Why Eli Roth should not use TTS-Systems for anonymization

Yamini Sinha\textsuperscript{1}, Jan Hintz\textsuperscript{1}, Matthias Busch\textsuperscript{1}, Tim Polzehl\textsuperscript{2}, Matthias Haase\textsuperscript{3}, Andreas Wendemuth\textsuperscript{4}, Ingo Siegert\textsuperscript{1}

\textsuperscript{1}Mobile Dialog Systems, Institute for Information Technology and Communications, Otto von Guericke University Magdeburg, Germany
\textsuperscript{2}Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI)
\textsuperscript{3}Human-Machine-Interaction, University of Applied Science Magdeburg-Stendal
\textsuperscript{4}Cognitive Systems Group, Institute for Information Technology and Communications, Otto von Guericke University Magdeburg, Germany

\{yamini.sinha, jan.hintz, matthias.busch, andreas.wendemuth, ingo.siegert\}@ovgu.de, tim.polzehl@dfki.de, matthias.haase@h2.de

Abstract

This paper evaluates the impact of using TTS-based speaker anonymization with objective and subjective methods. A pre-trained automatic speaker verification (ASV) VGGVox model (95.66\% recognition rate on Voxceleb 1), enrolled with human voices, is tested on the anonymized voices obtained from eSpeak TTS, to objectively verify the anonymization. We used one of the benchmark datasets for a speaker verification task with 1,251 speakers and over 100,000 utterances, consisting of spontaneous speech called VoxCeleb1. Upon anonymizing 40 speakers from the VoxCeleb1 test dataset, the objective evaluation shows that ASV systems, if presented with synthetic speech samples, are vulnerable to false acceptance. Experimental results show that after anonymization, approximately 6\% of the TTS speaker samples were falsely accepted as the counterfeited human speaker. This confusion about a TTS speaker as a human speaker may lie in the accuracy of the ASV model and the similarity metric used. Furthermore, we examined these confused speaker pairs against non-confused speaker pairs using a subjective measure (listener’s ratings) with 200 test subjects. In the subjective evaluations using a crowd-sourced platform, no significant results could be concluded, as human raters were unsure whether or not the voices were similar.

Index Terms: Speaker anonymization, Text-to-Speech, Speaker verification

1. Introduction

Perceived privacy is key in the use of digital voice assistants. Compared to passwords or biometric features like fingerprints or facial recognition, this is important because voice exhibits several features at a time and has a very dominating role in everyday life [1]. Trust in voice assistant technology and service providers is fundamental, especially for non-users [2, 3, 4]. Recently, many studies on voice anonymization have been initiated (e.g., Prajapati et al. [5]). Especially regarding increasing user acceptance and trust, anonymization represents a promising approach. However, there is a lack of studies that empirically investigate the influence of anonymization on acceptance and trust. Research on anonymization services on the Internet has shown that they have a significant influence on user behavior. Users’ privacy concerns have a significant impact on the extent to which they trust privacy technologies [6]. Users must trust the anonymization process and the provider to protect their voice. If users do not trust the system, they might not use a voice assistant in the worst case.

Using a Text-to-Speech (TTS) system to hide a speaker’s identity sounds like a naïve approach to this challenging problem at first. One would expect that a computer-generated voice that is not targeted at a certain speaker should possibly be matched correctly, however during previous experiments [7, 8], we still noticed a high false acceptance rate, falsely identifying the TTS-generated voice as some human speakers from the database.

The aim of this paper is to generate hypotheses that will be further evaluated regarding the confusion of TTS and human voices for speaker identification. This is not only helpful for the speaker verification and authentication research (to develop countermeasure methods) but also for the speaker anonymization utilizing the speaker conversion approach (selecting a proper target speaker).

2. Related Work

There are different approaches to speaker anonymization. These rely on either voice modulation (e.g., McAdams coefficient [9], vocal-track length transformation [10]) or voice conversion (e.g., x-vector based [11]). Both methods focus on the alteration of non-linguistic information while preserving linguistic information of the speech. The voice modulation methods alter the voice on parameters such as pitch and tone, whereas, voice conversion changes the voice-id of a speaker \(x\) to the voice-id of a target speaker \(y\). For example, to anonymize a speaker, the x-vector embedding can be swapped and set as the new target speaker. The voice is then synthesized based on the target x-vector using a waveform model [12].

In this paper, we focus on using an alternative approach, by combining Automatic Speech Recognition (ASR) and Text-to-Speech (TTS), i.e. the text of the spoken utterance of the source speaker is recognized first, and then synthesized to fit the characteristics of a certain target speaker. By this approach, it can be assumed that under all circumstances a speaker-anonymization can be achieved, as the original speaker and TTS speaker do not share the same speaker characteristics (as they are true independent speakers). Especially due to the current advancements in ASR and TTS systems, with ASR reaching human parity [13] and natural-sounding TTS voices are just around the corner where even actual TTS systems deceive human listeners in daily interactions [14].

One of the key aspects of anonymization is verifying if both the anonymization of the speaker and the preservation of the con-
tent was successful. This can be achieved through objective and subjective methods. In terms of objective methods, Automatic speaker verification (ASV) and automatic speech recognition (ASR) systems can be used to evaluate if the speaker’s identity was successfully hidden while the content (and intelligibility) remain preserved [15]. On the other hand, a subjective evaluation can be obtained through the listening tests with human assessors, rating on content, speaker identity and intelligibility.

3. Experimental Setup

In the following, we present each of the parts of our experimental approach in more detail. The setup consists of four principal parts, namely the dataset, the anonymization process, the automatic speaker verification, and the subjective evaluation. An overview can be seen in Figure 1.

3.1. Dataset

The datasets most commonly used for ASV in the community are VoxCeleb1 [16] and VoxCeleb2 [17], consisting of more than 7,000 speakers with a total of over one million utterances. The speakers are actors or celebrities, of several ethnicities, accents, and gender, uttering spontaneous speech, extracted from interview videos uploaded on YouTube. Metadata describing the gender and nationalities, of each speaker, is also available. In this experiment, we used the Voxceleb1 dataset consisting of 1,251 speakers with pre-defined split into 1,211 speakers for training and remaining 40 speakers for test.

As the utilized speaker verification model is pre-trained on VoxCeleb1 train data, tests are run on all 40 speakers from the test data of VoxCeleb1. The ASV system is enrolled with 90% of the speech samples from all the 40 speakers and the remaining 10% of the speech samples is used to a) validate the performance of the ASV model, and b) anonymize and objectively evaluate the success of anonymization.

3.2. Anonymization Process

To anonymize the selected test samples of the VoxCeleb1 dataset, we first use automatic speech recognition (ASR) to transcribe the input speech into text. For this process, we use the Google ASR service via an API call that autodetects the language of the speech and outputs a text in the respective language. For a real application in terms of anonymization, a local service should be used. We made use of the Google ASR to have a reliable ASR performance [18, 19].

This information is then passed to a Text-to-Speech module. As in previous experiments [7, 8], we use eSpeak NG which is based on a formant synthesis method. eSpeak is open-source, cross-platform, and compact software for speech synthesis that converts text to phonemes, using pitch and length information. It provides support for over 100 languages including English, German, French, Spanish, etc. with various accents and dialects. Unlike other synthesizers, based on human speech recordings. The synthesized speech from eSpeak is created using additive synthesis and an acoustic model, often resulting in synthesized speech that may be characterized as “not as natural” or “not smooth”. Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech.

3.3. Automatic Speaker Verification

As an initiative to develop voice privacy tools, [20] introduced speaker anonymization tasks and its evaluation method. From this collection of tools, the Automatic Speaker Verification (ASV) system is used as a benchmark objective evaluation methods for speaker verifiability before/after anonymization. This work uses VGGVox [16], an ASV system, which uses a VGG-GNet [21] as backbone architecture. VGGVox speaker verification model primarily aims to replace techniques of traditional hand-crafted features by a CNN architecture where the features required for the speaker recognition are chosen by the model. This minimizes the pre-processing of the audio data and preserves valuable information in the process. To achieve consistency, the in the current investigation generated audio input given to the model is converted into a single channel, 16-bit stream at a 16kHz sampling rate. Spectrograms of size 512x300 for 3 seconds of speech are generated, using a hamming window of width 25ms and step size of 10ms, in a sliding window fashion. To improve the classification accuracy, mean and variance normalization of each frequency bin of the spectrum is performed. Without further speech-specific pre-processing, these spectrograms are used as the input to the CNN. Due to its convolutional design, the CNN can perceive stimuli at specific locations, which has proven beneficial for analyzing spectral representations of audio signals, and extracting a robust set of features.

The ASV first registers the voiceprints of the enrolled speakers. The test speakers’ TTS voices are then compared to the enrolled speakers. This is done by extracting speaker embeddings of both the enrolled speakers and the test speakers. A distance measure such as Euclidean or cosine function is used to calculate the difference between the speaker embeddings. The closer the distance, the higher the probability that the audio is verifiably the enrolled speaker. This provides an objective evaluation of speaker anonymization.

3.4. Subjective Evaluation (Crowd-based)

While the objective evaluation metric measures the difference in acoustic features using a learned model, subjective evaluations are performed on the basis of listeners’ ratings. The results from these two evaluation methods may not be congruent or

---

1https://github.com/a-nagrani/VGGVox
2https://github.com/espeak-ng/espeak-ng
To evaluate the given ASR-TTS anonymization from a subjective view, we conduct a crowdsourcing study with 200 participants on the German Crowdee platform. Each participant was given 10 comparison tasks between a TTS-generated voice sample, and the original human voice sample. Each pair was created in a way, that the original speech file of a VoxCeleb1 speaker was compared to a synthesized version of a speech file of that speaker, thus containing different content and wording. Only a small margin of the test samples is being falsely rejected.

The set of pairs was chosen at random out of nine lines of stimuli where the pairs of stimuli do not follow the above-mentioned arrangement but contain speakers of different gender between the original and synthetic voices. In this way, the similarity scores were obtained on a seven-point Likert-scale (see Figure 2) using a slider with discrete positions and fixed labels at positions 1: “very different”, 4: “unsure”, and 7: “very similar”. The set of pairs was chosen at random out of nine original speakers (6 female and 3 male speakers) and 10 TTS voices (5 female and 5 male speakers).

To execute quality control on the rating procedure, we utilized a platform function involving the automated set-up and online scoring of trapping instances. In more detail, two traps were hiddenly included in the series of 10 comparisons per crowd task, where the pairs of stimuli do not follow the above-mentioned arrangement but contain speakers of different gender between the original and synthetic voices. In this way, the similarity rating between a male and a female voice should always result in a low score rating. For any rating beyond a value of 3, i.e. including the “unsure” answer option, we increase an accumulating error counter overall tasks of a rater by 1. At a value of 1, raters were explicitly notified to work more carefully in order not to get excluded. Raters were excluded when exceeding a given threshold set to the value of 2, which after all resembles a rather conservative setup. In addition, answers showing a particularly short working time were also rejected. This way, ratings of unmotivated contributors were excluded from the follow-up analysis.

### 4.1. Objective Evaluation

In the objective evaluation, we aim to validate the "TTS-anonymized" voices using a pre-trained ASV-model. We claim that an anonymization is successful, if the TTS-anonymized voices are not accepted as any of the previously enrolled speakers. Therefore, we first validate, that the ASV model can correctly identify the originally enrolled speakers using their test samples, as described in Sec. 3.3. Our results are in line with the results reported in [16]. They achieved a speaker identification performance of 92.1% Top-5 accuracy, and an Equal Error Rate (EER) for the speaker verification task of 7.8% on the VoxCeleb1 data. In our experiments with the VoxCeleb1 data, we achieved a recognition rate of 95.66% for non-anonymized speech. As expected, in the experiments conducted for this paper most of the enrolled human speakers are identified correctly when using non-anonymized original speech samples, as it can be seen in Figure 3. Only a small margin of the test samples is being falsely rejected.

Figure 2: Questionnaire on Crowdee platform

As the TTS synthesis of eSpeak is not as mature as other state-of-the-art systems in terms of naturalness, the test subjects were explicitly instructed to ignore errors and artifacts, like disfluencies, unnatural rhythm, and strange intonation or pitch melodies. They were asked to concentrate on the mere essential sound of the synthetic voice and their voice impression, and always judge if the synthetic voice underlying the synthetic speech sounds in comparison to the original voice.

The similarity scores were obtained on a seven-point Likert-scale (see Figure 2) using a slider with discrete positions and fixed labels at positions 1: “very different”, 4: “unsure”, and 7: “very similar”. The set of pairs was chosen at random out of nine original speakers (6 female and 3 male speakers) and 10 TTS voices (5 female and 5 male speakers).

To execute quality control on the rating procedure, we utilized a platform function involving the automated set-up and online scoring of trapping instances. In more detail, two traps were hiddenly included in the series of 10 comparisons per crowd task, where the pairs of stimuli do not follow the above-mentioned arrangement but contain speakers of different gender between the original and synthetic voices. In this way, the similarity rating between a male and a female voice should always result in a low score rating. For any rating beyond a value of 3, i.e. including the “unsure” answer option, we increase an accumulating error counter overall tasks of a rater by 1. At a value of 1, raters were explicitly notified to work more carefully in order not to get excluded. Raters were excluded when exceeding a given threshold set to the value of 2, which after all resembles
In the case of our implemented speaker anonymization, the ASV system is still enrolled with the voices of the human speakers but tested using speech samples generated by the eSpeak TTS, as described in Sec. 3.2.

Contrary to the assumption made in the introduction, that TTS-generated samples are not recognized as any of the human speakers, we observe that the TTS-voices are, in fact, able to more or less fool the system by confusing as one of four human speakers of the test set, see Figure 4, with approximately 6% recognition rate (falsely recognized). Especially, the speaker with the ID 10283 can nearly always confuse with several TTS voices.

The reason for this false acceptance is due to the general processing of the speaker embeddings (x-vectors) within the verification system. To decide, if an actual utterance belongs to the desired enrolled speaker (with known voiceprint), the speaker embeddings of the utterance under consideration are compared to the speaker embeddings of the voiceprint. This comparison is based on a distance measure. To illustrate this problem, the cluster embeddings of all 40 human speakers are depicted in Figure 5. The colored speaker embeddings belong to the four speakers that could be confused with a TTS voice, and are represented by a color ranging from yellow to red. It can be seen that all speaker embeddings are (evenly and more or less sharply) clustered within the space and that for most speakers the embeddings only have slight but noticable variations. This behavior is as expected for human voices, due to different pronunciation, emotional tone, etc. [22].

The cluster of the pink speaker (•), ID 10283, aka Eli Roth, is located at the upper middle of the embeddings space and has a large spreading, while other highlighted speakers (which are partly confused with our TTS-voices) have smaller variation.

The zoom-in shown in Figure 6 presents a cutout of the embedding space, now only showing the human speakers of interest, in the red-yellow color range (••••) and the TTS speaker set in the green-blue color range (••••••••••). Here we can see that many of the TTS embeddings are close to the speaker embeddings of ID 10283, aka Eli Roth (•), while having a larger distance to any of the other human speaker clusters. This proximity is the reason for the "misbehavior" of the ASV system. An utterance is accepted to belong to a speaker, if it has the closest distance
to that speaker than to any other speaker. This is depicted by the orange arrows in Fig. 6.

Although, the ability of a TTS-voice fooling the ASV system could be seen as an accident, this is still a severe problem, especially, as our TTS-system produces quite unnatural voices and the selection of the 40 test speakers is quite arbitrary. The location of the speaker embeddings is dependent on the model parameters of the ASV system and can differ with varying model implementations, the same applies to the synthesized voices. Thus, the observation we made for the pink speaker (•), ID 10283, aka Eli Roth, must not remain for other ASV models. Instead, other human speakers may then be subject to confusion.

But the question is, if these original speakers actually “sound” similar to the confused TTS-voices, i.e. do the speaker embeddings have something in common with audible differences? Therefore, we performed a listener evaluation, as described in Section 3.4.

4.2. Subjective Evaluation

![Confused vs. Non-confused Speaker Similarity Ratings](image)

Figure 7: Similarity ratings generated by listener evaluations of TTS and original voices for the two groups of speakers.

For the listener evaluation, samples with a huge similarity in the speaker embeddings space were chosen, to be compared against the TTS-generated audio files. These speakers are termed as confused speakers. Furthermore, original samples and TTS-generated samples for speakers that are not confused (non-confused speakers, aka having a large distance in the embedding space) are used for comparison.

The crowdsourced experiments show that on average, the human raters were unsure whether the voices were (acoustically/audible) similar for both groups of speaker comparisons. This can be seen in the direct comparison of the average similarity ratings given by the 200 test subjects. Figure 7 shows that the voices that had high false acceptance during the objective evaluation (confused speakers) have been rated very similar to those that were not falsely accepted as human speakers (non-confused speakers).

5. Conclusion and Outlook

Through the objective and subjective evaluation presented in the current paper, we now know that the problem of speaker confusion regarding anonymization using ASR+TTS-based voice conversion lies in the ASV system, its feature space, and distance metrics. This confusion shows how easy it can be to spoof an unprotected verification system, especially as the “attacks” were not targeted at a certain speaker, as well as the implicit dangers if TTS is used for speaker anonymization. The ISO/IEC norm 30107-1:2016 on biometric presentation attack detection [23] also highlights the vulnerability of ASV Systems. ASVSpoof [24] is an ongoing series of challenges, that focuses on the problem of attacks against ASV systems and their countermeasures. Especially voice conversion, speech synthesis, and audio replay are identified as the biggest threats to ASV system security.

The question now remains on how to avoid the behavior of falsely accepting a TTS voice as an enrolled human speaker. The solution cannot lie in a better description of the embedding space, as it is practically impossible to have a well-balanced embedding space that covers a big variety of speakers and also includes all TTS voices to avoid confusion later on.

It is obvious, that the use of Euclidean distance to separate different identities might not be discriminative enough. The clusters in the feature map of the TTS voices seem more compact, but if such a cluster is too close to a human speaker, it is always matched to it. Thus, using a probability density function to describe the compactness of the speakers’ voiceprints may be beneficial. But, it may still be possible to “confuse” human (original) voices and TTS (anonymized) voices, in cases where the embeddings of a TTS voice are “near” to a human speaker, see the speaker embeddings highlighted with • and • in Fig. 6.

Regarding speaker-conversion based anonymization, this means that although the process itself should lead to anonymization the target ID is chosen at random, the selected target speaker should ideally lay further away in the feature space, so it is not associated with the original speaker in any way.

To analyze this problem, our future research will include broader tests on other ASV and TTS systems such as Google, Amazon, and Azure TTS and the identification of acoustic characteristics that are the cause for the proximity of the corresponding speaker embeddings.

6. Acknowledgments

This research has been partly funded by the Federal Ministry of Education and Research of Germany in the project Emonymous (project number S21060/A) and partly funded by the Volkswagen Foundation in the project AnonymPrevent (AI-based Improvement of Anonymity for Remote Assessment, Treatment and Prevention against Child Sexual Abuse).

7. References


