Stealing your vocal identity from the internet: cloning Obama's voice from found data using GAN and Wavenet

Jaime Lorenzo-Trueba[1,a)] Wang Xin[1] Junichi Yamagishi[1,2]

Abstract:
With the rise of machine-learning speech processing techniques, the risk of having your vocal identity stolen from online audio clips has risen considerably. In this research we study how GAN-based technologies can allow us to significantly improve the vocal quality of found audios that suffer from noise or reverberation, and the consequences that processing has from the speech generation point of view when using state-of-the-art Wavenet-based voice generation techniques.

Keywords: Speech privacy, found data, speech enhancement

1. Introduction

With the latest changes and improvements in machine learning technologies the problem that was once science-fiction is becoming more of a real nowadays problem: having your identity stolen. In that line we have seen many different approaches: generating bots that mimic human social interaction [1], tools that can generate photo-realistic videos of text-reading [2], [3], video-based impersonation [4], image-based impersonation [5] and even claims that your voice can be cloned with as little as one minute of audio [*1].

In the particular case of voice identity, it is easy to understand that there are large amounts of speech-content sources publicly available, and it is likely that data for almost anybody can be found one way or another. If we are talking about public personalities such as Obama (a common target for identity theft research [2], [3]), we can think that the amounts of public data available is immense.

Fortunately, such content is commonly recorded in non-professional acoustic environments such as homes and offices. Moreover, the recordings are often carried out using consumer devices such as smartphones, tablets, and laptops. Therefore, the speech recordings of the content are of typically poor quality and contain a large amount of ambient noise and room reverberation. But, real applications, such as speaker adaptation of speech synthesis or voice conversion, have been designed to work only on clean data of optimal acoustic quality and properties, and thus the quality of the systems trained with such found data is limited.

In this research we want to study the viability of training a speech enhancement system using purely publicly available data, and then using such tools for enhancing the speech quality of found data with the aims of, in the near future, training proper speech synthesizers or voice conversion systems using state-of-the-art tools.

Concretely, we want to answer two questions. First, we want to know what kind of data provides us with the best speech enhancement systems when using it on found data: should we use training data targeting the environmental conditions of the target speech? Or should we use as much data as possible? Second, we want to figure out if reducing the noisiness of the found data actually provides any perceptual benefit, as the spectral manipulations caused by the enhancement process might be inducing a drop in naturalness that would affect our future speech applications.

2. GAN-based speech enhancement

Generative adversarial networks consist of two "adversarial" models: a generative model G that captures the data distribution and a discriminative model D that estimates the probability that a sample came from the training data rather than G. This GAN structure has already been used successfully for the task of speech enhancement [6], [7], and for this research we work on an improved version of SEGAN [6] that attempts to make the training process more robust and stable by introduce a modified training strategy for SEGAN's generator.

2.1 Robust SEGAN training

For the considered robust training modification, we assume that we have some pre-trained baseline speech enhancement models (which may be simpler signal processing methods or easier-to-train neural networks than GANs), and that we can access to enhanced speech signals using the baseline speech enhancement models or methods. The modification, then, is to compute the content loss of the initial iterations of the generator model based on the baseline enhanced speech instead of on clean speech.

Additionally, a skip-connection was added around the gener-

[1] National Institute of Informatics, Tokyo, Japan
[2] The University of Edinburgh, Edinburgh, UK
[a)] jaime@nii.ac.jp

[*1] https://lyrebird.ai/

Table 1 Description of the different training speech enhancement corpora considered.

Corpus Name	Acronym	#Files	Total Time
VCTK	clean	11572	8h54min56s
Noisy VCTK	n	11572	8h54min56s
Reverberant VCTK	r	11572	8h54min56s
Noisy Reverberant VCTK	nr	11572	8h54min56s
Device Recorded VCTK	DR	11572	8h54min56s

ator so that the task of the generator module is not to generate enhanced speech from scratch but to generate a residual signal that refines the input speech [8]. With this, we expect to encourage the generator to learn the detailed differences between clean and enhanced speech waveforms.

A more detailed explanation of this improved process is soon to be published.

3. Speech Corpus

For this particular work we considered two kinds of speech corpus. First of all the corpus to train the speech enhancement module, which was selected from publicly available data so that the training process can be replicable. Secondly we extracted a number of different Obama's public intervention to use as a source of our cloned voice.

3.1 Corpus for speech enhancement

For the speech enhancement corpus we always relied on a subset (28 speakers 14 male and 14 female, all with British accent and around 400 utterances per speaker) of the centre for speech technology research (CSTR) VCTK corpus[*2][9] as the clean speech, and different noisy iterations to train the speech enhancement signal to face different possible noisy or reverberant environments (see table 3). All of the distorted corpora were recorded as a collaboration between CSTR and the National Institute of Informatics of Japan, and are publicly available in the DataShare repository of University of Edinburgh.

3.1.1 Device-recorded VCTK

The device-recorded (DR) VCTK[*3][10] consists on re-recording of the high-quality speech signals of the original VCTK by playing them bach and recording them in office environments using relatively inexpensive consumer devices. With this corpus we expect to be able to learn the nuanced relationships between high quality and device-recorded versions of the same audio.

For the re-recording, eight different microphones were used, and it was carried out in a medium-sized office under two background-noise conditions (i.e. windows either opened or closed). In total this resulted in 16 different conditions.

3.1.2 Noisy, Reverberant and Noisy and reverberant VCTK

We considered three other artificially corrupted variations of the CSTR VCTK corpus: Noisy VCTK[*4][11], Reverberant VCTK[*5][12] and Noisy and reverberant VCTK[*6][13]. Having such a diverse portfolio of possible speech corruptions has the aim of providing our speech enhancement systems with the capacity of learning to target the different possiblities, be it plain noisy, reverberation compensation or a mixture of both.

Table 2 Characterization of the used Obama's found data.

Sources	Public speeches, interviews, debates
Total length (including silences)	3h 7min 39s
Minimum segment duration	0.54s
Maximum segment duration	24.4s
Average segment duration	5.4s

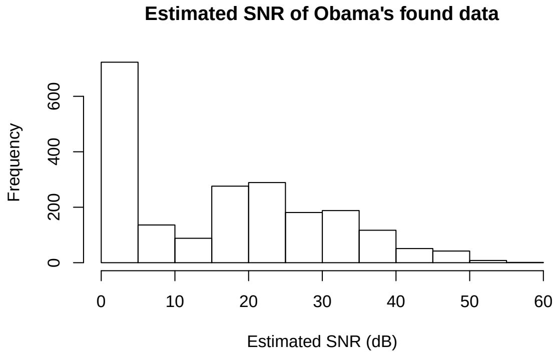

Fig. 1 Histogram of Obama's found data estimated SNR.

As mentioned beforehand, since all the datasets are based on the CSTR VCTK corpus, speakers and utterances of the Edinburgh noisy speech dataset are similar to those of the DR-VCTK dataset presented above.

3.2 Obama's found data

Obama's data was found online, mainly in YouTube videos with transcriptions as part of the description, from diverse sources such as interviews, political meetings, with very diverse recording conditions and environments, ranging from very noisy with large amounts of reverberation to not so noisy or not so reverberant samples, never achieving recording studio standards. The audio channel was split from the video, automatically segmented on long pauses, and down-sampled to 16kHz. The transcription was copied over as text files. Table 2 shows a brief characterization of the data.

The histogram of the SNR in dB estimated using NIST SNR tool [*7] can be seen in figure 1. There it is evident how the vast majority of the considered speech presented very low SNR when compared to conventional speech generation corpus standards.

4. Enhancing Obama's found data

The main aim of this research is to be able to train proper waveform generation models that can replicate a target speaker identity, in this case the very recognizable Obama's voice. For that, we only want to consider easily available low-quality resources, as introduced in section 3.2. But, that kind of training data is commonly known to be too poor to provide reasonably good speech synthesis systems. For that reason, we wanted to try applying the aforementioned speech enhancement technique to improve our low-quality found data to the point that it can allow us to train proper TTS systems. On top of that, as our proposed approach

[*2] http://datashare.is.ed.ac.uk/handle/10283/1942
[*3] https://datashare.is.ed.ac.uk/handle/10283/2959
[*4] https://datashare.is.ed.ac.uk/handle/10283/2791
[*5] https://datashare.is.ed.ac.uk/handle/10283/2031
[*6] https://datashare.is.ed.ac.uk/handle/10283/2826

[*7] https://www.nist.gov/information-technology-laboratory/iad/mig/nist-speech-signal-noise-ratio-measurements

Table 3 Description of the data sources of the different speech enhancement models. To understand the meaning of the sources acronyms please refer to table 1.

SOURCES	#Files	Total Time
DR	11572	8h54min56s
n	11572	8h54min56s
r	11572	8h54min56s
nr	11572	8h54min56s
DR+n	23144	17h49min52s
DR+nr	23144	17h49min52s
All (DR+n+r+nr)	46288	35h39min44s

is GAN-based, by manipulating the hidden variable at generation time, it allows us to obtain different enhanced representations of the same waveform, significantly boosting the amount of training data for the waveform generation systems.

4.1 Design of the speech enhancement models

As we had a large amount of free, publicly available resources for training our speech enhancement models, we wanted to study what would be the best training regime strategy. All in all, we trained 7 speech enhancement systems with the amounts and sources of training data seen in table 3.

The motivation for trying this large amount of possible configurations was clear. We expect each single category (without combination of data sources) to be specialized at enhancing their particular kind of disturbance (e.g. n should be good at cleaning noise, r at cleaning reverberation, DR at compensating for the low quality recording devices...). Then, it followed that, as most of the found data will come from noisy poor quality sources, it made sense to combine DR with the different noisy corpora. Finally, as it has been proven many times that having as much varied data as possible helps the neural networks generalize better, we also wanted to consider the mixture of all of our different corpora.

4.2 Training of the speech enhancement models

Similar to the original SEGAN training strategy, we extracted chunks of waveforms with a sliding window of 2^{14} samples at every 2^{13} samples (i.e. 50% overlap). At testing time, we concatenated the results at the end of the stream without overlapping. For the last chunk, instead of zero padding, we pre-padded it with the previous samples. For batch optimization, RMSprop with 0.0002 learning rate and batch size of 100 was used. The modified SEGAN model converged at 120 epochs.

For selecting the pre-enhancement method, our preliminary experiments, applying Postfish[14] and HRNR[15] sequentially showed better quality enhanced samples. We used this compound method to generate the baseline models explained in 2.1.

4.3 Objective evaluation of the speech enhancement

After training the speech enhancement models we proceeded to boost the noisy found data under consideration. Then, in order to quantify the impact of the enhancement process we estimated the SNR once again using NIST tool, and the results can be seen in table 4. It must be noted that SNR is most likely not the best way to measure the consequences of the enhancement process, but as we lack a clean reference we had limited availability of tools.

Table 4 Average SNR in dB estimated with NIST tool for the results of the different speech enhancement models.

SOURCES	average SNR (dB)
Obama source	17.2
n	49.8
r	22.7
nr	43.1
DR	28.24
DR+n	40.1
DR+nr	41.37
all (DR+n+r+nr)	37.89

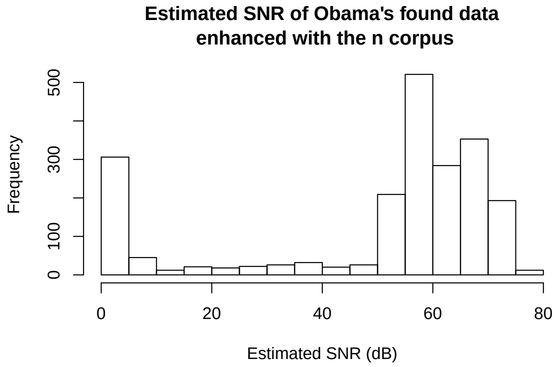

Fig. 2 Histogram of Obama's found data estimated SNR after enhancing with noisy VCTK.

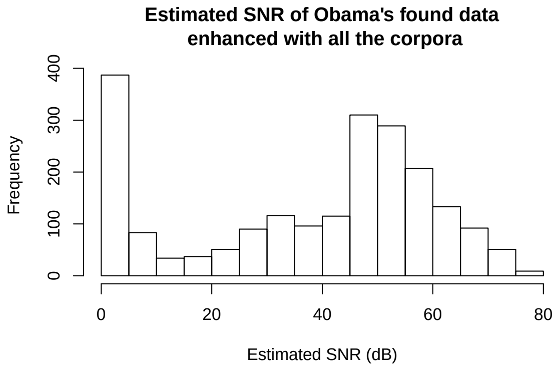

Fig. 3 Histogram of Obama's found data estimated SNR after enhancing with all the VCTK variants.

The SNR estimation results show us a clear picture: the enhancement process, regardless of which training data was used for the model, improved the average SNR of Obama's data. Not only that, we can see how the speech enhancement models trained with noisy data in particular (i.e. n, nr and its mixtures) perform considerably better than the other two possiblities (i.e. r and DR). This most likely has to do with the fact that they will reduce the noise levels in the signal, which is what the SNR measure targets, and not with an actual increase in perceptual quality of the voice. We can see the SNR histograms of the output of enhancing with n and all in figures 2 and 3 respectively.

4.4 Perceptual evaluation of the speech enhancement

As mentioned in the objective results (section 4.3), SNR estimation is most likely not the best way to measure the improvements of our approach. And, as the final objective of our research is to produce high quality synthetic speech, be it through speech synthesis or through voice conversion, it makes sense to evaluate

Table 5 Results of the perceptual evaluation in MOS score. Non statistically-significant differences are marked with *.

SOURCES	Quality	Cleanliness
Obama source	3.58*	2.42
n	2.73	3.35
r	3.55*	3.17
nr	3.11	3.42*
DR	3.51	3.31
n+DR	3.26	3.02
nr+DR	3.30	3.34
all (n+r+nr+DR)	3.41	3.40*

perceptual quality from the point of view of human users.

In that line, we carried out a crowd-sourced perceptual evaluation with Japanese native listeners. In this evaluation we presented the listeners with a set of 16 screens, which each corresponding to 1 of the 8 evaluated conditions (original plus the 7 enhanced versions) times 2 utterances. The evaluators were asked two question: first to rate their perceived speech quality of the samples in MOS scale, and second to rate how clean of noise, reverberation or artifacts the speech samples were, also in MOS scale (with 1 being very noisy and 5 being clean speech).

They participants were able to listen to the sample in each screen as many times as they wanted, but they could not proceed to the next sample until they answered both questions, without being allowed to go back. The test were selected based on their length, evaluating in the all the utterances between 5.0 and 5.9 seconds long (i.e. 530 utterances). In total this meant that 265 sets for evaluating all the evaluation utterances, which was done 3 times over for a total of 795 sets. They participants could repeat the task up to 8 times to guarantee that we collected at least 100 different listeners. In the end a total of 129 listeners took part in the evaluation (72 male, 57 female).

4.4.1 Results

The results of the perceptual evaluation can be seen in table 5, and they too show a clear picture. In the case of Obama's found data there is a clear perception of noisiness and related factors (2.42), even if the perceived quality is reasonably high (3.58). It is also noteworthy to say that this is about 1 point less of the average quality MOS of clean natural speech, most likely due to the poor conditions on which these sources were recorded.

In the case of the enhanced versions, we can see how we managed to consistently improve the cleanliness of the source data, with different degrees of success depending on which source data was used. Most noteworthy is the results of the "noisy-reverberant" model (3.42), which provided the biggest improvements in cleanliness. We can assume this to be because the found recordings present in general both noise and reverberation, so a speech enhancement system targeting this condition gives the best improvement in that field. Similarly happens to the "all" model, which we assume performed comparably because if was trained in all possible situations.

On the other hand, we can also see how there is a cost to applying speech enhancement techniques, as there is a consistent degradation in the perceived quality of speech. This means that speech enhancement, in the pursuit of cleanliness, produces a considerable drop in the naturalness in the outputted speech. This could potentially mean that the approaches that provided the biggest improvements in SNR, such as the "noisy" model with a quality rating of 2.73 or the "noisy-reverberant" model with a quality rating od 3.11 could not be the best way to producing clean speech for further speech processing.

In the end there seemed to be a trade-off between quality degradation and cleanliness improvements, which were not encouraging. But, if we look at the results of the "all" model, combining all possible sources of data, we see that it was capable of providing one of the best cleanliness scores (3.40) with one of the smallest quality degradation (0.17 degradation). This strongly hints that having trained our speech enhancement system in a variety of degradation conditions gave the system enough generalization capability and enough knowledge of human speech to reduce the noisiness while maintaining as far as possible voice naturalness.

5. Building the Wavenet vocoder

Building a state-of-the-art data-driven vocoder such as Wavenet places the first challenge when trying to use this found data: it is not easy to gather large amounts of good enough data for the process. And this is were the advantage of having used another data-driven speech enhancement system comes into play. As hinted in the introduction of section 4, we can take advantage of the generation process of our GAN-based speech enhancement system to generate multiple versions of the enhanced speech of the noisy data, effectively multiplying the amounts of training data available for our system.

For this particular research we are currently training our own Wavenet vocoder based on the enhanced Obama's speech. The Wavenet vocoder works at a sampling rate of 16kHz. The μ-law companded waveform is quantized into 10 bits per sample. Similar to the literature [16], the network consists of a linear projection input layer, 40 blocks for dilated convolution, and a post-processing block. The k-th dilation block has a dilation size of $2^{\mathrm{mod}(k,10)}$, where $\mathrm{mod}(\cdot)$ is modulo operation.

The acoustic features, which are fed to every dilated convolution block, contain the 80-dimensional mel-spectrogram plus 1 additional component specifying which of the different speech enhancing models had produced that speech waveform. The choice of mel-spectrogram as the main acoustic feature was determined by the expected limitations of traditional features (e.g. F0, Mel-generalized Cepstrum and aperiodicity bands), as the estimation of F0 is very problematic in both the original noisy speech signals and the enhanced signals. We also considered an increased number of mel bands compared to other approaches [17] (80 vs. 60) with the expectation that it will help the vocoder cope better with the corrupted or noisy segments.

The tools for this implementation are based on a modified CURRENNT toolkit [18], and can be found online too [*8].

6. Conclusions and future work

We have introduced a number of publicly available and known data-sets that proved to be extremely useful for training potent speech enhancement models. These models were applied to a corpus low quality, and considerably degraded found data (with

[*8] http://tonywangx.github.io

Obama's identity), and were capable of very significantly improving the SNR of the data. Not only that, but after carrying out a perceptual evaluation, we also saw how the obtained models can also significantly improve the perceptual cleanliness of the source speech without significantly degrading the naturalness of the voice as is common after applying speech enhancement techniques. This worked best when the speech enhancement system was trained using the largest amount of data available, thus covering a large amount of environmental and recording conditions, improving the generalization capabilities of the system.

The training of the Wavenet vocoder is still ongoing work, as is the training of the network for predicting the mel-spectrogram from text to build a complete text to speech (TTS) synthesis system and a competing CycleGAN-based voice conversion (VC) system. We also plan to do an evaluation comparing the Obama's impersonation capabilities of TTS vs. VC using the trained Wavenet vocoder in the future.

References

[1] Adams, T.: AI-Powered Social Bots, *arXiv preprint arXiv:1706.05143* (2017).

[2] Kumar, R., Sotelo, J., Kumar, K., de Brebisson, A. and Bengio, Y.: ObamaNet: Photo-realistic lip-sync from text, *arXiv preprint arXiv:1801.01442* (2017).

[3] Suwajanakorn, S., Seitz, S. M. and Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio, *ACM Transactions on Graphics (TOG)*, Vol. 36, No. 4, p. 95 (2017).

[4] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C. and Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2387–2395 (2016).

[5] Karras, T., Aila, T., Laine, S. and Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation, *arXiv preprint arXiv:1710.10196* (2017).

[6] Pascual, S., Bonafonte, A. and Serrà, J.: SEGAN: Speech Enhancement Generative Adversarial Network, *CoRR*, Vol. abs/1703.09452 (2017).

[7] Donahue, C., Li, B. and Prabhavalkar, R.: Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition, *arXiv preprint arXiv:1711.05747* (2017).

[8] Kaneko, T., Kameoka, H., Hojo, N., Ijima, Y., Hiramatsu, K. and Kashino, K.: Generative adversarial network-based postfilter for statistical parametric speech synthesis, *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2017)*, pp. 4910–4914 (2017).

[9] Veaux, C., Yamagishi, J. and King, S.: The voice bank corpus: Design, collection and data analysis of a large regional accent speech database, *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*, IEEE, pp. 1–4 (2013).

[10] Sarfjoo, S. S., Yamagishi, J. et al.: Device Recorded VCTK (Small subset version) (2017).

[11] Valentini-Botinhao, C. et al.: Noisy speech database for training speech enhancement algorithms and TTS models (2017).

[12] Valentini-Botinhao, C. et al.: Reverberant speech database for training speech dereverberation algorithms and TTS models (2016).

[13] Valentini-Botinhao, C. et al.: Noisy reverberant speech database for training speech enhancement algorithms and TTS models (2017).

[14] Montgomery, M.: Postfish by Xiph.org, Available: https://svn.xiph.org/trunk/postfish/README (2005).

[15] Plapous, C., Marro, C. and Scalart, P.: Improved signal-to-noise ratio estimation for speech enhancement, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 6, pp. 2098–2108 (2006).

[16] Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K. and Toda, T.: Speaker-Dependent WaveNet Vocoder, *Proc. Interspeech*, pp. 1118–1122 (online), DOI: 10.21437/Interspeech.2017-314 (2017).

[17] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. et al.: Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions, *arXiv preprint arXiv:1712.05884* (2017).

[18] Weninger, F., Bergmann, J. and Schuller, B.: Introducing CURRENNT–the Munich open-source CUDA RecurREnt neural network toolkit, *Journal of Machine Learning Research*, Vol. 16, No. 3, pp. 547–551 (2015).