INTERNATIONAL AUDIO LABORATORIES ERLANGEN
A joint institution of Fraunhofer IIS and Universität Erlangen-Nürnberg

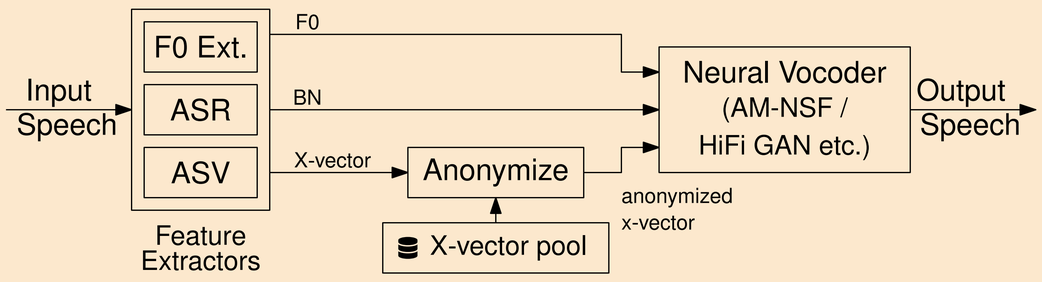
Evaluation of the Speech Resynthesis Capabilities of the VoicePrivacy Challenge Baseline B1

Ünal Ege Gaznepoglu, Nils Peters (ege.gaznepoglu@audiolabs-erlangen.de)

1. Introduction

Problem: Speaker anonymization systems sound unnatural

- Many systems build upon the VPC Baseline B1 [1]
  - Similar speech representation (F0, x-vector, bottleneck features)
  - Same vocoder (AM-NSF, HiFiGAN etc.)



Question: How to improve the naturalness? Which block is the culprit?

2. Methodology

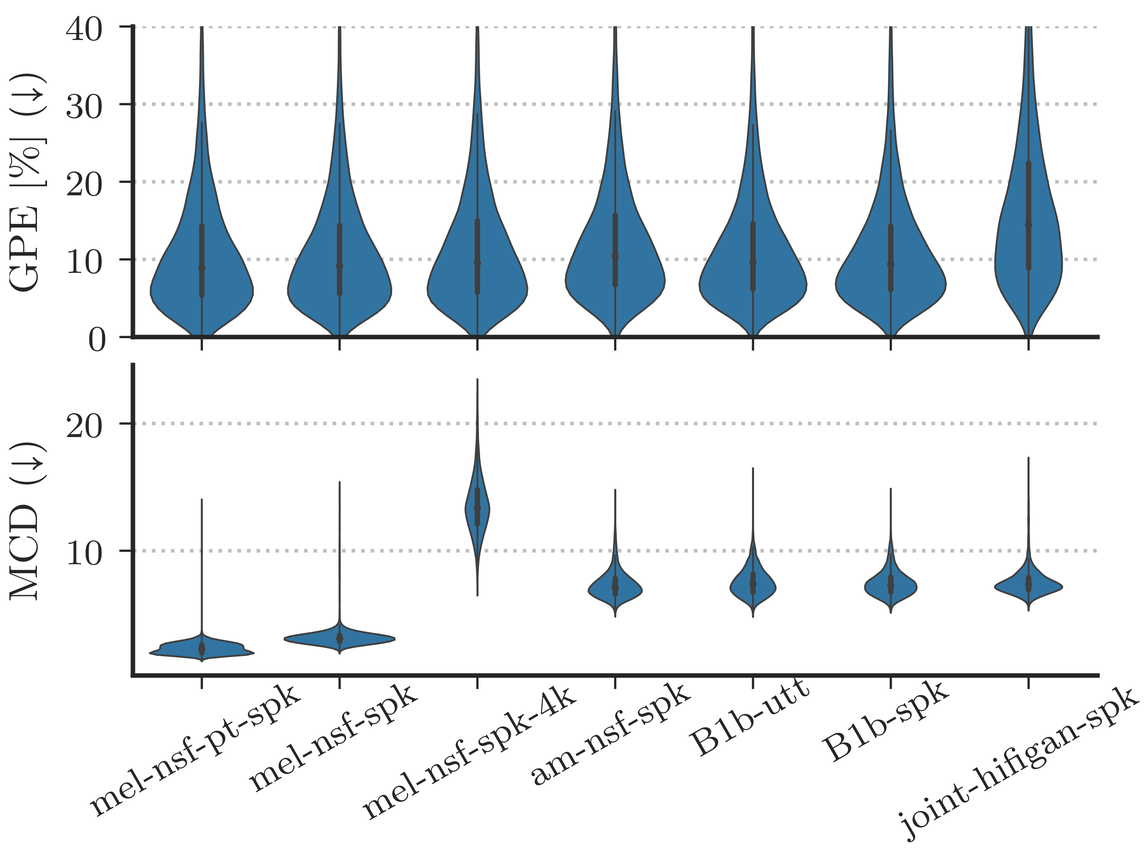
- Four objective, intrusive metrics
  - Mel-cepstral distortion (MCD): $\frac{6.1419}{T}\sum_{t=1}^{T}\sqrt{\sum_{i}(C_{t,i}-\hat{C}_{t,i})^2}$
  - Scale-invariant signal-to-noise ratio (SI-SNR): $\frac{|\alpha s|^2}{|\alpha s-\hat{s}|^2}$, for $\alpha=\frac{\hat{s}^T s}{\|s\|^2}$
  - Gross pitch error (GPE): $\frac{\text{num. of frames whose error} > 20\%}{\text{num. of correctly identified voiced frames}}$
  - Perceptual evaluation of speech quality (PESQ)
- A non-intrusive estimate of PESQ by torchaudio-SQUIM [2]
- Subjective evaluation with a MUSHRA-like listening test

3. Evaluated Systems

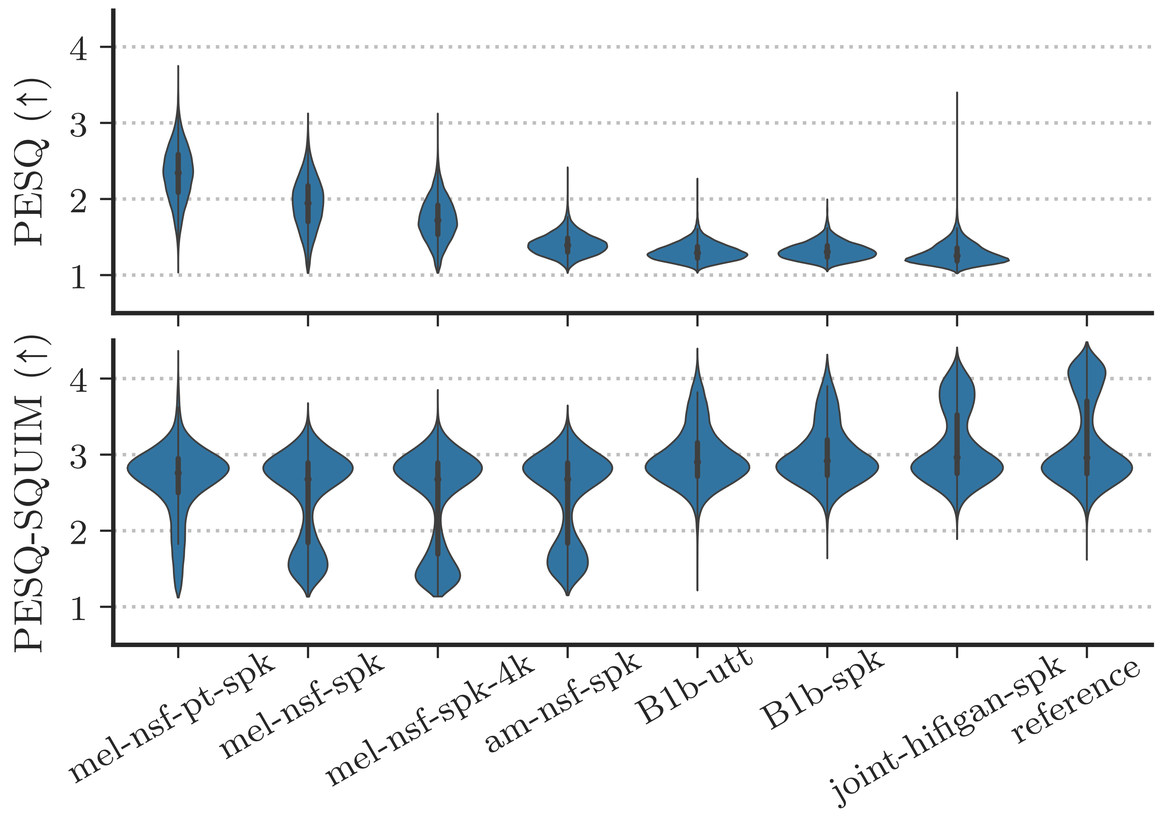
We bypass the anonymization block and resynthesize VoicePrivacy Challenge (VPC) test datasets [3] with the following B1 variants:

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

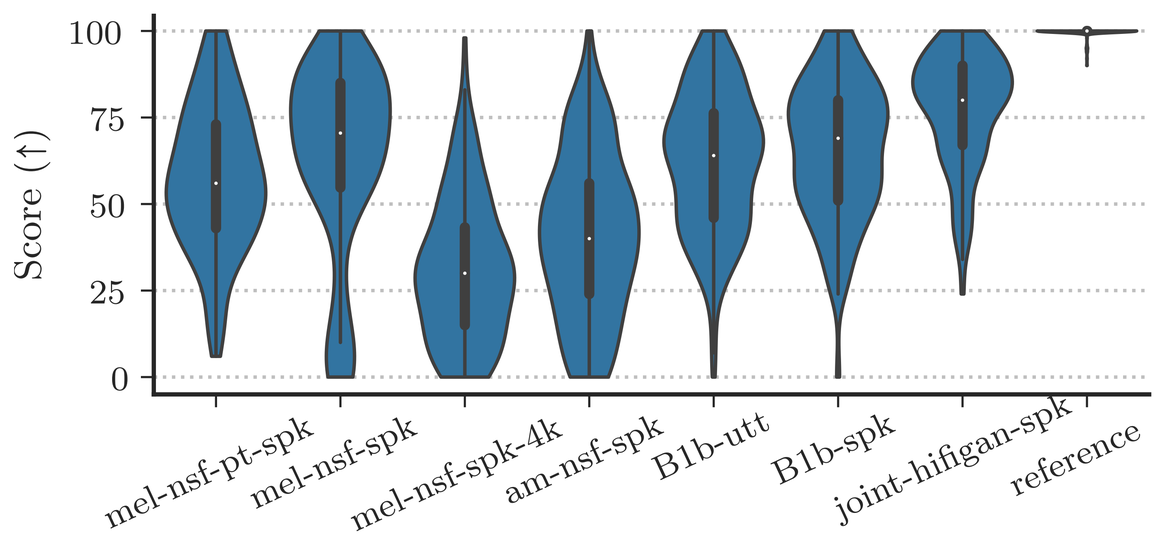
4. Objective Evaluation Results (a)



4. Objective Evaluation Results (b)



5. Subjective Evaluation ($n \geq 14$)



Subjects commented `joint-hifigan-nsf` "Americanizes" the voices, unable to retain speaker characteristics

6. Conclusions

- 2020 systems (`mel-nsf`, `am-nsf`) perform better w.r.t. intrusive measures we considered (GPE, MCD, PESQ)
- Performance gap between copy synthesis (`mel-nsf`) and other systems → utilized representation causes additional information loss.
- torchSQUIM-PESQ estimate came the closest to predicting subjective preferences, whereas intrusive metrics could not predict them
- PyTorch and C implementations of `am-nsf` are not equivalent

Possible improvements:

→ Additional incentive to help vocoders learn the speaker space

→ Trying an improved speaker embedding (e.g., ECAPA as done by [4])

References

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

[2] A. Kumar *et al.*, "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," in *Proc. ICASSP*, 2023.

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

[4] S. Meyer *et al.*, "Speaker anonymization with phonetic intermediate representations," in *Proc. Interspeech Conf.*, 2022.

Friedrich-Alexander-Universität Erlangen-Nürnberg

Fraunhofer IIS