

ILST Take-Home Exam 2 — Speech Processing for Data Protection: Speaker Anonymization

Daan Brugmans¹

¹S1080742

daan.brugmans@ru.nl

Abstract

1000 characters maximum. ASCII characters only. No citations. Provide a concise summary of your project, including key objectives and findings.

Index Terms: *speaker anonymization, voice privacy, private speech, adversarial anonymization*

1. Introduction

The popularity of Automatic Speech Recognition (ASR) and Automatic Speaker Verification (ASV) technologies among the general public continues to rise. As the usage of such state-of-the-art (SOTA) speech technologies continues to rise, there has developed an increased need for the protection of speech data. This is because speech is a biomarker: aside from the linguistic contents, an utterance contains information about the speaker, such as their sex and age category. With governmental bodies implementing consumer data protection laws, such as the EU's GDPR, and speech data containing information that may be considered sensitive, researchers have turned towards developing new methods for protecting the speech data of consumers.

One type of speech data protection is called *speaker anonymization*, where an utterance's speaker is made anonymous by obfuscating, removing or replacing the information in the utterance pertaining to the speaker, while keeping the linguistic information intact. In this review, I will briefly discuss the current field of speaker anonymization research, and touch on some state-of-the-art developments. For a more thorough review of the field, I recommend the work of Champion [1], which has guided me in my research. I will attempt to answer the following research question:

What kinds of approaches to speaker anonymization are currently researched in state-of-the-art research?

2. Method

2.1. Search Procedure

I mainly used the Semantic Scholar database for finding modern research papers. With Semantic Scholar, I searched for papers using the keyword "speaker anonymization" published from 2019 onwards. I read some of the most cited papers from that search result, then snowballed into reading other papers that those papers cited. In total, I considered 19 candidate papers for this review, of which I chose 8 to include in the final draft (including Champion's [1] paper).

While selecting which papers to include, I focused on distilling the selection of papers to one that showcases all types of approaches to speaker anonymization. Every type of speaker

anonymization approach is represented by an amount of papers roughly corresponding to how popular that type of approach is in state-of-the-art research, and the papers are all commonly cited and established methods for their type of approach. For the two newer types of approach, I also included a paper published that isn't as well known, but showcases an approach that is relatively new.

2.2. Models

All speech-based models in the papers are neural networks, as are almost all speaker anonymization approaches. For the latter, both supervised and unsupervised networks are used. Supervised speaker anonymization networks are more common in older papers, and include DNNs and U-Nets. Unsupervised networks seem to be popular in newer works, with (V)AE's and GANs often used, as well as variations of them, such as the CycleGAN-VAE and HiFiGAN; Flows are used only in the newest paper.

2.3. Datasets

There is a small set of speech datasets of which a subset is almost always included across all papers. These include LibriSpeech, VoxCeleb, and VCTK. LibriSpeech and VCTK are datasets of English speech, with VCTK including different accents, and VoxCeleb is an audio-visual dataset of multilingual speech. Other datasets are also used, such as TIMIT (English), VCC 2016, AISHELL (Mandarin) and CommonVoice (Italian and French).

2.4. Metrics

Speaker anonymization systems are often tested by measuring changes in ASR and ASV models. When measuring the ASR performance of the system, the Word Error Rate (WER) is used. When measuring the ASV performance of the system, the Equal Error Rate (EER) is a popular metric, and is a balanced metric of the true positive rate and false positive rate of the ASV system's predictions. Individual papers also use specific metrics relevant for their research.

3. Results

Current state-of-the-art research on speaker anonymization can be split up into three types of approaches: signal processing approaches, voice conversion approaches, and adversarial attack approaches.

3.1. Signal Processing Approaches

In signal processing-based speaker anonymization, an utterance is anonymized by directly altering the speech with some effect. This is the most classic approach to anonymizing speakers, and is currently the most commonly used in practice, for example when anonymizing interviewees. The main advantage of this approach is that it is resource inexpensive and easy to apply: no datasets have to be collected for training neural networks, which don't have to be run on dedicated hardware. However, this method of speaker anonymization is also the least effective, as the effects applied to the speech signal often significantly worsen the utterance's intelligibility and can be relatively easy to reverse, de-anonymizing the speech. As a result, research on this type of speaker anonymization is limited.

The work by Patino et al. [2] is a popular modern example of signal processing-based speech anonymization techniques. In this paper, the authors use the McAdams coefficient for anonymizing utterances, which is a variable used in a popular formula for adding timbre to speech. This approach to speaker anonymization has the benefit of not requiring any computationally expensive techniques or a wealth of data. The authors found that changing the McAdams coefficient is an effective way of greatly improving speaker anonymity while preventing big losses in speech accuracy.

3.2. Voice Conversion Approaches

Voice Conversion is a more modern approach to speaker anonymization. Voice Conversion (VC) is the task of converting the voice of an utterance from the source speaker to the target speaker by changing the characteristics of the source speaker's voice. A VC-based speaker anonymization approach, then, requires that the information pertaining to the speaker in an utterance can not only be extracted, but also independently from the linguistic information in the utterance. By only changing the information pertaining to the speaker, without altering linguistic information, the speaker of an utterance can be changed. By changing the speaker from the source speaker to a target speaker, the speech is essentially anonymized, as it cannot be traced back to the original speaker.

The work by Fang et al. [3] is a foundational example of modern speaker anonymization techniques through the changing of X-vectors. By training a neural net on an ASV task (learning to recognize a speaker from a speech input), the network learns to internally develop a representation of the speaker identity, which it stores in one of its layers. By extracting the embedding at this layer for a particular speaker's speech input, which is called the X-vector, one has a vector representation of a speaker's identity. In their paper, Fang et al. perform speaker anonymization by taking the average X-vector from a speaker's speech utterances, comparing it to an existing database of other X-vectors, from which a new average X-vector is calculated to create a "new" speaker identity, which is then used in an audio generator. By separating the linguistic contents of a speech utterance from the speaker information (using the X-vector), and supplying a different X-vector alongside the original linguistic information of the speech, a new speech utterance can be generated that retains the original message's linguistic contents using a "new", anonymous, speaker.

Yoo et al. [4] present another example of speaker anonymization using Voice Conversion techniques. The authors train a CycleVAE-GAN model that learns to generate an utterance that is linguistically the same as the input but uses a different speaker identity, expressed using an X-vector. The authors

try various methods for generating such speaker identity vectors to optimize for both anonymity and accuracy. They also train a DNN and a GMM as speaker identification models. Their results are mixed, showing that the balance between speaker anonymity and speech intelligibility seems to be a trade-off.

Meyer et al. [5] also propose a GAN-based approach. A speaker embedding (that includes an X-vector) is extracted from the source audio and is fed to a GAN, which generates a new speaker embedding. By using this speaker embedding in the TTS component, the speech has become anonymized. The authors find that their architecture has improved WER and EER than comparable state-of-the-art models.

3.3. Adversarial Attack Approaches

Adversarial attack-based speaker anonymization is the newest type of anonymization technique, and is currently actively researched. Adversarial attacks are attacks on neural networks that create and use adversarial samples that alter the network's behavior. When most neural networks learn from data, they not only learn behavior that is both based on concrete information present in the data and helps them to fulfill the learning objective, but also behavior that doesn't use concrete information present in the data, yet still helps the network to fulfill the learning objective anyway. Adversarial samples exploit a network's "irrelevant" learned behaviors. Because these behaviors are not based on concrete information present in the data, adversarial examples can change the overall behavior of neural networks without degrading the network's performance on both clean and adversarial samples. Adversarial attack-based speaker anonymization approaches, then, use this property of neural networks to generate speech utterances whose speaker cannot be identified by the neural network. The main advantage of this approach is that an utterance can be anonymous to a neural network, whose behavior is exploited such that it cannot recognize the correct speaker, while the utterance sounds nearly identical to the original to the human ear.

Chen et al. [6] use the pretrained state-of-the-art TTS model YourTTS to generate adversarial speech samples using the Fast Gradient Sign Method (FGSM), a popular method for creating adversarial samples in any DNN. By applying the FGSM attack iteratively or in a one-shot fashion, an adversarial version of the original speech is generated where perturbations are introduced. While these perturbations are barely susceptible to the human ear, they fool neural networks into behaving abnormally, preventing neural ASV systems from performing speaker verification properly. The authors' method then works as a speaker anonymization method for neural networks.

Deng et al. [7] present V-Cloak, an example of one-shot adversarial speaker anonymization. The authors of this paper train a custom Wave-U-Net to generate adversarial speech samples. This is a one-shot approach that contrasts the common iterative approach to generating adversarial samples. V-Cloak is trained on 3 respective loss functions that measure speaker anonymity, intelligibility, and naturalness. The authors find that their model outperforms comparable state-of-the-art speaker anonymization models for both English and Chinese speech in both anonymity and intelligibility. A user study conducted by the authors also found that V-Cloak's naturalness was generally best among state-of-the-art models.

O'Reilly et al. [8] propose VoiceBlock, an Encoder-Decoder-based system that is trained to anonymize user speech on the fly. Their model's encoder extracts various features about a speaker that are passed to the bottleneck layer, the output of

which is passed to a decoder, which filters the input. The output of this decoder contains adversarial perturbations that prevent neural ASV systems from recognizing speakers correctly while maintaining intelligibility to human listeners. VoiceBlock is able to process one chunk of audio in 0.25 seconds or less on a desktop CPU and is designed to be used in real-time applications, such as telephone communications.

4. Discussion

The trend of using machine learning techniques, especially neural network-based methods, in state-of-the-art research on speaker anonymization approaches seems to align with the ever-rising popularity of machine learning approaches in the field of speech technology in general.

Research on signal processing techniques for speaker anonymization seems to have become less popular as the popularity of neural networks has started to rise. There exists more research on signal processing for speaker anonymization prior to 2019, a period in which state-of-the-art neural architectures such as the transformer, conformer, GAN, VAE, and Flow were still developing or didn't exist yet. As in the past few years these neural architectures have led to breakthroughs throughout the field of speech technology as a whole, proving to provide a great wealth of new opportunities, research on speaker anonymization has also shifted towards the creation of neural solutions. This may be opposed to the opportunities that lie in new signal processing-based research, which may be lacking, since they used to be more heavily researched in the past, and amount of approaches possible seems much more limited than in neural approaches. Although signal processing-based approaches have been proven to increase anonymity of speech, this often goes hand-in-hand with a noticeable degradation in intelligibility and naturalness of speech, which is often distorted or even human-like. Patino et al.'s [2] work can be considered a signal processing technique with a relatively low loss in intelligibility. This is in contrast to neural methods, especially adversarial attack-based methods, which are able to maintain both intelligibility to both ASR models and human listeners as well as naturalness to the human ear. For these reasons, I expect that we will not see many signal processing-based approaches to speaker anonymization put into practice in the near future.

The neural network-based approaches to speaker anonymization, voice conversion and adversarial attacks, continue to develop and have seen new advancements in the past few years of research.

For voice conversion-based approaches, the introduction of the X-vector has not only proven that a speaker's identity and vocal characteristics can be extracted from an utterance, but has also given rise to the idea that different characteristics of a speaker's voice can be independently extracted, changed, and anonymized, a process known as disentanglement. X-vectors have since become a popular method of anonymizing speakers, whether that is by directly changing or generating an X-vector and generating speech using the new X-vector (Fang et al. [3], Yoo et al. [4]), or by incorporating the X-vector into an architecture that generates anonymous speech (Meyer et al. [5]). The disentanglement property of utterances has been further researched in the past years, resulting in new options for altering specific characteristics of a speaker's voice, such as flipping the perceived sex of the voice or making it non-binary, or changing the perceived age category of the speaker. Outside the field of speaker anonymization, neural voice conversion methods have proven increasingly popular among the general pub-

lic, and are often used for purposes of entertainment or even the spreading of misinformation. The familiarity of voice conversion amongst the general public may be a reason to expect that voice conversion-based speaker anonymization may be adopted in practice more easily by the general public. Additionally, I expect that the relevance of voice conversion-based approaches will remain for now, despite advances in adversarial attack-based approaches. This is because the adversarial attack-based approaches' main advantage, the ability to anonymize speakers for neural (ASV) systems only, also means that they aren't viable approaches for anonymizing speech that is meant to be heard by a human listener.

Adversarial attack-based approaches have proven to be a valuable development in the protection of speaker identity when utterances are processed digitally. The fact that neural networks have the property of learning "irrelevant" behavior that can be exploited to change overall model behavior has made it possible to anonymize speakers in a way that prevents systems from automatically extracting and collecting speaker identities from speech data, while ensuring that the original speaker is still easily recognized by human listeners. I argue, then, that a system of this type of approach to speaker anonymization has the most potential to be implemented as a way of protecting the data privacy of consumers: for example, they can be embedded into consumer products such as phones and automatically anonymize any utterance that a consumer records. However, then, it would be vital to ensure that the chosen method can be run swiftly on mid-range consumer hardware. Existing work such as the work done by O'Reilly et al. [8] shows that such a method may be feasible to implement.

I expect that for both neural approaches to speaker anonymization, further developments will be made as the state-of-the-art of speech technologies continue to develop, and that those new SOTA models will be incorporated into new speaker anonymization models. One such example is the work by Chen et al. [6], who use a pretrained state-of-the-art Text-to-Speech model to generate adversarial samples. Much work done since 2019 has focused on both predictive and generative neural architectures that have been longer established, such as DNNs, CNNs, RNNs, GANs, and VAEs. I expect that in the coming years, we can expect new research on neural speaker anonymization approaches that incorporate neural architectures that were developed more recently, such as transformers, conformers, flow-based models, and diffusion-based models, and make use of state-of-the-art neural models, such as the multilingual Whisper ASR model to test anonymization performance across languages and TTS models such as YourTTS, as is used by Chen et al. [6].

5. Conclusion

- Summarize the key insights from your chosen topic.
- Highlight any gaps in current research and suggest avenues for future exploration.
- Personal thoughts about the course and the activities done in relation with the project are welcomed.
- No references are needed in this section.

6. References

- [1] P. Champion, "Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques," phdthesis, Université de Lorraine, Apr. 2023. [Online]. Available: <https://hal.univ-lorraine.fr/tel-04218098>

- [2] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker Anonymisation Using the McAdams Coefficient," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 1099–1103. [Online]. Available: https://www.isca-archive.org/interspeech_2021/patino21_interspeech.html
- [3] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker Anonymization Using X-vector and Neural Waveform Models," in *10th ISCA Workshop on Speech Synthesis (SSW 10)*. ISCA, Sep. 2019, pp. 155–160. [Online]. Available: https://www.isca-archive.org/ssw_2019/fang19_ssw.html
- [4] I.-C. Yoo, K. Lee, S. Leem, H. Oh, B. Ko, and D. Yook, "Speaker Anonymization for Personal Information Protection Using Voice Conversion Techniques," *IEEE Access*, vol. 8, pp. 198 637–198 645, 2020, conference Name: IEEE Access. [Online]. Available: <https://ieeexplore.ieee.org/document/9247219>
- [5] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, "Anonymizing Speech with Generative Adversarial Networks to Preserve Speaker Privacy," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 912–919. [Online]. Available: <https://ieeexplore.ieee.org/document/10022601>
- [6] S. Chen, L. Chen, J. Zhang, K. Lee, Z. Ling, and L. Dai, "Adversarial Speech for Voice Privacy Protection from Personalized Speech Generation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 11 411–11 415, iSSN: 2379-190X. [Online]. Available: <https://ieeexplore.ieee.org/document/10447699>
- [7] J. Deng, F. Teng, Y. Chen, X. Chen, Z. Wang, and W. Xu, "V-Cloak: Intelligibility-, Naturalness- & Timbre-Preserving Real-Time Voice Anonymization," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 5181–5198. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity23/presentation/deng-jiangyi-v-cloak>
- [8] P. O'Reilly, A. Bugler, K. Bhandari, M. Morrison, and B. Pardo, "VoiceBlock: Privacy through Real-Time Adversarial Attacks with Audio-to-Audio Models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 30 058–30 070, Dec. 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/c204d12afa0175285e5aac65188808b4-Abstract-Conference.html