Can we steal your vocal identity from the Internet?

Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data

Jaime Lorenzo-Trueba, Fuming Fang, Xin Wang, Isao Echizen (NII, Japan),
Junichi Yamagishi (NII, Japan / UoE, UK), Tomi Kinnunen (UEF, Finland)

Research questions

1. Can we construct TTS and VC attacks using *noisy found real audio data* (such as Youtube) instead of clean data?
2. Is speech enhancement useful?
3. Are such the attacks easy or difficult to detect?

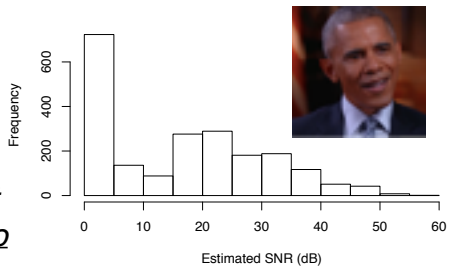
Found audio

3 hours of speech

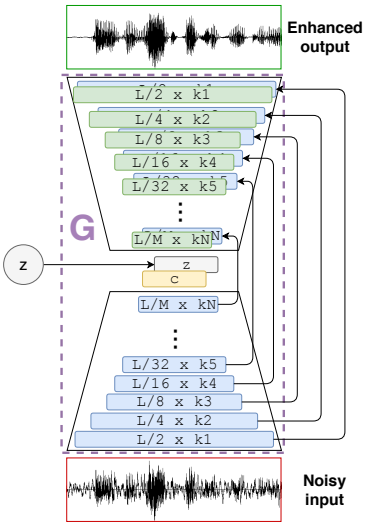
Found online

Interviews, public speeches etc

Noisy reverberant compressed audio



Speech enhancement GAN



1-D CNN based speech enhancement using raw speech waveforms

Least Square GAN criterion

G: *U-net* encoder-decoder architecture

D: Similar to encoder of G

Databases (28 speakers):

Output: VCTK corpus

Input: Noisy, Reverberant and/or Device-recorded versions of VCTK

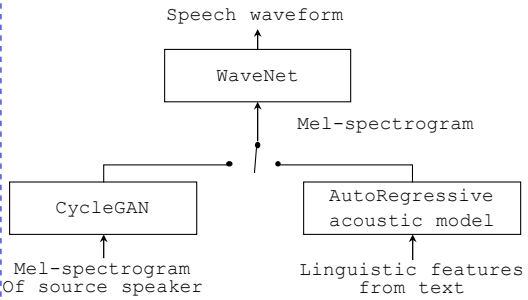
Result

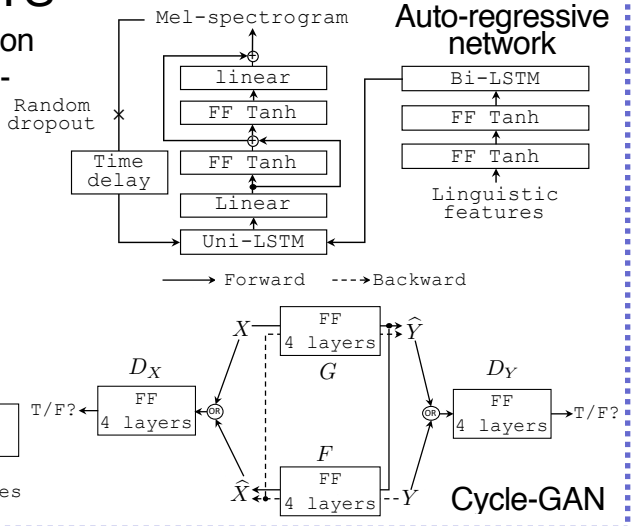
VCTK datasets	SNR (dB)	Subjective judgement (MOS) Quality	Cleanliness
-	17.2	3.58*	2.42
Noisy	49.8	2.73	3.35
Reverberant	22.7	3.55	3.17
Noisy, **R**everb.	43.1	3.11	3.42*
Device-recorded	28.2	3.51*	3.31
D+N	40.1	3.26	3.02
D+NR	41.4	3.30	3.34
All	37.9	3.41	3.40*

Crowd source listening test using 129 subjects. * mark indicates non-statistically significant differences. Other differences are significant.

Voice conversion & TTS

TTS and VC systems using a common Wavenet vocoder where 80-dim mel-spectrogram was estimated using either deep auto-regressive network (TTS case) or cycle-GAN (VC case)





	Quality	Similarity
Obama	4.40	4.70
Copy synthesis	2.45	2.99
VC1	2.66	**1.56**
VC2	2.67	**1.55**
VC3	2.83	**1.56**
TTS1	2.49	**1.43**
TTS2	2.51	**1.40**
TTS3	2.63	**1.45**

Number of subjects is 103. See a table below for configs for VC and TTS.

Anti-spoofing

Countermeasures

CQCC-GMMs

Trained using

- ASVspoof 2015
- VCC 2016 (used for benchmark of VCC 2018)

EER in percentage

	ASVspoof2015	VCC2016
Copy synthesis	4.63	8.46
VC1 (jpn2eng)	2.32	**1.09**
VC2 (eng2eng)	2.16	**0.00**
VC3 (mix2eng)	2.25	**1.01**
TTS1 (Noisy)	1.60	**0.00**
TTS2 (All)	2.01	**0.00**
TTS3 (All+reverb)	0.79	**0.00**

Conclusion

1. It is not easy to train TTS and VC attacks using noisy found real audio data at all!!
2. Speech enhancement is useful perceptually, but, increases detectable artifacts
3. The attacks using the noisy found data were easy to be detected by the countermeasures

All speech databases (clean, noisy, reverberant, device-recorded VCTK) are publicly available at Datashare.

Codes for SEGAN and TTS are also publicly available.