaluate the resynthesis capabilities. These metrics are all intrusive, meaning that their computation requires access to a reference signal.

3.2.1. Mel-cepstral distortion (MCD)

MCD is used to measure the signal similarity in a perceptual sense. The implementation we use is provided by [19].

3.2.2. Scale-invariant signal-to-noise ratio (SI-SNR)

SI-SNR is used to measure the signal similarity [20]. The reference signal is first projected on the estimated signal, to obtain a scaling coefficient. Then the signal-to-noise ratio is computed. The implementation we use is a NumPy port of [21].

3.2.3. Perceptual evaluation of speech quality (PESQ)

PESQ is an intrusive measure introduced by ITU to predict the subjective speech quality evaluations. We use the implementation in `python-pesq` [22].

3.2.4. Gross pitch error (GPE)

GPE is a metric for F0 extractor evaluation. In our work, we use it to compare the synthesized F0 to the original. We adopt the definition in (1), also used in a previous work of us [5].

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

MCD, SI-SNR and GPE have the advantage that they could be computed on smaller segments.

3.3. torchaudio-SQUIM

In addition to the intrusive metrics, we also tested the torchaudio-SQUIM [12], which provides non-intrusive estimates of the intrusive metrics. We use their PESQ prediction and report numbers for all the classes as well as for the reference signals. If these estimates correlate well with their intrusive complements or with user preferences, SQUIM could be also tested for evaluating anonymized speech.

3.4. Subjective listening test

We conducted a MUSHRA-like listening test on 18 subjects of varying listening test experiences, using webMUSHRA software [23]. We randomly picked eight utterances from `libri-test` and six from `vctk-test`, (7 male and 7 female speakers, utterance lengths between [5.5, 8] seconds), which are available at ¹. The users are presented each synthesis output and asked to rate the naturalness using the following prompt, inspired by the VPC subjective test [13].

You will listen to a series of audio samples, comprising of both original recordings (referred to as reference) and versions that have been resynthesized using different neural vocoders, resulting in varying degrees of artifacts. Your task is to rate the naturalness of each recording.

Naturalness: Please judge how much audio degradation you can hear in each file. You need to select a score in the interval [0, 100], where higher numbers correspond to a more natural sounding audio, a 0 corresponding to 'severely degraded' and a 100 to 'no degradation at all'. For this score please only consider the sound characteristics and not the content. Also note that the reference contains some background noise. Finally, the deviations from original speaker's voice also count as degradations.

4. Results and Discussion

4.1. Objective evaluation

4.1.1. Signal similarity metrics

Figure 2 depicts the SI-SNR and MCD results. We saw no significant differences during visual inspection of the "per-dataset" and "per-gender" distributions. Therefore we display averages over datasets and gender instead.

Copy synthesis, e.g., `mel-nsf-pt-spk`, outperformed the others, but `mel-nsf-pt-spk` and `mel-nsf-spk`, two implementations of the same system, behaved differently. PyTorch copy-synthesis achieved better SI-SNR and MCD. The anchor, i.e., `mel-nsf-spk-4k`, attained comparable SI-SNR but the worst MCD. Other vocoders attained a similar MCD, standing between the copy synthesis and the anchor. We interpret the discrepancy between `mel-nsf` and synthesis from the representation as a sign of inadequacies of the utilized speech representation, resulting in some further information loss on top of the artifacts due to NSF. `am-nsf-spk` performed slightly better than other vocoders, indicating the AM contributed to the resynthesis performance.

4.1.2. F0 similarity

In a similar manner, Figure 3 depicts the GPE results. Behavior across female and male recordings are shown this time.

¹<https://audiolabs-erlangen.de/resources/2023-VPC-resynth-eval>

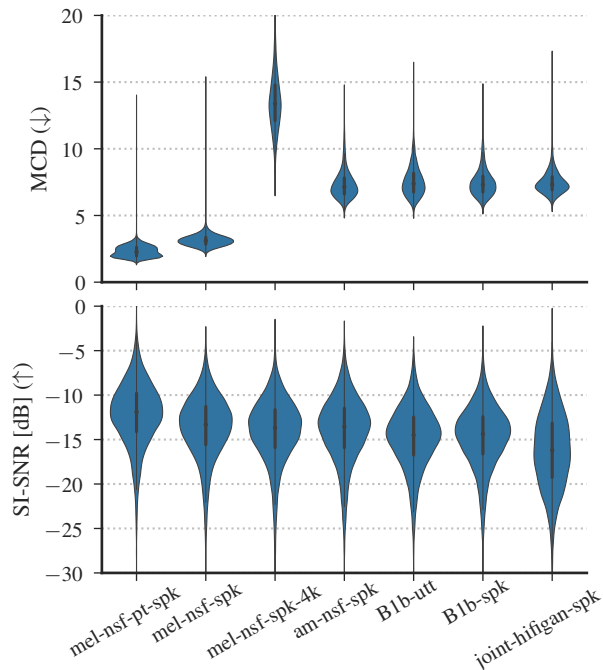


Figure 2: Evaluation results for signal similarity metrics.

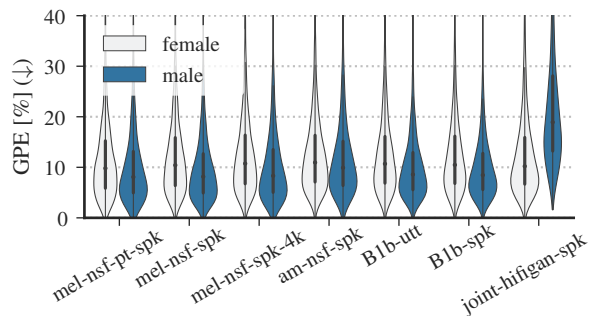


Figure 3: F0 similarity evaluation.

NSF-based systems maintained a certain standard in terms of F0 preservation, due to the source-filter model. HiFiGAN takes some extra liberty whilst synthesizing the signals, thus attained significantly higher GPE and this probably explains why it attained the worst SI-SNR too. An interesting outcome is that female speech has slightly higher GPE for NSF, but HiFiGAN corrupts pitch significantly more for male speakers.

4.1.3. PESQ and torchaudio-SQUIM

Finally, we compare the PESQ computations as well as PESQ estimates by torchaudio-SQUIM in Figure 4.

PESQ with respect to the reference (left) shows a similar, but a more pronounced version of the trend in the SI-SNR plots. PyTorch copy-synthesis, i.e., `mel-nsf-pt-spk`, attained the best PESQ scores. The anchor performed better than the variants that synthesize from the speech representation (e.g., `B1b`). `am-nsf-spk` performed slightly better than other vocoders, again hinting the joint AM-NSF approach introduced in 2022 causing minor degradation. `joint-hifigan-spk` performed the worst. On metrics that take the perceptual aspects into account, such as MCD and PESQ, `B1b-spk` performed better than `B1b-utt`, which imitates the vocoder training scenario. This may indicate an underfit. A number of factors could have caused this, such as an insufficient representation com-

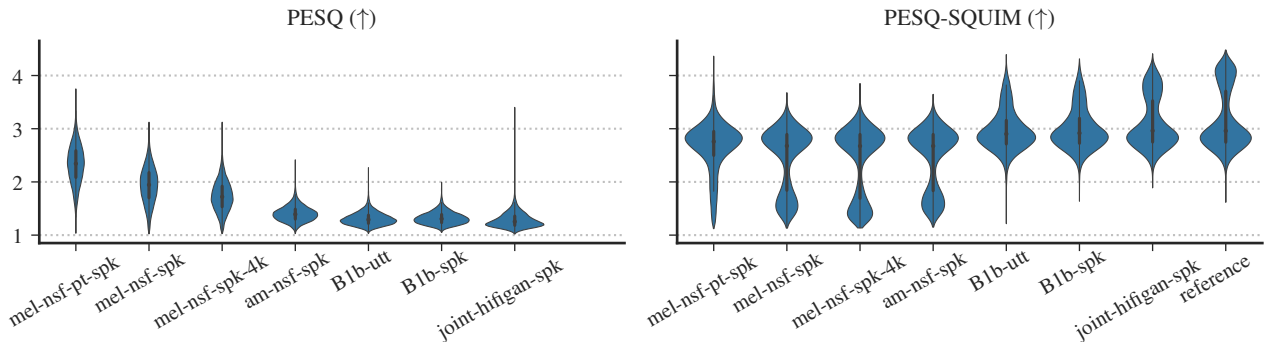


Figure 4: Evaluation results for PESQ and torchaudio-SQUIM estimate of PESQ.

plexity, lack of augmentation (augmenting x-vectors might help the vocoder to better learn the neighboring relations of the x-vectors) or may simply indicate that the training procedure has been cut off too early. The NSF was trained using L1 loss [6] on the magnitude spectrogram, which could be substituted with a perceptual loss to improve the performance.

Interestingly, the PESQ scores exhibit a greater inter-utterance variance for `mel-nsf-spk` variants. Additional investigations are required to understand this phenomenon. In particular it is crucial to understand if a confounding variable affects the scores, as previously shown by [18] with PESQ for factors such as loudness and alignment.

Torchaudio-SQUIM estimates of PESQ showed a different behavior. Systems `joint-hifigan-spk`, `B1b-utt` and `B1b-spk` achieved better performance with torchaudio-SQUIM evaluation. The systems have the HiFiGAN discriminators in common, which possibly explains the outcome. Some systems, e.g., `am-nsf-spk`, showed an unexpected bimodal distribution that is not explained by the gender or the dataset, which needs further investigations. The SQUIM estimates for the reference signals, depicted by the right-most violin plot in the Figure 4, again show a bimodal behavior.

4.2. Subjective listening test

The listening test responses are filtered such that the answers for an utterance, whose reference was rated with less than 90 points, are removed. This results in at least 14 subjects rating each utterance. The ratings are presented in Figure 5.

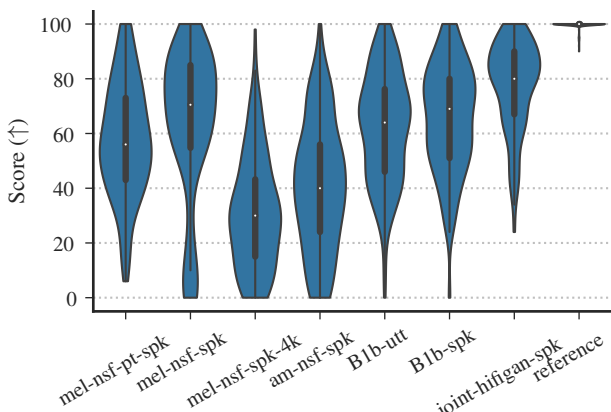


Figure 5: The subjective evaluation results.

Subjects reported that some `mel-nsf-spk` utterances had a severe muffling effect, often at their beginnings, rendering the part of the utterance completely unintelligible. In contrast, the `mel-nsf-pt-spk` was reported to suf-

fer from random impulsive artifacts, somewhat like a "sizzling frying pan constantly accompanying the recordings". `joint-hifigan-nsf` was reported to change the accents, "Americanizing" the voices, and the identity perception was different to what reference or other systems evoked. Otherwise the speech was reported to sound most human-like.

Now turning to the analysis of the gathered scores, most subjects were able to identify the reference stimuli and grade accordingly. Removed answers constitute less than 10% of the acquired data. The anchor `mel-nsf-spk-4k` was rated the lowest whereas `joint-hifigan-spk` was rated the best, except it compromises on the speaker identity. Systems other than `joint-hifigan-spk` exhibited a higher inter-utterance variance. Notwithstanding the few utterances with unintelligible segments causing a second modality at the bottom, the copy-synthesis, `mel-nsf-spk` was rated the second best, followed by `B1b` variants. `mel-nsf-pt-spk` and `am-nsf-spk` were rated only slightly better than the anchor.

4.2.1. Predictability of the subjective test scores

Intrusive metrics could not predict the outcome that `joint-hifigan-nsf` would be perceived the most natural, `B1b` performing better than `am-nsf-spk` and `mel-nsf-spk` outperforming `mel-nsf-pt-spk`. Only evaluation we ran that anticipated this outcome was torchaudio-SQUIM. We conclude that, the reference being available causes the intrusive evaluation to focus on the differences in signals that our subjects did not consider. Among the objective metrics, MCD was relatively successful.

Comparison of the PyTorch-based `mel-nsf-pt-spk` and C-based `mel-nsf-spk`, our subjects rated the latter better. The subjects penalized non-stationary artifacts less. To conclude, even though the objective metrics we utilize in this paper contribute to understanding how the blocks interact, these are not sufficient to explain the subject preferences completely.

4.3. Future work

We think it would be worthwhile to study the effects of using additional speaker embeddings such as ECAPA [24], as multiple systems in the literature utilized it and performed well in the VPC 2022. Part of ECAPA's success comes from a better temporal pooling strategy using attention. However, VPC simply uses temporal averaging to obtain the utterance-level x-vectors, and mere utterance averages to obtain speaker-level x-vectors, so modifications to these aspects are worth investigating.

Some of the metrics we used, e.g., MCD, GPE and SI-SNR allow computation on very small segments, unlike PESQ. The time segments with the reported muffling effect could be

automatically located with these and further analysis could be conducted. Also for these metrics, different temporal pooling strategies could be experimented with.

5. Conclusion

In this paper, we have investigated the reproduction capabilities of the VoicePrivacy Challenge Baseline B1 by utilizing a diverse set of objective evaluation metrics. Our subjective and objective evaluation results indicate that the copy synthesis scores better than the synthesis from representations, likely indicating the speech representation is causing additional information loss and yielding unnatural sounding output. Previous studies found that a more recent speaker embedding could help improve the anonymization performance, and our results hint that it could also improve the synthesized speech quality. In addition, the vocoder training scheme may benefit from a number of changes to bolster its understanding of the speaker embedding space.

The objective metrics we utilize in this work show limited effectiveness in evaluating the system behavior for anonymization, primarily because they are intrusive and no references are available for anonymization, and metrics we evaluated partially align with the listening test subject preferences. Torchaudio-SQUIM's PESQ implementation performed relatively well, and it does not require a reference, so the voice anonymization evaluations may benefit from it.

6. References

- [1] N. Tomashenko *et al.*, "Introducing the VoicePrivacy initiative," in *Proc. Interspeech Conf.*, 2020.
- [2] Ingo Siegert and Sebastian Stober. "AnonymPrevent - AI-based improvement of anonymity for remote assessment, treatment and prevention against child sexual abuse." (2021), [Online]. Available: <https://forschung-sachsen-anhalt.de/project/anonymprevent-based-improve-ment-anonymity-25374>.
- [3] N. Tomashenko *et al.*, "The VoicePrivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, 2022.
- [4] 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.
- [5] Ü. E. Gaznepoğlu and N. Peters, "Deep learning-based f0 synthesis for speaker anonymization," in *the forthcoming Proc. of the 31st European Signal Processing Conference (EUSIPCO)*, 2023.
- [6] F. Fang *et al.*, "Speaker anonymization using x-vector and neural waveform models," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019.
- [7] P. Champion, D. Jouvét, and A. Larcher, "Speaker information modification in VoicePrivacy 2020 toolchain," in *VoicePrivacy Challenge Submission*, 2020.
- [8] R. Khamsehashari *et al.*, "Voice privacy - leveraging multi-scale blocks with ECAPA-TDNN SE-res2next extension for speaker anonymization," in *2nd Symp. on Security and Privacy in Speech Communication*, 2022.
- [9] S. Meyer *et al.*, "Speaker anonymization with phonetic intermediate representations," in *Proc. Interspeech Conf.*, 2022.
- [10] S. Meyer *et al.*, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in *Proc. IEEE Spoken Lang. Tech. Workshop (SLT)*, 2022.
- [11] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2020.
- [12] A. Kumar *et al.*, "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
- [13] N. Tomashenko *et al.*, "2nd VoicePrivacy challenge evaluation plan." (2022), [Online]. Available: <https://arxiv.org/abs/2203.12468>.
- [14] C. O. Mawalim, K. Galajit, J. Karnjana, and M. Unoki, "X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system," in *Proc. Interspeech Conf.*, 2020.
- [15] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," presented at the Proc. NeurIPS Conf. 2020.
- [16] X. Wang and J. Yamagishi, "Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
- [17] A. Veysov, *Silero VAD: Pre-trained enterprise-grade voice activity detector (VAD)*, <https://github.com/snakers4/silero-vad>, 2021.
- [18] Z. Qiao, L. Sun, and E. Ifeachor, "Case study of PESQ performance in live wireless mobile VoIP environment," in *IEEE 19th Intl. Symp. on Personal, Indoor and Mobile Radio Communications*, 2008.
- [19] J. Sternkopf and S. Taubert. "Mel-cepstral-distance." (2022), [Online]. Available: https://github.com/jasminsternkopf/mel_cepstral_distance.
- [20] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?" In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [21] N. S. Detlefsen *et al.*, "TorchMetrics - measuring reproducibility in PyTorch," *Journal of Open Source Software*, vol. 7, no. 70, 2022.
- [22] M. Wang, C. Boedekker, R. G. Dantas, and A. Seelan. "Pesq: Python wrapper for PESQ score (narrow band and wide band)." (2022), [Online]. Available: <https://github.com/ludlows/python-pesq>.
- [23] M. Schoeffler *et al.*, "webMUSHRA — a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.
- [24] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech Conf.*, 2020.