Anonymization of Stuttered Speech – Removing Speaker Information while Preserving the Utterance

Jan Hintz1,∗, Sebastian Bayerl2,*, Yamini Sinha1, Suhita Ghosh3, Martha Schubert1 Sebastian Stober1, Korbinian Riedhammer2, Ingo Siegert1

1Mobile Dialog Systems, Otto-von-Guericke-University Magdeburg, Germany
2Technische Hochschule Nürnberg, Germany
3 Artificial Intelligence Lab (AILab), Otto-von-Guericke-University, Magdeburg, Germany

Abstract

Concealing the identity through speaker anonymization is essential in various situations. This study focuses on investigating how stuttering affects the anonymization process. Two scenarios are considered: preserving the pathology in the diagnostic/remote treatment context and obfuscating the pathology. The paper examines the effectiveness of three state-of-the-art approaches in achieving high anonymization, as well as the preservation of dysfluencies. The findings indicate that while a speaker conversion method may not achieve perfect anonymization (Baseline 27.25% EER and F0 Delta 32.63% EER), it does preserve the pathology. This effect was objectively evaluated by performing a stuttering classification. Although this solution may be useful in a remote treatment scenario for speech pathologies, it presents a vulnerability in anonymization. To address this issue, we propose an alternative approach that uses Automatic Speech Recognition and text-based speech synthesis to avoid re-identification (48.27% EER).

Index Terms: speech recognition, speech synthesis human-computer interaction, computational paralinguistics, dysfluencies, voice conversion, anonymization

1. Introduction

Stuttering is a varied speech disorder that can be characterized by its core symptoms; Blocks, prolongations, and repetitions of sounds, syllables, and words [1]. Having a stutter will eventually limit a person’s ability to communicate effectively, which also affects the use of voice technology. Such user interfaces are benefiting from ever-improving models for speech recognition and understanding, as well as more intelligent assistant functions. Moreover, voice assistants have the potential to drive forward inclusion and provide equal access to information and communication for people with voice disorders [2]. Especially in the sector of electronic health (eHealth), these applications have the chance to substantially reduce the costs of therapy and increase access to therapy. A prerequisite for these applications is the ability to automatically process and detect disordered speech, which is a sub-field of pathological speech research with a focus on developing systems for the automatic processing, screening, and monitoring of non-normal speech recordings [3].

As with other pathology-related speech processing, privacy is a major concern because speech recordings are highly sensitive personal data. A person using speech interfaces could, for example, be identified by their unique way of stuttering, which enables user profiling or identification in everyday interactions [4, 5]. Therefore, it is important to account for privacy concerns when collecting, transmitting, and analyzing pathological speech data. eHealth applications, e.g., a remote stuttering therapy application, need to preserve the pathology-related patterns in speech. Thus, balancing the benefits of speech data analysis with the protection of personal data is a challenging task. In another application, critical speech patterns, such as voice disorders, should be masked to prevent the identification of individuals based on speech impairments. This leads to a conflict of interests. On the one hand, speech data can provide valuable insights and help improve applications or deliver more inclusive speech technology. On the other hand, it is important to ensure that an individual’s privacy rights are not violated. This calls for techniques that can preserve what is being said while protecting the identity of the person speaking; Anonymizing speech data on the user’s device before it is transferred for analysis presents a possible solution.

To evaluate different approaches to speaker anonymization, Tomashenko et al. [6] introduced the voice privacy challenge. They also introduce the Word Error Rate (WER) as a usability metric for anonymized speech, to measure intelligibility. While this is a relevant criterion with non-disturbed or non-pathological speech, the error introduced by stuttering affects the off-the-shelf automatic speech recognition systems (ASR) results significantly [4]. Based on these considerations, we classify elements of stuttering before and after anonymization to see if the obfuscation process affected the pathology, instead of relying on the already faulty ASR outcome for non-anonymized data.

It has to be noted that stuttering is particularly hard to detect automatically, as there is no permanent physical change to the speech production process of a person who stutters (PWS). It is however a task, situation, and environment-dependent occurrence of different types of dysfluencies [7]. Stuttering detection of different dysfluencies using w2v2vec 2.0 (W2V2) embeddings was conducted by [8]. A recent stuttering classification benchmark was introduced by the ComParE challenge [9]. Most contributions used W2V2-based systems. [10] and [11] used pre-trained W2V2 models as feature extractors. Grözse et al. used fine-tuning of different W2V2 models for stuttering classification, yielding an unweighted average recall of up to 62.1% on the eight-class problem [12]. Another W2V2-based system could be further improved by cross-language training and fine-tuning using additional training data [13]. Using stuttering classification to showcase the efficacy of anonymization sheds light on many pathological speech-processing tasks due to the large differences in the characteristics of the core symptoms.

This paper examines for the first time how different approaches to anonymization can be used in such a remote scenario to ensure anonymity and (at the same time) to grant the possibility for therapy. Furthermore, we apply a novel voice-conversion method that better preserves dysfluencies.
2. Methods

2.1. Dataset

The Kassel State of Fluency Challenge dataset (KSF-C) for stuttering classification was recently introduced by the ComParE challenge [9]. The dataset is based on the Kassel State of Fluency (KSoF) dataset, which consists of three-second-long clips extracted from recordings of 37 speakers (28 male / 9 female) in different stages of therapy for stuttering [14, 15]. We acquired the 214 full recordings from the authors from which the segments were extracted, including time-aligned high-quality verbatim transcripts used for analysis in [15]. The full recordings are used for the experiments involving speech recognition. The stuttering classification task considers eight classes; Blocks, fillers, modifications, prolongations, sound repetitions, word repetitions, no dysfluencies, and a garbage class (background noise, music, etc.). The dataset (KSF-C) is limited to unambiguously labeled clips, i.e., clips with only one labeled dysfluency, which allows us to focus on the effects of anonymization on stuttering classification and not on the intricacies of multilabel classification. Using the challenge dataset also allows for experimental results to be compared to existing baseline results and challenge contributions.

In addition to the 4,601 clips in the KSF-C dataset, we used the 22,100 unambiguously labeled clips, from two English datasets with compatible stuttering labels in the training of the classification systems; SEP-28k-Extended and FluencyBank [16, 17, 18]. The three datasets containing stuttered speech are available for research purposes.

2.2. Anonymization

We propose two distinct methods for anonymization: 1) Voice conversion, where the voice of a speaker is altered to sound like a target speaker and 2) speech re-synthesis, where the speech is generated by a Text-to-Speech (TTS) system from an ASR result.

2.2.1. StarGAN

StarGAN-v2VC [19] is a generative adversarial network (GAN) based approach for speaker anonymization, where the speaker embeddings are learned during the training.

- **Speaker Encoder**: The speaker encoder (SE) extracts the embeddings of the speakers $s' \in S$. Provided a reference utterance of a speaker $s'$, such that $s' \neq s$, the SE learns the representations of the speaker $h_{s'}$, through the style re-construction loss, as mentioned in [19]. The speakers $s$ and $s'$ should not be the same, so the model is forced to learn separate embeddings for different speakers through the style diversification loss, also explained in [19]. This loss ensures that the converted samples sound different for different target speakers.

- **Content Encoder**: This encoder learns representations of the actual content, which is not dependent on the speaker or the speaker’s style.

- **Generator**: The generator $G$ takes three inputs, a source mel-spectrogram, target speaker style-code, and latent pitch representations, to produce the converted mel-spectrogram. The converted mel-spectrogram bears the style/timbre of the target speaker and the linguistic content of the source.

- **Discriminator and Speaker Classifier**: The architecture has a discriminator $D$, as present in any GAN model, which learns the representations for the real and fake samples. $D$ has shared layers that learn the representations of the real and fake samples common for all speakers, which are followed by a binary speaker classifier, which classifies whether the sample is real with respect to the speaker $s' \in S$. The additional classifier has the same architecture as the $D$, which helps in learning representations that are specific to a particular speaker, i.e., real with respect to a specific speaker and not in a general speech sample. We incorporate two additional losses, which facilitate the model to capture the features specific to stuttering data. 1) Change in pitch or pitch-modulation is a prosodic feature, which is found to capture subtle changes in intonation, pauses, or ‘trembling voice’, pertinent to model stuttering in the converted sample. We compute a loss based on the subtle fluctuations between the source and converted samples. 2) During moments of stuttering, a person may speak more softly or loudly than usual, and these features can be captured by a change in intensity. Therefore, we incorporate a loss computed on the change in intensity between the source and the converted samples.

2.2.2. Re-Synthesis

The speech re-synthesis approach converts speech to text (STT) and resynthesizes the text to speech (TTS) as seen in Figure 2.

![Figure 2: Re-Synthesis anonymization approach](image)

- **STT**: Whisper is a state-of-the-art automatic speech recognition model with an encoder-decoder transformer architecture that was trained on large amounts of multilingual speech data and, therefore, more robust [21]. The model is trained using multitask learning with weak supervision, detecting no-speech, English transcription, non-English transcription, and any-to-English speech translation, making it an ideal out-of-the-box system to transcribe stuttered speech. For our experiments, we used the medium-sized model with the language set to German and the decoding beam size set to 5 with temperature fallback to transcribe the audio files used in the ex-
2.3. Evaluation

2.3.1. Speaker Verification

For speaker verification, we utilize ECAPA-TDNN[28] embeddings extracted by a model of the same name, included in the Nemo Toolkit. The model was trained with an end-to-end approach using angular softmax loss, yielding embeddings of size 192. We are following the evaluation scenario “ignorant” [29]. The 214 full original recordings as described in Section 2.1 are enrolled into the system. The enrolled embeddings are compared against all segment embeddings by calculating the Euclidean distance. The EER is used as an anonymization metric as described by [6].

2.3.2. Stuttering Classification

In order to check the preservation of stutter-typical dysfluencies after anonymization, we use the KSF-C challenge task [9]. The challenge task was evaluated using unweighted average recall (UAR) as the deciding metric. We report detailed recall results to be able to discuss the preservation of certain dysfluency patterns. The current state-of-the-art classification system is a large wav2vec 2.0-based system and is described in more detail in [13]. The model used in our experiments uses the same architecture and was trained using multitask learning on a cross-language and cross-dataset task utilizing a combination of the three datasets containing stuttered speech described in Section 2.1. We follow the implementation and training details of the multi-task and multi-language system described in [13].

The evaluation of the stuttering classification considers three cases: directly classifying anonymized stuttering data using a model trained on stuttering data that was not anonymized before, and secondly, models fine-tuned using the anonymized stuttering data from the different systems and subsequently running the same evaluation. The third case considers applying the model fine-tuned on anonymized data and applied to the original data. This is done to check different properties of the data. First, we want to check if detecting dysfluencies in anonymized data is possible using existing models. Secondly, it gives insight if models can be fine-tuned utilizing anonymized data to detect dysfluencies in anonymized data, given enough dysfluency information is preserved, and thirdly if we can use these fine-tuned systems trained on anonymized data to detect dysfluency in the original training data. Implementation and fine-tuning hyperparameters used for fine-tuning are identical to the ones described in [13].

### Table 1: Speaker Verification Results

<table>
<thead>
<tr>
<th>Re-synthesis</th>
<th>KSF C</th>
<th>StarGAN Baseline</th>
<th>StarGAN F0 Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER(%)</td>
<td>8.24</td>
<td>27.25</td>
<td>32.63</td>
</tr>
</tbody>
</table>

The confusion patterns for the classification of anonymized data with non-adapted systems for baseline and the improved system (Figure 3) look similar, with lots of confusion towards the garbage class. However, it is good that the confusion mostly happens with the garbage class and not with other classes. This shows that the classification is not random and that the model works as intended for the Baseline and StarGAN anonymization techniques. Besides sound repetitions, there is almost no recall of the other classes. The data anonymized using the SST-TTS technique is almost exclusively confused with the no dysfluencies class, which is also expected as the model tries to create normal and fluent speech.

The fine-tuned stuttering classification system, using the stuttering data anonymized with either the baseline or the improved model, outperforms the baseline system by a substantial margin [9] and even outperforms some submissions to the challenge [11, 30, 10]. The overall results include the recall value for the garbage class in the UAR, which somewhat diminishes the results of keeping the dysfluency information. Because the garbage class cannot be anonymized Meaningfully, as all systems are trained on voice, and garbage mostly involves background noise and does not typically involve speech. Therefore, it is unsurprising that the garbage class is not recalled using the SST-TTS anonymized data, which is a desirable outcome, as no speech is disregarded.

The overall poor performance on word repetitions is also not surprising, using no textual information for their recognition, and the training data is strongly imbalanced, with only 3.6% of segments labeled as word repetitions. There were even challenge contributions that did not achieve a recall of > 0% for word repetitions [13]. This issue could potentially be solved with data augmentation techniques, which can be considered as future work.

Surprisingly, the overall best recall performance was achieved by the model previously fine-tuned on the baseline-anonymized data and reapplied to the original data. This could be due to a normalization effect of the anonymization, as it removes at least some speaker variance from the stuttered speech data.

In a remote therapy scenario, keeping the modified speech pattern intact is important. It enables generating feedback regarding the correct application of the speech technique. The high recall rate on the modified class with up to 96% is desirable in the context of speech therapy, while at the same time concealing the speaker’s identity when transferring speech data to a third party.

To our surprise, the TTS anonymization produced data, with which we still could train dysfluency classification systems, that beat the baseline of the compared challenge data. Even including the recall of 0% on the garbage class, which for
latter. It would also
be of high interest to see how humans would rate the level of
anonymization and pathology preservation.

\begin{table}
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline
\hline
\hline
None [13] & Org & Org & 62.78 & 47.73 & 86.75 & 31.25 & 91.24 & 65.45 & 83.33 & 22.22 & 74.24 \\
\hline
StarGAN Baseline & Org & Anon & 35.50 & 21.59 & 19.28 & 81.25 & 35.65 & 9.09 & 72.73 & 11.11 & 33.33 \\
& Anon & Anon & 52.66 & 71.02 & 63.86 & 25.00 & 89.12 & 43.64 & 57.58 & 5.56 & 65.53 \\
& Anon & Orig & 60.70 & 69.89 & 85.54 & 12.50 & 91.24 & 70.00 & 53.79 & 33.33 & 69.32 \\
\hline
StarGAN F0 Delta 2 & Org & Anon & 35.39 & 16.48 & 16.87 & 81.25 & 45.02 & 10.91 & 62.88 & 11.11 & 38.64 \\
& Anon & Anon & 46.38 & 59.66 & 55.42 & 12.50 & 94.26 & 42.73 & 50.76 & 0.00 & 55.68 \\
& Anon & Orig & 56.01 & 64.20 & 81.93 & 12.50 & 96.07 & 74.55 & 47.73 & 5.56 & 65.53 \\
\hline
Re-Synthesis & Org & Anon & 14.15 & 6.82 & 1.20 & 12.50 & 0.60 & 0.00 & 2.27 & 0.00 & 89.77 \\
& Anon & Anon & 22.25 & 30.11 & 25.30 & 0.00 & 50.15 & 8.18 & 7.58 & 5.56 & 51.14 \\
& Anon & Orig & 48.74 & 35.80 & 85.54 & 0.00 & 88.82 & 79.09 & 53.79 & 5.56 & 41.29 \\
\hline
\end{tabular}
\caption{Unweighted average recall (UAR) and detailed per-disfluency recall results for experiments using a stuttering classification model trained on cross-lingual training data (Orig), and results for stuttering classification models fine-tuned with the respective anonymized stuttering data (Anon).}
\end{table}

obvious reasons cannot be resynthesized with speech models, the UAR would even go up to 48 %, improving over the baseline by a substantial margin after retraining and applying the model to the non-anonymized data. One reason for this result is that the Whisper ASR system produces transcripts with artifacts that preserve stuttering information. The system does not only decode existing words but also part words, like other modern ASR systems. In cases where synthesis can reasonably re-synthesize repetitions of single sounds, stuttering information would even be preserved acoustically. Two examples of these artifacts can be found in the transcription of files containing sound repetitions:

1. “TT-TT-TTT tat”
2. “Und ar-ar-ar-ar-ar-ar-ar-are” 

- This puts even more emphasis on text normalization prior to re-synthesis if pathology information should be removed.
- Also shows that unlike previously assumed, STF-TTS might still be an option in cases where pathology information should be preserved.

Results of the models fine-tuned on anonymized data and applied to the original data could make data donations easier, as fully anonymized data donations would be possible, lowering the consent threshold when generating datasets containing pathological speech. It could mean that using a small base dataset of real speakers could be utilized to initialize a model that is then further fine-tuned on anonymized data.

4. Conclusion

Our research shows that the two presented approaches to speech anonymization have different advantages. While re-synthesis grants the user high anonymity, it is difficult to convey information about the speech disorder as word repetitions and prolongations – essential parts of stuttering – are masked by the anonymization process. Our results suggest that a voice conversion using our modified version of StarGAN gives the best trade-off between anonymization and preserving stuttering for speech therapy purposes. This has yet to be confirmed for other speech pathologies.

For future works, a phoneme-based transcript and synthesis, as well as ASR systems specifically adapted for stuttered speech, could be interesting approaches to preserve speech pathologies better while granting high anonymity. It would also be of high interest to see how humans would rate the level of anonymization and pathology preservation.
5. Acknowledgements
This research has been funded by the Volkswagen Foundation in the project AnonymPrevent (AI-based Improvement of Anonymity for Remote Assessment, Treatment, and Prevention against Child Sexual Abuse).

6. References