IDIAP RESEARCH REPORT

ADJUSTABLE DETERMINISTIC PSEUDONYMIZATION OF SPEECH

S. Pavankumar Dubagunta Rob Van Son
Mathew Magimai.-Doss

Idiap-RR-12-2021

AUGUST 2021

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Adjustable Deterministic Pseudonymization of Speech

S. Pavankumar Dubagunta*[1,2], Rob J.J.H. van Son[3,4], and Mathew Magimai.-Doss[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne (EPFL), Switzerland
[3]Netherlands Cancer Institute (NKI-AVL), Amsterdam, Netherlands
[4]ACLC, University of Amsterdam, Amsterdam, Netherlands

July 2021

Abstract

While public speech resources become increasingly available, there is a growing interest to preserve the privacy of the speakers, through methods that anonymize the speaker information from speech while preserving the spoken linguistic content. In this paper, a method for *pseudonymization* (reversible anonymization) of speech is presented, that allows to obfuscate the speaker identity in untranscribed running speech. The approach manipulates the spectro-temporal structure of the speech to simulate a different length and structure of the vocal tract by modifying the formant locations, as well as by altering the pitch and speaking rate. The method is deterministic and partially reversible, and the changes are adjustable on a continuous scale. The method has been evaluated in terms of (i) *ABX* listening experiments, and (ii) automatic speaker verification and speech recognition. *ABX* experimental results indicate that the speaker identifiability among forced choice pairs reduced from over 90% to less than 70% through pseudonymization, and that

*Corresponding author. Email: dspavankumar@gmail.com. Address: Idiap Research Institute, Centre du Parc, Rue Marconi 19, Martigny, Valais, Switzerland, CH-1920.

de-pseudonymization was partially effective. An evaluation on the VoicePrivacy 2020 challenge data showed that the proposed approach performs better than the signal processing based baseline method that uses McAdams coefficient and performs slightly worse than the neural source filtering based baseline method. Further analysis showed that the proposed approach: (i) is comparable to the neural source filtering baseline based method in terms of phone posterior feature based objective intelligibility measure, (ii) preserves formant tracks better than the McAdams based method, and (iii) preserves paralinguistic aspects such as dysarthria in several speakers.

Keywords speech privacy, speech pseudonymization, speech signal processing, speech features.

1 Introduction

The availability of large speech corpora in combination with advanced statistical techniques improved speech technology tremendously [Ardila et al., 2019, Panayotov et al., 2015, Zhang et al., 2017, Ning et al., 2019]. But speech recordings pose a possible privacy risk. More and more resources of speech data are shared on public platforms each day. While personally identifiable information such as name, age etc. of the speaker can be easily hidden, speech itself remains as a personal identifier of the speaker. With the increased use of speaker verification technologies, sensitive information related to speakers could be extracted from their speech and lead to harm [Korshunov and Marcel, 2017, Kucur Ergunay et al., 2015]. This is especially true when the speakers have medical conditions, are minors, or the spoken content is sensitive. But these are also groups that might benefit from improvements in speech technology tailored to their needs.

The privacy risks resulting from sharing speech recordings would be mitigated if the probability of speaker (re-)identification could be reduced while retaining useful linguistic and paralinguistic features. Speech anonymization methods, thus, aim at decoupling the hazardous identity of the speaker from the interesting linguistic and paralinguistic aspect of the speech. That is, anonymization removes the information about *who spoke it* from the speech while preserving *what was spoken* and *how it was spoken*. The "perfect" anonymization procedure would correspond to having the spoken text read by another speaker in the exact same manner. And some current speech anonymization applications work using components of a speech recogniser coupled to a neural network based speech synthesizer [Fang et al., 2019, Mawalim et al., 2020]. However, such an approach only preserves the verbal content of the

speech, and at best some of the prosodic aspects. Such an approach may not be able to preserve paralinguistic features of interest, such as the expressed emotions, articulation changes depending on the speaking skills or pathological conditions etc., and in general may not preserve the linguistic detail. Thus, such an anonymization may not be useful in scenarios, such as (i) dysarthric patients uploading their speech for evaluation, (ii) children or language learners submitting their utterances for evaluation, where preserving paralinguistic information is important. An alternate way could be to use signal-processing approaches that directly alter the spectral properties of the original utterance for anonymization based on prior knowledge. Such an approach that uses the McAdams coefficient [Patino et al., 2020] exists. It is based upon short-time linear prediction analysis, where a constant exponentiation is applied to the angle of the complex poles, thereby expanding or contracting the timbre or the spectral envelope at the formant locations [Patino et al., 2020]. However the method performs inferior to that of the neural based approach in terms of automatic speech recognition (ASR) and automatic speaker verification (ASV). However, signal processing based approaches have the advantage over most statistical and machine learning approaches that the changes made and their effects observed can be explained and, ideally, controlled; hence there is an interest in improving such controllable approaches. A downside of the existing speech anonymization applications is also the degraded quality of the transformed (anonymized) speech [Srivastava et al., 2020] which reduces their usefulness. The research community has acknowledged these problems, and in 2020 a special challenge for improving anonymization of speech has been organised [Tomashenko et al., 2020a, Tomashenko et al., 2020b].

The literature on data anonymization (e.g., [Rubinstein and Hartzog, 2016, Stalla-Bourdillon and Knight, 2017, Finck and Pallas, 2020]) can be crudely summarised as "anonymous data is not useful, useful data is not anonymous". This is also likely to be true for speech anonymization transforms. Therefore, reversible anonymization of speech, also called *pseudonymization*, could shift the risk-benefit balance for sharing speech corpora towards more sharing, and is therefore of potential interest. Contrary to "true" (i.e. irreversible) anonymization, pseudonymization is a more practical approach to anonymization, since it assumes that data can be re-identified, in principle, with the help of additional information that is hidden during the anonymization. The risk of re-identification of pseudonymized data is then the risk that the hidden information can be reconstructed by an attacker. Pseudonymization of speech will always be a trade-off between the risk of re-identification and usefulness. Thus, there is a need to develop such pseudonymization methods so that several speech applications can benefit from the privacy benefits they offer.

This work aims at developing a pseudonymization approach that is adjustable

in the level of information removed from the speech, while still preserving relevant features enough to make the resultant speech useful. The proposed approach uses a series of signal processing steps to transform a given speaker's speech to tailor to a desired vocal profile (cf. [Kung, 2018]), configurable in terms of the formant frequencies, fundamental frequency and speaking rate.

We conduct three sets of studies to demonstrate the potential of the proposed pseudonymization approach:

1. First, we validate the proposed approach through *ABX* pilot tests. These studies are carried out to ascertain how well the proposed approach obfuscates the speaker identity for expert and naive listeners.

2. Second, we validate the proposed approach in the framework of VoicePrivacy 2020 challenge [Tomashenko et al., 2020c, Tomashenko et al., 2020a] by studying it against two anonymization approaches, a neural source filtering based approach and signal processing-based McAdams approach. We also perform ablation experiments to investigate which part of the proposed approach (related to the source, system or speaking rate) plays a crucial role in obfuscating speaker identity.

3. Third, we conduct studies that extend beyond the scope of VoicePrivacy 2020 challenge. In the VoicePrivacy 2020 challenge, the preservation of intelligibility is assessed through automatic speech recognition. Such a method can be prone to errors related to the availability of a suitable language model and a pronunciation lexicon. So, we propose the utilization of a recently proposed phone posterior feature-based intrusive objective speech intelligibility approach [Ullmann et al., 2015]. We also investigate the ability of the proposed pseudonymization method to preserve general articulatory features of speech by comparing the formant track movements measured on the anonymized and original recordings. Finally, investigations on speech anonymization have primarily laid emphasis on the preservation of intelligibility. However, speech also contains information other than the spoken message and speaker identity, such as paralinguistic information. So, we investigate the ability of the proposed pseudonymization approach to preserve such information through a dysarthric speech classification study.

The rest of the paper is organised as follows. Section 2 describes the signal processing approach to anonymization developed for adjustable deterministic pseudonymization of speech. Section 3 describes the listening experiments conducted using human listeners. The experimental setup for the 2020 VoicePrivacy Challenge

and the results are described in Section 4. Section 5 describes additional analysis in terms of intelligibility measurement based on dynamic time warping (DTW), formant measurement in pseudonymized speech and experiments on dysarthria prediction. Section 6 presents a discussion and Section 7 concludes the paper.

2 Proposed Pseudonymization Method

As illustrated in Fig. 1, the proposed pseudonymization approach consists of estimating the speaker characteristics and obfuscating them by providing a different set of characteristics (referred to as the *target* speaker) to modify the utterances. As we see later, the same pseudonymization module can be used to de-pseudonymize the utterances, upon the knowledge of the original speaker's characteristics.

Two sources of speaker variation useful for speaker identification can be distinguished, viz. *inherent* features, i.e., those that derive from a speaker's anatomy and physiology, and *learned* features [O'Shaughnessy, 2000]. This study aims at hiding the global and inherent features of speakers, i.e., the vocal tract related spectral features (cf. [Almaadeed et al., 2016]) and some learned features, i.e., pitch and speaking rate. This translates to making changes in speech that relate to vocal tract length, average formant frequencies and intensities, pitch, and speaking rate. The pieces of information thus hidden will be the original values of these quantities and the extent of the changes. The corresponding steps can be summarised as:

1. Change the speaking rate and fundamental frequency, and
2. Simulate a different vocal tract for the speaker.

The perceived acoustic length of the vocal tract of each speaker is changed to that of a desired speaker by changing the playback speed of the utterance. Specifically, the vocal tract length corresponds to formant values as follows: an increase of vocal tract length by a factor a induces a formant shift by a factor $1/a$. In the remainder of this paper, the estimated vocal tract length (VTL) will be represented by the neutral first resonance frequency ϕ. A speaker's $\hat{\phi}$ is estimated from the first four formant frequencies according to Eq. 20 of [Lammert and Narayanan, 2015] using the proposed extension (Table 3, ibid.):

$$\hat{\phi} = 229 + 0.030\phi_1 + 0.082\phi_2 + 0.124\phi_3 + 0.354\phi_4 \tag{1}$$

where $\phi_i = F_i/(2i-1)$ can be considered as estimates of VTL from individual formants F_i. Speakers do not only differ in vocal tract length, but also in the vocal

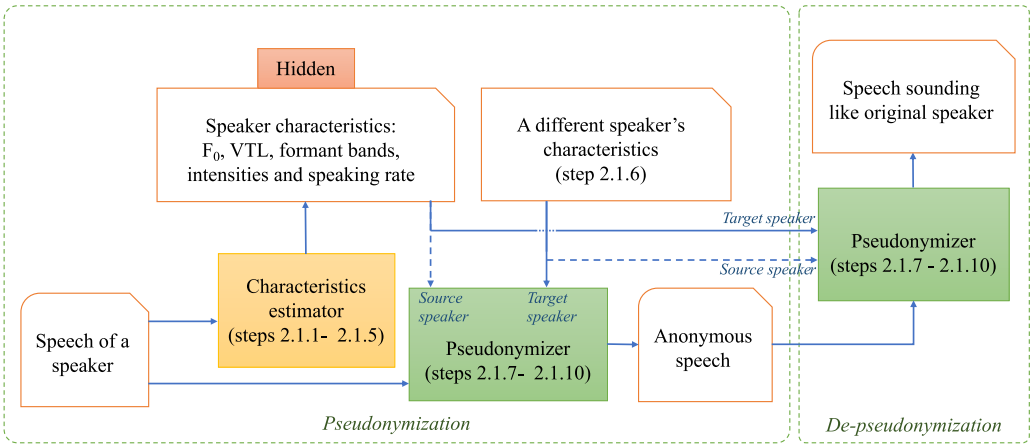

Figure 1: Illustration of speech pseudonymization, with the steps elaborated in Sec. 2.1.

tract structure, defined by the global position of the formants, their bandwidths and intensities. The below section describes how each of these quantities are estimated and are used to pseudonymize speech.

2.1 Steps involved

2.1.1 Intensity normalisation

Normalise the intensity of each utterance to 70 dB (relative to 20 μPa).

2.1.2 Identifying the vowel segments and estimating the speaking rate

Estimate the speaking rate by automatically locating syllables from speech using peaks in the signal energy, that are preceded and succeeded by dips in energy as cues [De Jong and Wempe, 2009, van Son et al., 2018]. The number of syllables normalised by the duration per speaker gives the speaking rate. This method requires no transcriptions.

2.1.3 Formant track estimation for each vowel region

1. Use short-time processing with a Gaussian-like window of 25 ms, repeated every 6.25 ms (see Sec. 2.2 for more details).

2. Formant track estimation in the vowel regions: Use linear prediction analysis and iterative formant estimation procedure from Lee [Lee, 1988].

2.1.4 Speaker-specific VTL and formant frequency estimation

1. In each vowel segment, look for the most neutral frame, i.e. the frame with (F_1, F_2) closest to $(500, 1500)$.

2. Attribute the formant estimates F_{1-5} from this closest frame to the entire vowel segment.

3. Estimate the VTL of the speaker in the vowel segment using Eq. 1.

4. Compute the speaker's VTL by taking the mean VTL across each speaker's vowel segments.

5. Compute the speaker formant frequencies F_i by taking the median across each speaker's vowel segments.

2.1.5 Speaker-level formant band intensity estimation

1. The frequency spectrum of each speaker is divided into several *formant bands* based on the estimated VTL[1] ϕ, as

$$B_i = \begin{cases} \left[0,\ \frac{\phi}{2}\right], & i = 0 \\ \left[\frac{\phi}{2},\ 2\phi\right], & i = 1 \\ \left[2(i-1)\phi,\ 2i\phi\right], & i = 2, 3, ..., 9 \end{cases} \tag{2}$$

   in Hz. Since F_i $(i = 1, 2, ..., 9)$ is typically around $(2i - 1)\phi$, the bands are centered around the corresponding formant frequencies (except B_0 and B_1).[2]

2. Use a passband Hann filter to isolate the information in each band. The filter has the following properties: (i) it is real-valued and operates on the complex short-time Fourier transform (STFT) of the input utterance, independently across each time step, (ii) the passband frequencies and 3dB bandwidth are defined as above, (iii) the transition from stop band to pass band and vice versa spans $(i - 1)\phi/5$ Hz.

[1] We will use the symbol ϕ to mean $\hat{\phi}$ hereafter.

[2] E.g. $\phi = 500$ (a typical male value). So, $F_3 \approx 2500$ Hz, $B_3 = [2000, 3000]$ Hz, $F_4 \approx 3500$ Hz, $B_4 = [3000, 4000]$ Hz, and so on. $B_1 = [250, 1000]$ Hz, $B_0 = [0, 250]$ Hz.

3. Use the above filter on each utterance and measure the mean intensity per speaker per formant band, I_i, from the filtered utterances for $i = 0, 1, 2, ..., 5$.

2.1.6 Target parameters for pseudonymization

To pseudonymize the formants, the target frequencies, represented in terms of VTLs $\phi_i = F_i/(2i-1)$, can be randomly chosen in the range $\phi_i \pm 40$ and $\phi_i \pm 75$ Hz for F_{0-1} and F_{2-5}, respectively, and the intensities can be randomly chosen in the range 64±4.5, 67±2.5, 58±4.5, 50±8, 47±10, 45±9 dB (I_{0-5}±2SD), where SD denotes standard deviation. These values were chosen based on ranges found in the speakers in the IFA corpus [Van Son et al., 2001] (5M/5F, see Experiment 1, Section 3.1.1). In an alternative setting where a given speaker is to be pseudonymized to a specific target speaker, the parameters ϕ, ϕ_i ($i = 1...5$), I_i ($i = 0...5$) and speaking rate can be pre-computed across several of the target speaker's utterances (preferably over 300 seconds spoken in a comparable style) and used.

2.1.7 VTL shifting

This is a time-domain processing method.

1. We have a VTL estimated for the current speaker and a VTL estimate for the target speaker: determine the factor $a = \phi^{(current)}/\phi^{(target)}$.

2. Resample the utterance to F_s/a, where F_s denotes the original sampling frequency (and consider that the sampling frequency is still F_s). This corresponds to a frequency scaling by $1/a$ to the original utterance's spectrum.

2.1.8 Duration and pitch change

This is a time-domain processing method. Estimate F_0 by using the standard autocorrelation method. Adjust the duration and fundamental frequency to match the desired duration (determined by the target speaker's speaking rate) and fundamental frequency using pitch synchronous overlap-add method [Moulines and Charpentier, 1990].

2.1.9 Formant band shifting

This is a frequency-domain processing method. For each formant, we aim at masking $\phi_i^{(current)}$ and shifting it to the frequency $\phi_i^{(target)}$ by modifying its intensity appropriately.

1. Use the steps of VTL shifting (from Sec. 2.1.7), by using $\phi_i^{(current)}$ and $\phi_i^{(target)}$ in the place of $\phi^{(current)}$ and $\phi^{(target)}$ respectively, to create a VTL shifted version of the current utterance, where the formant i is now at $F_i^{(target)}$.

2. Extract the band $B_i^{(current)}$ (Eq. 2) from the VTL shifted spectrogram using a Hann filter as described in Sec.2.1.5[3]. Use $I_i^{(target)}/I_i^{(current)}$ as the filter's gain.

3. Use a complementary bandstop Hann filter with unit gain on the current utterance's spectrogram to mask $\phi_i^{(current)}$ in the band.

4. Add the extracted band to the current spectrogram so that it now has $\phi_i^{(target)}$ (and then discard the VTL shifted spectrogram).

5. Repeat the above steps for each desired formant.

2.1.10 Additional processing to hide the speaker identity

Additional anonymizing steps consist of (i) exchanging the B_4 and B_5 bands by using the Hann filter method described above and (ii) adding modulated pink noise at the speaker's B_{6-9} bands to mask these formants. These steps were not used in the human listening experiments in Sec. 3.

Finally, reconstruct the corresponding utterance by taking inverse STFT. Note that, except for the overlap-add synthesis step and noise insertion, all the steps in this process are deterministic and reversible.

2.2 Implementation

The software is available on GitHub [van Son, 2020d, van Son, 2020c]. The program *Praat* [Boersma and Weenink, 2017] has two commands *Change gender...* and *Change speaker...* that use the same algorithm to perform the respective operations. This study uses the *Change gender...* command internally because it has options suitable for the proposed approach. In these commands, the desired new pitch is set as an absolute value, but it depends on correct pitch measurement in the source speech. Both commands work on the vocal tract length and duration by a *Formant shift ratio* and a *Duration factor*. To implement a change to a specified target vocal tract length and duration, or speaking rate, the estimated vocal tract length and speaking rate of the source speaker have to be supplied.

[3] $\phi_i^{(target)}$ is largely present in $B_i^{(current)}$, as this band heavily overlaps with $B_i^{(target)}$, but not always.

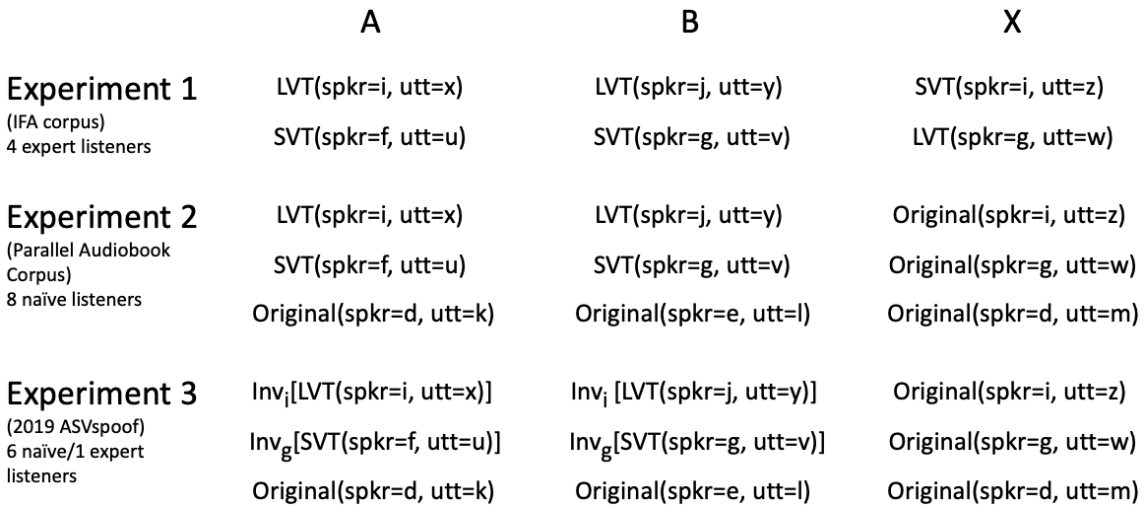

Figure 2: ABX listening experiments. The subjects had to identify which of the two utterances, *A* or *B*, was spoken by the speaker in *X*. LVT()/SVT(): Stimulus pseudonymized to a Long/Short Vocal Tract length, Original(): Original recording as stimulus, $\mathrm{Inv_i}$[]: Inverse, de-pseudonymized to the parameters of speaker 'i'. spkr: Speaker number, utt: Utterance number. For example, $\mathrm{Inv_g}$[SVT(spkr=f, utt=u)] indicates a stimulus created from utterance 'u' from speaker 'f', pseudonymized to a Short Vocal Tract length, and then de-pseudonymized to the parameters of speaker 'g'. See the text for details.

VTL is determined using the *Praat* `robust` formant option [Boersma and Weenink, 2017]. Speaking rate is determined by the syllable rate determined from a modified version of a script by De Jong and Wempe [De Jong and Wempe, 2009] taken from [van Son et al., 2018].

To pseudonymize an utterance, the original values of the VTL (ϕ), median formant frequencies, pitch, and speaking rate are transformed to the chosen values of the (synthetic) target speaker. The `PseudonymizeSpeech.praat` script [van Son, 2020c] presented above can calculate these on-the-fly using a collection of speech recordings or can use a database of pre-calculated values. Pseudonymization examples are available with the script, also consult the manual at [van Son, 2020d].

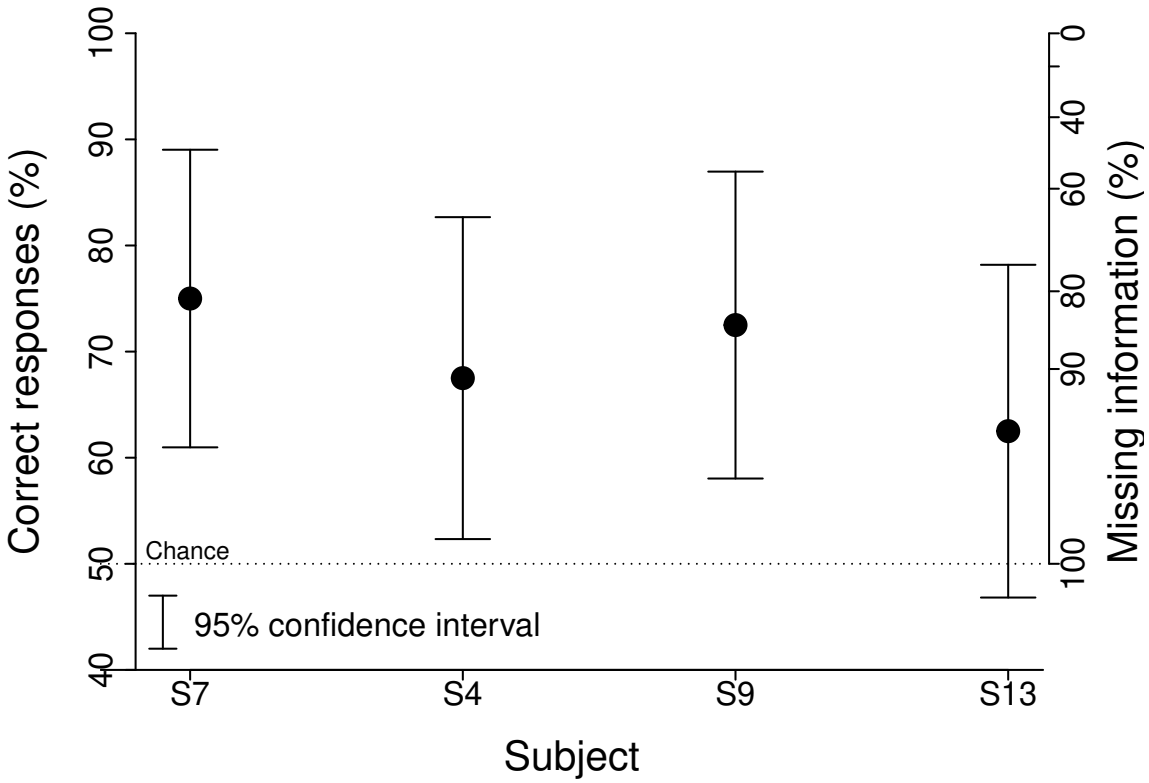

Figure 3: Speaker identification in Experiment 1 by four expert subjects (S7, S4, S9, S13) with correct responses (left) and missing information (right, 100% = 1 bit). Confidence intervals from Student distribution. Overall mean correct: 69%, 95% conf int. [61, 78]%. No differences were found in responses to male and female speakers.

3 Listening experiments

We conducted *ABX* pilot listening experiments where subjects have to identify which of the two utterances, *A* or *B*, was uttered by the speaker in *X*. These experiments were designed to test the efficacy of the proposed approach, in terms of the following questions.

1. Can experts identify a speaker from pseudonymized speech?
2. How does pseudonymization affect the reliability of speaker identification by naïve listeners?
3. How resilient is the method to re-identification?

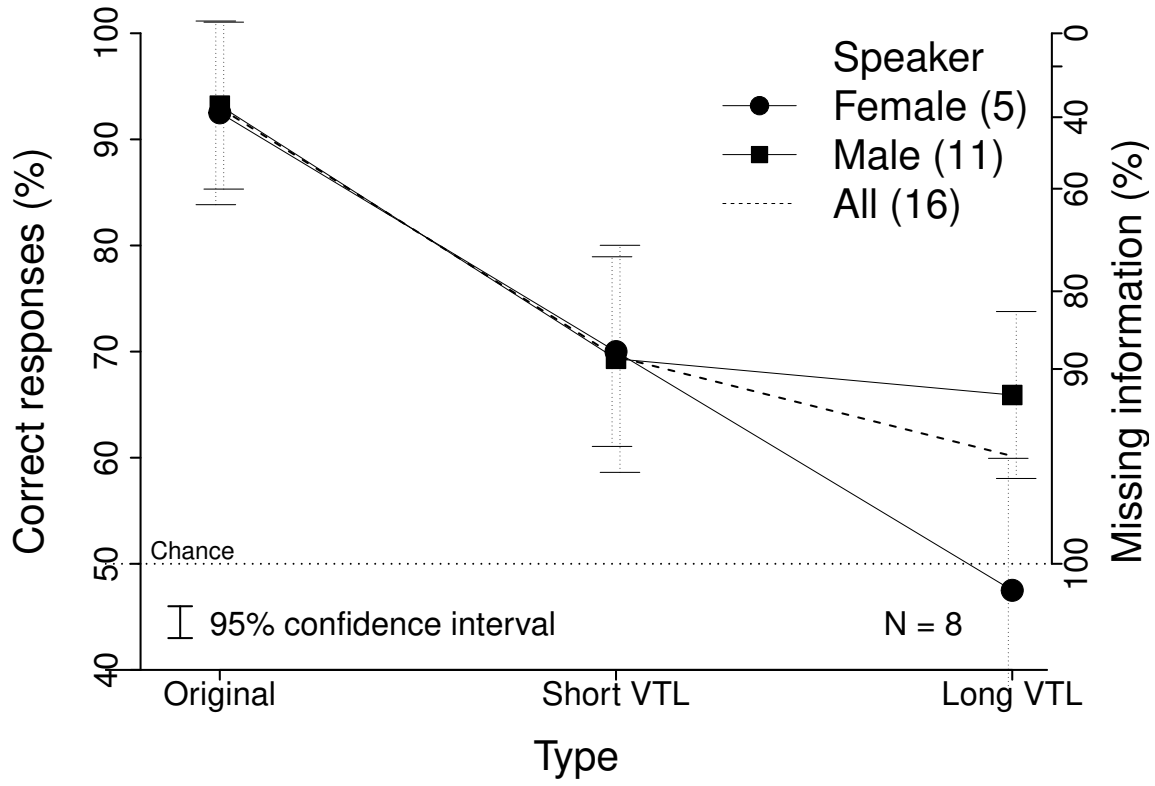

Figure 4: Speaker identification in experiment 2 by stimulus type and speaker gender. Original: AB are original recordings, Short VTL: AB pseudonymized to a short vocal tract length, Long VTL: AB pseudonymized to a long vocal tract length. N: Number of subjects. See also Fig. 3.

3.1 Experimental setup

Pseudonymized sentences and sentence fragments were produced by running the `PseudonymizeSpeech.praat` script [van Son, 2020d, van Son, 2020c] with target values for a male-like Long Vocal Tract Length (Long VTL) and a female-like Short Vocal Tract Length (Short VTL). Randomised values were used for the frequencies and intensities of bands B_0, B_{3-5} (see Section 2.1.6). Three ABX listening experiments were performed, where one choice, A or B, is uttered by the same speaker as sound X and the other is a *distractor*, see Figure 2. Fully functional offline copies of the experiments are available from [van Son, 2020b].

3.1.1 Experiment 1

Stimuli were *Pseudosentences* from the IFA corpus read by 10 Dutch speakers (5F) [Van Son et al., 2001]. In Experiment 1, the parameters of the male-like Long VTL targe are ϕ = 510Hz, F_0 = 120Hz, rate = 3.8 syll/s.; and those of the female-like Short VTL target are ϕ = 585Hz, F_0 = 185Hz, rate = 4.2 syll/s. Speaker profiles were derived from all pseudosentences read by that speaker. Long VTL and Short VTL pseudonymizations of the target speaker and a distractor were presented to 4 experts: 3 speech therapists and 1 linguist. In Experiment 1, both the X and the A and B sounds of each ABX stimulus were pseudonymized. When X was Long VTL in the ABX stimulus, A and B were Short VTL and when X was Short VTL, A and B were Long VTL. Each target speaker was presented once with a male and once with a female distractor.

3.1.2 Experiment 2

Sentence fragments with a maximum duration of 3s were selected from readings of *Treasure Island* taken from the *Parallel Audiobook Corpus* [Ribeiro, 2018] read by 16 speakers of British English (5F). In Experiment 2, the pseudonymization target values were somewhat lowered to adapt to the new corpus. The parameters of the male-like Long VTL target are ϕ = 500Hz, F_0 = 120Hz; and those of the female-like Short VTL target are ϕ = 575Hz, F_0 = 175Hz. Target speaking rate was always 4.0 syll/s. Speaker profiles were derived from all sentences in a single chapter, not used for selecting stimulus sentences. X was an original recording from the speaker to be recognized, A and B were both either Original recordings, or Long VTL or Short VTL pseudonymizations, one of which was from the same speaker as X. There were 16 ABX stimulus combinations for each condition, Original, Long VTL, and Short VTL, 48 ABX combinations in total. Each speaker was used only once as target speaker for each condition (not counting practice items). Distractors were selected at random irrespective of the gender. The genders of target speaker and distractor were the same (FF or MM) for 27 stimuli and different (FM or MF) for 21 stimuli. For this experiment, 8 "naive" listeners participated, recruited by email, not counting a subject that was dropped (see Section 3.2.2).

3.1.3 Experiment 3

All the sentences from the *Bonafide* recordings from the Logical Access part of the 2019 ASVspoof corpus [Yamagishi et al., 2019] were pseudonymized with the same pseudonymization target values as in Experiment 2. The procedure for Experiment 3

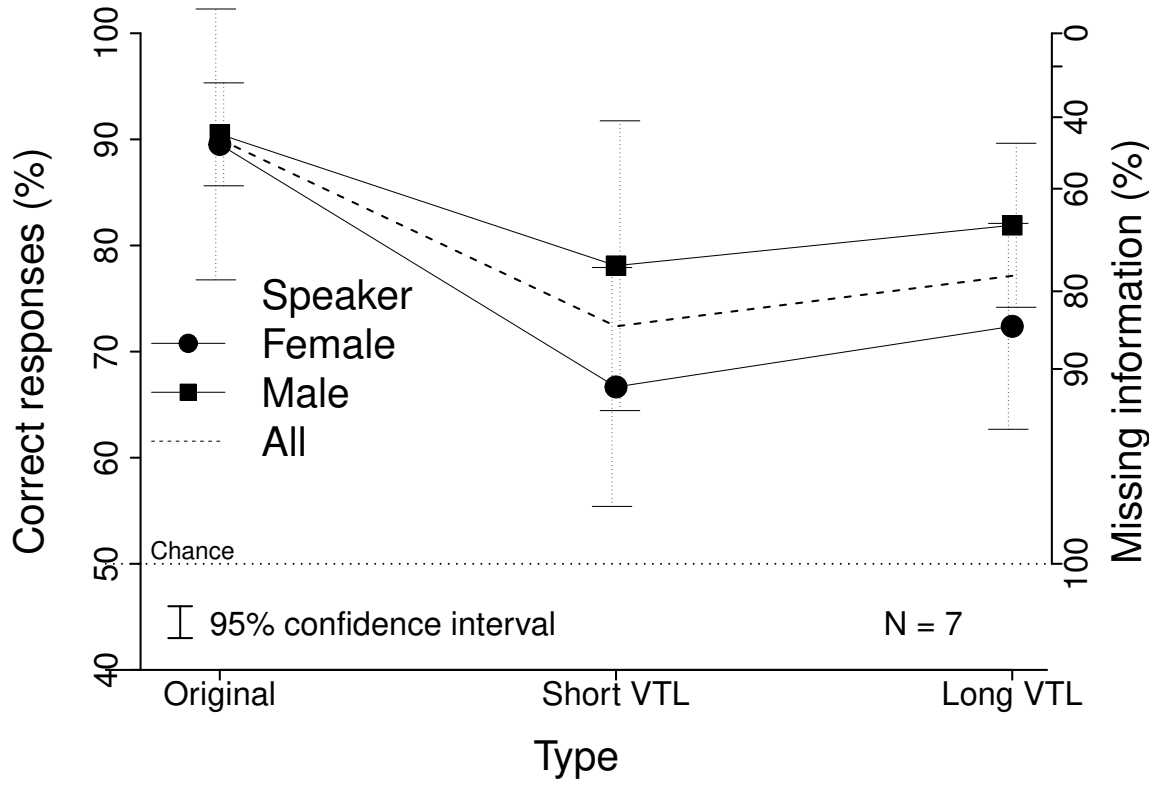

Figure 5: Identification after de-pseudonymization in experiment 3 by stimulus type and speaker gender. Speaker: Target speakers, 15F/15M for each Type, 90 in total. See also Fig. 4.

Table 1: Summary of *ABX* listening experiments. Sp.: Speakers.

Exp	**Corpus**	**Speech (≤3s)**	**Sp. F/M**	**Subjects**
1	[Van Son et al., 2001]	Pseudo sent.	5/5	4 experts
2	[Ribeiro, 2018]	sentences	5/11	8 naive
3	[Yamagishi et al., 2019]	sentences	45/45	6 naive/1 exp.

was the same as in Experiment 2. Speaker profiles were derived from all the sentences of that speaker. Gender information was available for 58 out of 107 speakers. A linear model based on the speaker profiles, with perfect fit on the known genders, was used to predict the gender of the other speakers. Sentence fragments with a maximum duration of 3s were selected as *ABX* stimuli from the target speaker and

Table 2: Speaker identification accuracy in experiments 2 and 3. Linear mixed effects models of influence of (de-) pseudonymization and speaker gender on identification for each stimulus type (see text). Ex: Experiment, p (Ex): p-value of difference between experiments, p (Gen): p-value of difference between speaker genders in combined experiment.

Stimulus	**Ex. 2 (sd)**	**Ex. 3 (sd)**	**p (Ex)**	**p (Gen)**
Original	93% (6)	90% (8)	>0.05	>0.05
Short VTL	70% (11)	73% (12)	>0.05	>0.05
Long VTL	60% (7)	77% (6)	0.009	0.012

distractor, and were all de-pseudonymized using the speaker profile of the target speaker. In the de-pseudonymization, the formant frequencies and band intensities of the transformed segments were not known (since they were randomly chosen). Therefore, only the vocal tract length, pitch, and speaking rate were transformed to the target speaker profile. Target speaker and distractor were always of the same gender, both male or female. This was done because pilot tests showed that mixed gender stimulus pairs were perfectly identified.

Each condition in Experiment 3, Original, Long VTL, and Short VTL, contained sentences from 15 male and 15 female target speakers and randomly selected distractors of the same gender, 90 *ABX* stimulus sets in total. In Experiment 3, each speaker was only used once as a target speaker and once as a distractor (not counting practice items). Sentences were selected at random from the corpus from each speaker, but no sentence recording was used twice in the experiment.

Subjects for experiments 2 and 3 were recruited over email with written instructions. Listening conditions in these two experiments were not supervised. As a quality assurance, only the responses from subjects who were able to correctly identify 70% of the target speakers in the original recordings (condition Original) were included in the analysis. Five subjects participated in both experiment 2 and 3, one subject participated in both experiments 1 and 3. Table 1 contains a summary of the three listening experiments.

3.2 Results and analysis

All the statistical analysis was done with R [R Core Team, 2019]. Missing information is calculated as the entropy $H = -\sum_{i=1}^{2} p_i \log_2 p_i$ (in %). Differences in identification between conditions and stimulus classes are tested using paired Student t-tests

(following [Fradette et al., 2003]). The stimuli and experiment are available in [van Son, 2020b] and the listener responses are available in [van Son, 2020a].

3.2.1 Experiment 1

The expert listeners reported that they found it difficult to believe that the target speaker was always among the response choices. The expert listeners identified the target speaker approximately 70% of the time (see Fig. 3, missing >80% of information H). The responses were better than chance and worse than perfect ($p \leq 0.006$ for both 90% and 50% correct, t-test, not shown). There were no statistically significant differences between listeners and no effects of speaker gender (not shown).

3.2.2 Experiment 2

Responses of one subject, who did not reach 70% correct identification on the original recordings, were dropped (subject removed, see above). On average, the speaker identification of the original recordings was over 90% correct (see Fig. 4). The naive listeners identified the target speaker approximately 70% of the time in the short VTL condition and somewhat less in the long VTL condition (missing >80% of information to identify the speaker). This was significantly less than in the original condition with unaltered speech ($p \leq 0.0001$, paired t-test by subject). The difference between the short and long VTL condition were not significant ($p > 0.05$). There is a tentative difference in responses to the (5) female speakers and the (11) male speakers for the Long VTL stimuli ($p = 0.027$, paired t-test). It appears that the female speakers are not identified above chance level in the long VTL condition. There seems to be an asymmetry in the effect of pseudonymization on male and female voices which we currently cannot explain.

In the responses from experiments 1 and 2, there is a tendency that comparison to a distractor of a different gender improves identification of the target speaker (not shown). However, partly due to the design of the experiments, this could not be verified ($p > 0.05$, paired t-test).

3.2.3 Experiment 3

All the subjects cleared the 70% correct criterion for the Original stimulus condition. Speaker identification of the original recordings in Experiment 3 was around 90% correct (see Fig. 5 and Table 2). De-pseudonymization, the inverse transform, was effective in reversing the pseudonymization towards a Long VTL target, increasing the identification from 60% to 78% correct (Table 2) with missing information $\leq$80%

(Fig. 5). However, the differences in identification between the Original and the de-pseudonymized stimuli was still significant ($p \leq 0.009$ paired t-test by subject). The difference in identification between male and female speakers was not significant in Experiment 3 ($p > 0.05$ for all stimulus types).

3.3 Modeling responses to listening experiments 2 & 3

The results of experiments 2 and 3 were combined in a linear mixed effect model to estimate the effects of the speaker gender and pseudonymization versus de-pseudonymization (Exp) on speaker Identification (I) for each stimulus type, i.e., Original, Short VTL and Long VTL. The full model was:

$$I \sim Exp + Gender + (Exp + Gender|Subject) \tag{3}$$

Subjects that participated in both the experiments were identified in the model. Statistical significance was determined using ANOVA on full model versus a model with the relevant fixed factor removed. No difference was found for the Original and Short VTL stimuli ($p > 0.05$). For the Long VTL target pseudonymizations, both the differences between male and female speakers and the differences between the experiments were statistically significant (see Table 2). Using the model of Eq. 3, the male speakers were identified 13% points more than female speakers and de-pseudonymization increased identification by 19% points (p-values in Table 2).

Experiment 3 only contained same gender comparisons between the target and distractor speakers, while Experiment 2 contained the same and mixed gender comparisons. Same gender comparisons could be seen as "more difficult" than mixed gender comparisons. Repeating the linear mixed effect modelling with only the responses to the same gender distractors gave the same results; no effect for Original and Short VTL stimuli ($p > 0.05$) and a consistent effect for de-pseudonymization and speaker gender for Long VTL stimuli ($p(Ex) = 0.008$, $p(Gen) = 0.024$, not shown) were observed. But the effect of de-pseudonymization only increases marginally (to 22% points). The overall effect of de-pseudonymization was found for both female and male speakers separately ($p \leq 0.012$, ANOVA, removing *Gender* from Eq. 3, not shown).

4 2020 VoicePrivacy Challenge experiments

Automatic evaluations of the proposed method were carried out as part of the VoicePrivacy 2020 challenge, using the data sets and experimental protocols set by

the challenge [Tomashenko et al., 2020a], and the performances were measured in terms of ASV and ASR systems' evaluation metrics. The ASV evaluation consists of an *enrollment* phase, where several speakers enrol into a system, and a *trials* phase, where each test speaker that claims to be a specific enrolled speaker has to be verified. For the anonymization experiment, each speaker is anonymized to two different speakers, one for enrollment and another for trials. Thus, a good anonymizer would increase the ASV error, while keeping the ASR error as low as possible.

4.1 Summary of the data sets and evaluation protocol

For evaluations on anonymization, the `dev` and `test` subsets of the VCTK and LibriSpeech corpora were used. As reference (and non-overlapping) speaker set for anonymization, `libri-other-500` subset of LibriSpeech was used. The anonymized speech is evaluated in terms of word error rate (WER) using a neural-network based ASR system trained with lattice-free maximum mutual information objective function [Povey et al., 2016] and in terms of equal error rate (EER) and log-likelihood ratio based costs C_{llr} and C_{llr}^{min} using an x-vector [Snyder et al., 2018] based ASV system, both provided by the challenge organisers. For more details about the data set and the experimental protocol, the reader is referred to [Tomashenko et al., 2020a].

4.2 Baselines provided by the challenge

The challenge provided two baseline systems: (i) neural source filtering (NSF) based and (ii) McAdams method based.

4.2.1 NSF baseline

The NSF approach [Fang et al., 2019] was built on the idea that any speech signal can be decomposed into three sets of features: those representing (i) the spoken content, (ii) the speaker and (iii) the speaker's fundamental frequency, and that speech can be synthesised back by combining these components. Mel-filterbank features or intermediate representations from an ASR neural acoustic model constitute the spoken content, whereas fixed length neural speaker embeddings, known as x-vectors, represent the speakers. Anonymization can be achieved by merely replacing the source speaker's x-vectors with those of the target speaker, which is chosen among a pool of reference speakers, typically the one who is *farthest* in terms of their x-vector *affinity* score. Thus, in this method, a given speaker's speech is first decomposed into its three constituents, then anonymized by replacing the x-vectors and finally converted back into a waveform using speech synthesis.

Table 3: ASV results for both development and test partitions (o-original, p-pseudonymized(F03-9), b1-NSF. b2-McAdams).

Data	Expt.	Dev. set (female)			Dev set (male)			Test. set (female)			Test set (male)		
		EER%	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$	**EER%**	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$	**EER%**	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$	**EER%**	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$
libri	o	8.67	0.30	42.86	1.24	0.03	14.25	7.67	0.18	26.79	1.11	0.04	15.30
	b1	36.79	0.89	16.35	34.16	0.87	24.72	32.12	0.84	16.27	36.75	0.90	33.93
	b2	23.44	0.62	11.73	10.56	0.36	11.95	15.33	0.49	12.55	8.24	0.26	15.38
	p	25.28	0.66	9.30	18.79	0.56	15.70	24.82	0.59	10.23	14.92	0.43	10.65
vctk common	o	2.33	0.09	0.86	1.43	0.05	1.54	2.89	0.09	0.87	1.13	0.04	1.04
	b1	27.91	0.74	7.21	33.33	0.84	23.89	31.20	0.83	9.02	31.07	0.84	21.68
	b2	11.63	0.37	43.55	10.54	0.32	25.00	14.45	0.47	42.73	11.86	0.35	28.23
	p	16.86	0.51	11.12	20.23	0.56	7.65	26.01	0.70	13.16	13.84	0.45	5.32
vctk different	o	2.86	0.10	1.14	1.39	0.05	1.16	4.94	0.17	1.50	2.07	0.07	1.82
	b1	26.11	0.76	8.41	30.92	0.84	23.80	31.74	0.85	11.53	30.94	0.83	23.84
	b2	15.83	0.50	39.81	11.22	0.38	23.09	16.92	0.55	41.34	12.23	0.40	25.06
	p	15.67	0.50	6.25	14.74	0.39	3.84	26.23	0.75	11.92	22.90	0.67	7.57

4.2.2 McAdams baseline

This is a signal processing method based on formant shifting. In this method, each utterance is analysed using short-time processing, where each segment is fit with an all-pole model on its power spectrum using linear prediction. The angles θ_i of the complex poles thus correspond to the formant frequencies, when the model order is appropriately chosen. The anonymization process involves shifting the formants non-linearly, by exponentiating the complex poles by a constant factor M, i.e. $\theta_i \rightarrow \theta_i^M$, where M is the McAdams coefficient. The resultant signal is then overlap-added across segments to reconstruct its corresponding pseudonymized utterance.

Contrasting the McAdams method with our proposed approach, a key difference is that the former allows only a single degree of freedom (i.e. by tuning M) in moving the formants, whereas the proposed approach allows each band of formants and F0 to be individually moved and their intensities adjusted, thus allowing several degrees of freedom.

4.3 Idiap-NKI Challenge entry

We followed the protocol set by the challenge, and evaluated ASR and ASV performances by pseudonymizing the given subsets of VCTK and LibriSpeech data sets. For pseudonymization, target speaker profiles were created using `libri-other-500` set

Table 4: ASR results in WER% for both development and test partitions (o-original, b1-NSF, b2-McAdams, p-pseudonymized(F03-9), s-LM_s, l-LM_l), and results of subjective tests (median scores with per-listener normalization, 0-1) for Nat-naturalness, and Int-intelligibility [Wang et al., 2020].

Expt.	**libri**				**vctk**				**subjective tests**	
	Dev. set		**Test set**		**Dev. set**		**Test set**			
	s	**l**	**s**	**l**	**s**	**l**	**s**	**l**	**Nat**	**Int**
o	5.24	3.84	5.55	4.17	14.00	10.78	16.38	12.80	0.74	0.70
b1	8.76	6.39	9.15	6.73	18.92	15.38	18.88	15.23	0.29	0.34
b2	12.19	8.77	11.77	8.88	30.10	25.56	33.25	28.22	0.31	0.36
p	8.82	6.48	8.04	5.87	21.99	18.23	23.32	18.89	0.38	0.39

of the LibriSpeech corpus. In a given subset, each speaker is pseudonymized to have the characteristics of a randomly chosen target speaker from the `libri-other-500` set. In ASV, this means that the enrollment and trials of the same speaker are often mapped to different target speakers (and we have not ensured that they are different in all the cases, since the probability of choosing the same speaker among 1000+ speakers is small). If only the trials sets are pseudonymized, ASV may indicate a higher error (indicating a better anonymization) due to acoustic mismatch introduced by the pseudonymization method. A higher equal error rate (EER) in ASV implies better pseudonymization of the speakers, and a lower WER on ASR implies better preserving of intelligibility.

The proposed method uses all the steps presented in Section 2.1. That is, the method changes the speaking rate, pitch and the B_0 and B_{3-5} bands and their intensities. The target values for pseudonymization are determined by selecting a random speaker from `libri-other-500` as the target speaker. In addition, the B_4 and B_5 bands are switched, and bands B_{6-9} are replaced with intensity modulated pink noise. For the sake of clarity, this pseudonymization method is referred to as *F03-9*.

4.4 Results

Tables 3 and 4 compare the ASV and ASR results, respectively, of the baseline anonymization methods using neural source-filtering (NSF) and McAdams, and the proposed pseudonymization method. In ASR, the proposed method gave a lower WER than the McAdams baseline, indicating better intelligibility, in all the cases.

Table 5: ASV results with ablation (pseudon - pseudonymized(F03-9), pseudon(2) - a repeat of 'pseudon' with a different random speaker set, after ensuring that all the enrollment and evaluation sets are pseudonymized to different speakers, no system - only source and speaking rate have been modified, no source - only system and speaking rate have been modified, no rate - only source and system have been modified, no-additional - no additional processing described in Sec. 2.1.10 has been applied, de-pseudon - pseudonymization reversed (the right portion of Fig. 1) for the *no-additional* experiment.).

Data	**Expt.**	**Dev. set (female)**			**Dev set (male)**			**Test. set (female)**			**Test set (male)**		
		EER%	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$	**EER%**	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$	**EER%**	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$	**EER%**	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$
	pseudon	25.28	0.66	9.30	18.79	0.56	15.70	24.82	0.59	10.23	14.92	0.43	10.65
	pseudon(2)	31.82	0.80	12.39	14.13	0.44	9.04	25.18	0.67	11.48	15.81	0.48	12.18
	no system	15.91	0.49	42.74	5.12	0.17	36.62	10.77	0.33	39.27	2.00	0.07	25.45
libri	no source	20.88	0.61	7.24	15.22	0.48	12.38	19.71	0.55	6.93	8.46	0.27	3.25
	no rate	21.16	0.61	7.48	15.22	0.49	12.39	19.53	0.56	7.47	8.46	0.28	3.30
	no additional	15.48	0.48	42.54	5.12	0.17	36.60	10.77	0.33	39.25	2.23	0.07	25.66
	de-pseudon	10.37	0.34	33.84	3.11	0.10	23.38	10.40	0.30	30.87	1.78	0.06	18.23
	pseudon	16.86	0.51	11.12	20.23	0.56	7.65	26.01	0.70	13.16	13.84	0.45	5.32
	pseudon(2)	24.42	0.67	14.56	25.36	0.72	10.71	25.43	0.70	10.01	15.82	0.46	4.67
vctk	no system	17.73	0.51	8.83	12.25	0.38	14.34	13.58	0.45	9.03	8.76	0.26	12.51
	no source	23.55	0.58	13.82	19.94	0.58	8.59	25.43	0.74	19.02	20.62	0.59	8.05
common	no rate	21.80	0.58	15.06	19.09	0.57	8.33	25.14	0.75	19.42	20.34	0.56	7.68
	no additional	17.44	0.51	8.88	12.25	0.37	14.19	14.45	0.45	9.13	9.04	0.26	12.50
	de-pseudon	6.40	0.24	2.14	8.55	0.22	5.48	7.23	0.22	1.48	4.52	0.13	3.83
	pseudon	15.67	0.50	6.25	14.74	0.39	3.84	26.23	0.75	11.92	22.90	0.67	7.57
	pseudon(2)	28.86	0.76	11.66	22.73	0.67	8.62	34.05	0.85	11.61	22.22	0.62	7.05
vctk	no system	17.97	0.52	10.79	2.33	0.09	1.14	14.87	0.45	5.98	11.83	0.35	14.64
	no source	27.68	0.70	11.78	5.11	0.18	3.09	22.27	0.65	12.90	27.55	0.68	12.81
different	no rate	24.48	0.66	11.36	5.26	0.19	3.15	21.35	0.63	12.03	24.11	0.61	9.48
	no additional	17.57	0.51	10.63	2.38	0.10	1.16	14.40	0.44	5.84	12.34	0.36	14.77
	de-pseudon	5.90	0.21	1.39	2.28	0.09	0.66	10.19	0.34	2.69	6.83	0.22	5.03

In ASV, the EER in all the cases except one (vctk-different female) is higher, implying a better pseudonymization, than the McAdams baseline. This is also indicated by a consistently higher or equal C_{llr}^{min} in all the cases. However, there is a room for improvement in comparison to the NSF baseline in terms of ASV performance, although it is fairer to compare the method with the signal processing based baseline.

We conducted ablation experiments to study the contribution of the individual steps proposed in Sec. 2, to study the effect of speaker selection in the proposed

Table 6: ASR results in WER% with ablation (pseudon - pseudonymized(F03-9), pseudon(2) - a repeat of 'pseudon' with a different random speaker set, after ensuring that all the enrollment and evaluation sets are pseudonymized to different speakers, no system - only source and speaking rate have been modified, no source - only system and speaking rate have been modified, no rate - only source and system have been modified, no-additional - no additional processing described in Sec. 2.1.10 has been applied, de-pseudon - pseudonymization reversed for the *no-additional* experiment, s-LM_s, l-LM_l).

Expt.	**libri**				**vctk**			
	Dev. set		**Test set**		**Dev. set**		**Test set**	
	s	**l**	**s**	**l**	**s**	**l**	**s**	**l**
pseudon	8.82	6.48	8.04	5.87	21.99	18.23	23.32	18.89
pseudon(2)	8.59	6.14	8.16	5.85	20.52	16.89	23.69	19.50
no system	7.30	5.21	6.87	5.07	18.00	14.34	20.38	16.42
no source	8.14	5.93	7.62	5.64	20.12	16.31	22.83	18.81
no rate	7.72	5.63	7.24	5.31	18.90	15.01	21.97	17.72
no additional	7.18	5.18	6.90	5.08	18.05	14.32	20.36	16.41
de-pseudon	6.85	4.95	7.03	5.27	17.43	13.61	20.37	16.08

approach and to study the effect of de-pseudonymization (the right part of Fig. 1). The individual steps of pseudonymization are: (i) the *source* part: pseudonymizing B_0 band, (ii) the vocal-tract *system* part: pseudonymizing the B_{3-9} bands, which also includes the *additional* processing of introducing modulated pink noise in B_{6-9} bands (Sec. 2.1.10) and exchanging B_4 and B_5 bands, and (iii) the *speaking rate* part. To study the effect of speaker selection, we repeated the proposed approach using a different set of random speakers, after ensuring that the enrollment and evaluation target speakers are always different. To be able to perform de-pseudonymization, we had to omit the irreversible additional processing step. Tables 5 and 6 show the results of all the ablation experiments. The results indicate that the vocal-tract system component plays the most prominent role in pseudonymization, and a significant part of it is due to the additional processing. The repeat study of the proposed approach indicates that ensuring different target speakers in enrollment and evaluation improved the ASV results in most cases, with two cases improving over those of the NSF method in Table. 3. However, a degradation can be seen in a few cases. This variability can be attributed to the differences in the target

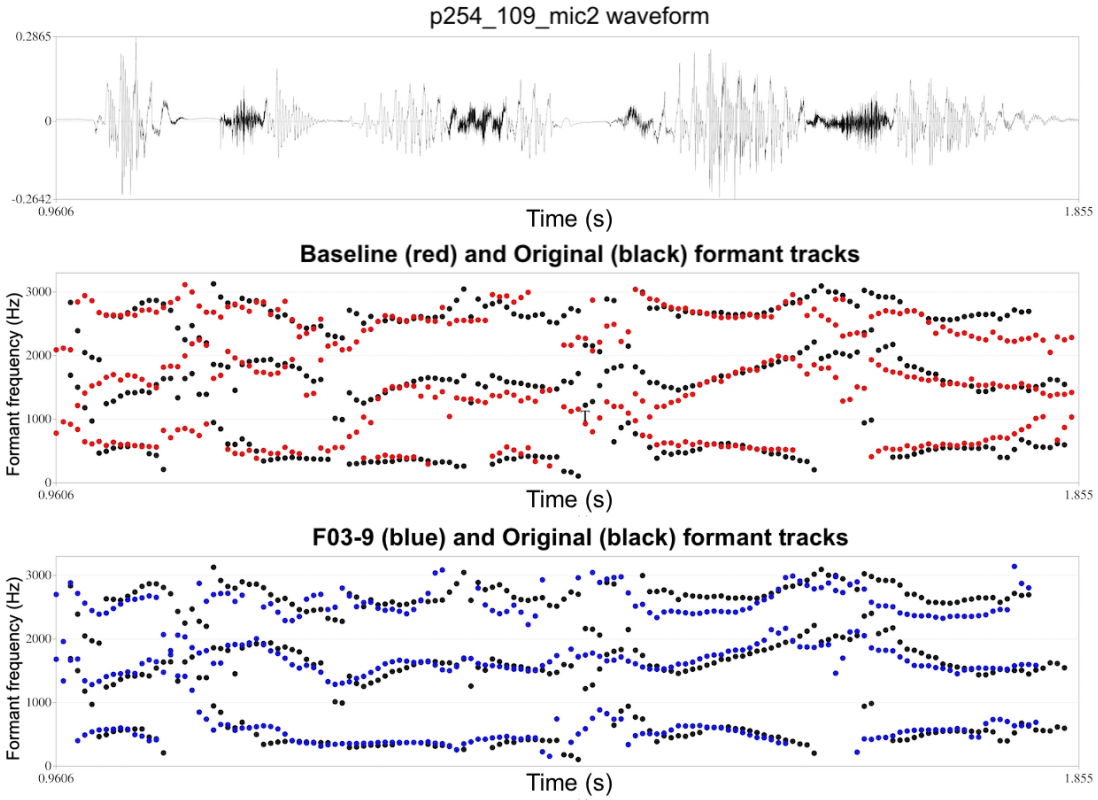

Figure 6: Example formant tracks for correlating formant values between pseudonymized speech and the original recordings. Top: waveform of sentence [but it is a pleasure] from speaker p254, center: F_1-F_3 formant tracks for McAdams Baseline (red) and Original (black) speech, bottom: id. for F03-9 pseudonymization (blue) and Original (black). Horizontal: Time, Vertical: Amplitude (top) and Frequency (mid and bottom).

speaker selection, indicating its importance in improving the ASV performance of the proposed approach, and this could be a good future direction. It can be seen that ASR performance is less sensitive to the target speaker selection. Finally, de-pseudonymization can be seen to be partially effective.

5 Beyond the VoicePrivacy challenge

In this section, we explore some directions in which the proposed method could be evaluated, viz. (i) measuring the intelligibility after pseudonymization by utilising the original utterances as references, instead of ASR, (ii) measuring the closeness of the formant tracks between the pseudonymized and original utterances, and (iii) measuring the extent of retaining pathological conditions such as dysarthria after the proposed pseudonymization.

5.1 Intelligibility measure based comparison of phone posterior sequences

The 2020 VoicePrivacy Challenge proposed WER of ASR system as a measure of intelligibility. However, ASR system performance gets affected by components such as Viterbi search, language model and pronunciation lexicon. Even if we presume that all the anonymization systems are compared using exactly the same language model, acoustic model and lexicon, the search heuristics can make a difference. So, here we propose to utilise an alternate objective intelligibility measure where the original reference speech and the anonymized speech are compared in the phone posterior feature space, employing only the acoustic model. First demonstrated in the context of using synthetic speech for template-based ASR using posterior features [Soldo et al., 2012] and then extended to speech intelligibility assessment in [Ullmann et al., 2015], the method consists of estimating sequences of phone posterior probabilities corresponding to the reference speech and the test speech, and comparing the two sequences using DTW with a local score based on Kullback Leibler (KL) divergence [Soldo et al., 2011, Soldo et al., 2012].

In this paper, the local score d_{jt} is computed as

$$d_{j,t} = \mathbb{KL}(\mathbf{y}_j \parallel \mathbf{z}_t) = \sum_{k=1}^{K} y_j^k \log\left(\frac{y_j^k}{z_t^k}\right) , \tag{4}$$

where $\mathbf{y}_j = [y_j^1, y_j^2, ..., y_j^K]$ denotes the phone posterior feature vector at time j in the reference sequence of length J, $\mathbf{z}_t = [z_t^1, z_t^2, ..., z_t^K]$ denotes the phone posterior feature vector at time frame t in the test sequence of length T and K denotes the number of phones.

We used the following dynamic programming recursion

$$D_{j,t} = d_{j,t} + \min\left(D_{j,t-1}, D_{j-1,t-1}, D_{j-2,t-1}\right) , \tag{5}$$

Table 7: Intelligibility in terms of DTW distances (b1-NSF, b2-McAdams, p-pseudonymized(F03-9)).

E	**libri**		**vctk**	
	Dev.	**Test**	**Dev.**	**Test**
b1	0.005650	0.005804	0.007050	0.007638
b2	0.008798	0.008082	0.010237	0.010273
p	0.005955	0.004463	0.006001	0.006100

where $D_{j,t}$ denotes the cumulative score at j, t. The additional skip transition from $D_{j-2,t-1}$ was allowed to accommodate for the duration changes between the reference and test utterances. The final score $D_{J,T}$ normalised by the path length yields a measure of intelligibility; the lower the score, the better the intelligibility.

We computed intelligibility scores in the following manner:

1. First, estimate the posterior probability of the clustered context dependent phones using the neural network-based acoustic model provided with the VoicePrivacy challenge and then marginalise the context-dependent information, position markers and lexical stress markers to estimate the posterior probabilities of context independent phones. The context independent phone posteriors are used as the posterior features, $\mathbf{y}_j$ and $\mathbf{z}_t$ for the DTW-based intelligibility score estimation.

2. Compare the intelligibility scores (DTW distances) for the proposed pseudonymization method (F03-9) and the baseline methods by averaging the scores of all the utterances in each method.

Results from Table 7 indicate that the intelligibility scores for the proposed pseudonymization method are comparable to those of the NSF baseline and better than the McAdams baseline. This indicates that the differences observed in the WER metric (Table 4) could be due to aspects such as search heuristics employed during decoding.

5.2 Measuring pseudonymized formant values

Formants are important in the study of speech because their values are linked to the shape of the vocal tract, and hence to the constellation and movements of the articulators [Dromey et al., 2013, Lee et al., 2015, McKell, 2016, Christensen, 2018].

Formant values are also related to the intelligibility of phonetic contrasts [Kent et al., 1989, Harper et al., 2017, Richardson and Sussman, 2017]. These relations are also relevant to the study of pathological speech, such as dysarthric speech [Sapir et al., 2010] and Parkinson's disease [Sapir et al., 2007]. To evaluate to what extent formant measurements can be preserved after pseudonymization, formant tracks before and after pseudonymization are compared (see Figure 6). To determine the preservation of formant tracks after pseudonymization, the first three formant tracks of pseudonymized speech samples are correlated to those of the original recordings, using the *Robust* formant tracking in *Praat* [Boersma and Weenink, 2019]. The same recordings from 60 speakers (30F/30M from vctk_dev and vctk_test) were used for the McAdams *Baseline* and *F03-9* pseudonymization. A higher average correlation coefficient (R) indicates that the pseudonymized speech would be more useful to study the acoustic effects of differences in articulation.

The results of the comparison are presented in Table 8. These results show that the average R of the pseudonymized formant values are higher for the *F03-9* pseudonymizations than for the *Baseline* method for all three formants. Correlation coefficients, R, for the *Baseline* method were between R=0.26 and R=0.60. Correlation coefficients for the *F03-9* method were 0.1-0.3 higher on average for all speakers, between R=0.56 and R=0.72 (R^2: 0.12-0.31 higher, highest values for F_3, $p \leq 10^{-7}$, paired Student t-test per speaker). There was a difference based on the speaker gender. For female speakers, the difference in R was 0.05-0.20 (highest values for F_3, $p \leq 10^{-2}$, *idem*), for male speakers, it was 0.14-0.42 (highest values for F_3, $p \leq 10^{-5}$, *idem*). The differences in R between *Baseline* and *F03-9* were larger for male than for female speakers for all three formants (two sample Student-t test, $p \leq 0.001$, 0.01, and 10^{-6} for F_1 - F_3, respectively).

5.3 Automatic dysarthria classification

The ability to investigate paralinguistic features after pseudonymization was evaluated on the example of dysarthric speech. Speech samples were taken from the TORGO corpus [Rudzicz et al., 2012]. The recordings from the head mounted microphone were used. Recordings from the directional microphone were added for two sessions, both session 2 of control speakers FC02 and MC04.

The control and dysarthric utterance recordings were pseudonymized as with the F03-9 method described above. However, for this experiment, the characteristics of a random speaker of the opposite gender was selected from the *Bonafide* recordings in the Logical Access part of the 2019 ASVspoof corpus [Yamagishi et al., 2019]. As altered, slow, speaking rate is an important symptom of dysarthria, the speaking

Table 8: Mean correlation coeff, R (SD), between formant tracks from Original and pseudonymized recordings, for all speakers (N=60). F_1, F_2, F_3: Correlation coefficients of the formants. F: Female speakers, M: Male speakers. Baseline (McAdams) & *F03-9*: Pseudonymization procedures, see text. Average number of recordings per speaker: 23.5±13.7 (F), 24.2±15.8 (M).

Group		F_1	F_2	F_3
F	Baseline	0.507 (0.158)	0.601 (0.198)	0.424 (0.287)
	F03-9	0.563 (0.194)	0.659 (0.161)	0.620 (0.202)
M	Baseline	0.490 (0.161)	0.571 (0.158)	0.264 (0.226)
	F03-9	0.655 (0.153)	0.716 (0.136)	0.688 (0.136)
Total	Baseline	0.499 (0.160)	0.586 (0.178)	0.344 (0.257)
	F03-9	0.609 (0.174)	0.688 (0.149)	0.654 (0.169)

rate of the pseudonymized utterances was not changed from the original value. The results of the ablation experiment in Section 4.4 show that not changing the speaking rate has only a low impact on ASV identification performance (see Table 5). The dysarthria classification was done with linear support vector machines (SVMs) trained, using a leave-one-out procedure, on eGeMAPS feature set that is commonly used in studying paralinguistic aspects [Eyben et al., 2016]. SVMs trained on the original recordings were used to classify the original utterances, whereas those trained on pseudonymized recordings were used to classify the pseudonymized utterances.

The dysarthria classifier did not perform very well (59% correct). Inspection of the results showed that this was most likely due to the low audio quality of some sessions. It also seemed to perform worse on some of the female speakers. It was decided to drop all sessions where dysarthria classification of the original recordings was below 70% correct. This left 15 (out of 30) recording sessions from a total of 10 (out of 15) speakers, 5 control (2F) and 5 dysarthric (1F) speakers. The two sessions recorded with the directional microphone were among those dropped for low classification performance.

The audio quality of the remaining utterance recordings was characterised by measuring the signal-to-noise ratio (SNR): the difference between the maximum and minimum intensities in each utterance (measured in 50ms bins). The SNR of the recordings from the control speakers, range [17-32] dB, was clearly lower than those of the dysarthric speakers, range [41-51] dB (see Table 9).

The results of the dysarthria classification evaluation after pseudonymization are

Table 9: Dysarthria classification results for original and pseudonymized recordings from the TORGO corpus [Rudzicz et al., 2012] (see text). Given are the percentage correct classification for the original and pseudonymized (Pseud.) recordings, the concordance (Conc.), i.e., the percentage identical classification for original and pseudonymized recordings. The overall Cronbach's alpha is acceptable, α=0.769. Without the two female control speakers FC01 and FC02, Cronbach's alpha is excellent, α=0.949. Spkr: Speaker, Pseud.: Pseudonymized recordings, Conc.: Concordance, N: number of utterances, SNR: mean Signal to Noise Ratio (dB) per utterance.

		Correct				
Group	Spkr	Original	Pseud.	Conc.	N	SNR
Control	FC01	98.2	47.6	49.4	164	25
	FC02	86.3	13.7	24.4	1000	32
	MC01	98.5	99.3	98.4	748	26
	MC02	99.1	98.7	98.3	464	26
	MC03	99.3	100.0	99.3	600	17
Dysarthric	F01	90.2	90.9	90.2	132	41
	M01	92.7	99.7	92.4	288	51
	M02	95.8	98.5	95.8	409	46
	M04	91.3	97.9	91.5	424	40
	M05	91.0	93.6	87.5	488	42
Total		94.2	84.0	82.7	471.7	35

mixed (Table 9). For the two female control speakers, FC01 and FC02, the performance is below 50%, at chance level. It is clear that the pseudonymization of utterances from these speakers degraded the speech too much and the classifier did not work. The results for the speech of the other speakers is excellent. This is summarised in the Cronbach's alpha values, which are acceptable for the whole group of 10 speakers (α=0.769), but are excellent (α=0.949) when the results for FC01 and FC02 are removed. It is currently unknown why the dysarthria classification did not work for the pseudonymized recordings of speakers FC01 and FC02. The audio quality of these recordings is not different from those of the male control speakers, and the classification performance for the original recordings does not deviate much from those of the other recordings. An analysis with low-SNR (SNR $<$ 30dB) utterances removed did not improve the results for FC01 and FC02 and did not affect the

results of the other speakers (not shown).

6 Discussion

Pseudonymization aims at protecting the privacy of the speakers. Whether or not the levels of protection are sufficient depends on the requirements of the application and the risks that an identification would pose. One objective of the proposed approach is to make pseudonymization deterministic and adjustable, i.e., gradual, on untranscribed recordings. It works in the spectro-temporal domain on any speech recording, and is intrinsically deterministic and reversible. The exception to reversibility is the overlap-add procedure to adapt the pitch and duration of the speech which is inherently "lossy", i.e., partially irreversible. But overlap-add is a well known, predictable, speech synthesis procedure. The aspects of the speech that are transformed as well as the extent of the changes can all be freely chosen. The only constraint is the quality of the resulting speech.

However, reversibility is not necessarily an advantage. It is clear that the ability to, partially, de-pseudonymize speech warrants extra attention. The current study explores one specific de-pseudonymization approach based on knowing the original pseudonymization target. An obvious way to prevent de-pseudonymization would be to obfuscate the target speaker selection.

Another important goal of pseudonymization of speech could be to allow the study of linguistic and paralinguistic aspects of speech without jeopardising the privacy of the speakers. It is not yet known which of such aspects can still be studied after pseudonymization and what the corresponding risks of re-identification are. In this study, the extent to which linguistic and paralinguistic features are preserved was estimated by comparing formant tracks after pseudonymization with the originals and by evaluating the results of an automatic dysarthria classifier on pseudonymized speech.

6.1 Listening experiments

All three *ABX* listening experiments showed reduced speaker identification after pseudonymization (Fig. 3 and 4) and also after de-pseudonymization (Fig. 5). After pseudonymization, more than 80% of the information necessary to make the choice between speaker *A* and *B* is lost (<70% correct identification, Fig. 3 and 4), compared to less than 40% missing information with the original recordings (>90% identification, Table 2). Reverting the transformation from known pseudonymization

targets can improve the recognition, especially for speech transformed to a Long VTL (Fig. 5 and Table 2).

The responses in both experiments 2 and 3 displayed an asymmetry between male and female voices. Female speakers were identified worse than male speakers after both pseudonymization and de-pseudonymization. This difference was statistically significant for the Long VTL condition when the responses in these experiments are combined (Table 2). This asymmetry was smaller, or absent, in the Short VTL condition (statistically not significant).

6.2 Automatic evaluations

Automatic evaluations on the VoicePrivacy challenge data showed that the method is better than the comparable signal-processing based McAdams method. However, there is still a significant gap in terms of ASV performance w.r.t. the NSF baseline. One factor could be that the former chooses the target speakers randomly, whereas the latter specifically chooses far away speakers. Future investigations could focus on identifying the areas of improvement that lead to closing this gap and improving beyond it. In terms of preserving the intelligibility, the proposed method showed comparable performance in terms of both the ASR and the proposed phone posterior based approach.

It is worth mentioning that, in the VoicePrivacy challenge, besides the objective evaluations, the organisers also conducted subjective evaluations, in which the proposed method showed promising results in terms of intelligibility, quality and dissimilarity of the pseudonymized speech w.r.t. the original speakers [Wang et al., 2020][4].

6.3 Formant values

The outcomes of the formant track analysis indicate that both the *Baseline* and the*F03-9* method preserve formant tracks to some extent. The *F03-9* pseudonymization better preserves F_{1-3} formant track movements than the McAdams *Baseline* method, sometimes with a considerable margin. The differences were more pronounced for male than for female speakers. The biggest differences were found in the F_3 tracks.

From these results, it is clear that it is possible to preserve at least some level of measurable formant track information after pseudonymization. However, there are

[4]We cite the presentation as it was the only reference available at the time of the submission of this article.

systematic differences between the two methods tested and the gender of the speakers in how well the formant track information is preserved. This shows that there is still room to optimise this feature in future speech pseudonymization methods.

6.4 Dysarthria classification

The TORGO corpus proved to be sub-optimal for the evaluation of automatic dysarthria classification of pseudonymized speech. Half of the recording sessions, including all recordings of 5 speakers, had to be dropped due to very low classification performance. For 8 out of the remaining 10 speakers, classification after pseudonymization performed excellent, with high concordance between original and pseudonymized audio. For two other speakers, the results after pseudonymization were essentially at chance level. What this shows is that there is indeed good potential to use pseudonymization to study paralinguistic aspects of (pathological) speech, at least for dysarthria. However, the pseudonymization method used in this study cannot yet be applied to all speakers.

7 Conclusions

A method to pseudonymize speech is described that is both deterministic and adjustable. The method can pseudonymize speech samples with only a few hundred seconds of speech of the source speaker by altering the voice source related, vocal tract system related and speaking rate information. *ABX* pilot listening tests demonstrated that the pseudonymized samples are largely unidentifiable for human listeners. However, the deterministic nature of the procedure compels caution and measures to counter re-identification should be considered before applying the procedure. An evaluation at the 2020 VoicePrivacy challenge showed that the method pseudonymizes utterances better than the McAdams method provided by the challenge and is inferior to the neural source-filter based baseline. However, in terms of a phone posterior feature-based intelligibility measure computed using only the acoustic model, the proposed method is comparable to the neural source-filter based baseline. Ablation studies analysing the role of different processing steps in the proposed approach revealed that the alteration of vocal tract system related information and the target speaker selection play a major role in anonymizing the speaker's identity. Furthermore, the studies also revealed that the pseudonymization process can be partially reversed, assuming the target speaker information such as, VTL, formant information are computable. A formant track analysis investigating the preservation of articulatory information in pseudonymized speech showed promising

results with a somewhat better correlation between the original and pseudonymized speech for the proposed method than for the baseline McAdams approach. Finally, in a case study on dysarthria, it was found that pathological speech evaluation after pseudonymization could be feasible; the results were, however, speaker dependent.

8 Acknowledgements

This work was partially funded by Hasler Foundation under the project Flexible Linguistically-guided Objective Speech assessment (FLOSS) and by Innosuisse under the project Conversation Member Match (CMM). The authors gratefully thank them for their financial support. The Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden), which contributes to the existing infrastructure for quality of life research.

References

[Almaadeed et al., 2016] Almaadeed, N., Aggoun, A., and Amira, A. (2016). Text-Independent Speaker Identification Using Vowel Formants. *Journal of Signal Processing Systems*, 82(3):345–356.

[Ardila et al., 2019] Ardila, R. et al. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

[Boersma and Weenink, 2017] Boersma, P. and Weenink, D. (2017). *Praat: a system for doing phonetics with the computer*.

[Boersma and Weenink, 2019] Boersma, P. and Weenink, D. (2019). Praat: Doing phonetics by computer (computer program). version 6.1.06.

[Christensen, 2018] Christensen, J. V. (2018). The association between articulator movement and formant histories in diphthongs across speaking contexts. MSc thesis, Brigham Young University.

[De Jong and Wempe, 2009] De Jong, N. H. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390.

[Dromey et al., 2013] Dromey, C., Jang, G.-O., and Hollis, K. (2013). Assessing correlations between lingual movements and formants. *Speech Communication*, 55(2):315–328.

[Eyben et al., 2016] Eyben, F. et al. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(02):190–202.

[Fang et al., 2019] Fang, F. et al. (2019). Speaker Anonymization Using X-vector and Neural Waveform Models. In *10th ISCA Speech Synthesis Workshop*, pages 155–160. ISCA.

[Finck and Pallas, 2020] Finck, M. and Pallas, F. (2020). They who must not be identified—distinguishing personal from non-personal data under the GDPR. *International Data Privacy Law*, 10(1):11–36.

[Fradette et al., 2003] Fradette, K., Keselman, H. J., Lix, L., Algina, J., and Wilcox, R. R. (2003). Conventional And Robust Paired And Independent-Samples t Tests: Type I Error And Power Rates. *Journal of Modern Applied Statistical Methods*, 2(2):481–496.

[Harper et al., 2017] Harper, S., Goldstein, L., and Narayanan, S. S. (2017). Quantifying labial, palatal, and pharyngeal contributions to third formant lowering in american english /ɹ/. *The Journal of the Acoustical Society of America*, 142(4):2582–2582.

[Kent et al., 1989] Kent, R. et al. (1989). Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. *Clinical Linguistics & Phonetics*, 3(4):347–358.

[Korshunov and Marcel, 2017] Korshunov, P. and Marcel, S. (2017). Presentation attack detection in voice biometrics. In Vielhauer, C., editor, *User-Centric Privacy and Security in Biometrics*, chapter 7. The Institution of Engineering and Technology, Savoy Place, London WC2R 0BL, UK.

[Kucur Ergunay et al., 2015] Kucur Ergunay, S., Khoury, E., Lazaridis, A., and Marcel, S. (2015). On the vulnerability of speaker verification to realistic voice spoofing. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–8. IEEE.

[Kung, 2018] Kung, S. (2018). A Compressive Privacy approach to Generalized Information Bottleneck and Privacy Funnel problems. *Journal of the Franklin Institute*, 355(4):1846–1872.

[Lammert and Narayanan, 2015] Lammert, A. C. and Narayanan, S. S. (2015). On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. *PLOS ONE*, 10(7):e0132193.

[Lee, 1988] Lee, C. . (1988). On robust linear prediction of speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(5):642–650.

[Lee et al., 2015] Lee, S.-H., Yu, J.-F., Hsieh, Y.-H., and Lee, G.-S. (2015). Relationships between formant frequencies of sustained vowels and tongue contours measured by ultrasonography. *American Journal of Speech-Language Pathology*, 24(4):739–749.

[Mawalim et al., 2020] Mawalim, C. O., Galajit, K., Karnjana, J., and Unoki, M. (2020). X-Vector Singular Value Modification and Statistical-Based Decomposition with Ensemble Regression Modeling for Speaker Anonymization System. In *Proceedings of Interspeech*, pages 1703–1707.

[McKell, 2016] McKell, K. M. (2016). The association between articulator movement and formant trajectories in diphthongs. MSc thesis, Brigham Young University.

[Moulines and Charpentier, 1990] Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5-6):453–467.

[Ning et al., 2019] Ning, Y., He, S., Wu, Z., Xing, C., and Zhang, L.-J. (2019). A Review of Deep Learning Based Speech Synthesis. *Applied Sciences*, 9(19):4050.

[O'Shaughnessy, 2000] O'Shaughnessy, D. (2000). Speaker Recognition. In *Speech Communications: Human and Machine*, pages 437–459. IEEE.

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *Proceedings of ICASSP*, pages 5206–5210. IEEE.

[Patino et al., 2020] Patino, J., Todisco, M., Nautsch, A., and Evans, N. (2020). Speaker anonymisation using the McAdams coefficient. Technical Report EURECOM+6190, Eurecom.

[Patino et al., 2020] Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., and Evans, N. (2020). Speaker anonymisation using the mcadams coefficient. *arXiv preprint arXiv:2011.01130*.

[Povey et al., 2016] Povey, D. et al. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Proceedings of Interspeech*, pages 2751–2755.

[R Core Team, 2019] R Core Team (2019). R: A language and environment for statistical computing.

[Ribeiro, 2018] Ribeiro, M. S. (2018). Parallel audiobook corpus. [dataset]. University of Edinburgh. School of Informatics. `https://doi.org/10.7488/ds/2468`.

[Richardson and Sussman, 2017] Richardson, K. and Sussman, J. E. (2017). Discrimination and identification of a third formant frequency cue to place of articulation by young children and adults. *Language and speech*, 60(1):27–47.

[Rubinstein and Hartzog, 2016] Rubinstein, I. S. and Hartzog, W. (2016). Anonymization and Risk. *Washington Law Review*, 91:59.

[Rudzicz et al., 2012] Rudzicz, F., Namasivayam, A. K., and Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46(4):523–541.

[Sapir et al., 2010] Sapir, S., Ramig, L. O., Spielman, J. L., and Fox, C. (2010). Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech. *Journal of speech, language, and hearing research*, pages 114–125.

[Sapir et al., 2007] Sapir, S., Spielman, J. L., Ramig, L. O., Story, B. H., and Fox, C. (2007). Effects of intensive voice treatment (the lee silverman voice treatment [lsvt]) on vowel articulation in dysarthric individuals with idiopathic parkinson disease: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*, pages 899–912.

[Snyder et al., 2018] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *Proceedings of ICASSP*, pages 5329–5333. IEEE.

[Soldo et al., 2012] Soldo, S., Magimai.-Doss, M., and Bourlard, H. (2012). Synthetic references for template-based asr using posterior features. In *Proceedings of Interspeech*.

[Soldo et al., 2011] Soldo, S., Magimai.-Doss, M., Pinto, J. P., and Bourlard, H. (2011). Posterior features for template-based asr. In *Proceedings of ICASSP*.

[Srivastava et al., 2020] Srivastava, B. M. L. et al. (2020). Evaluating Voice Conversion-based Privacy Protection against Informed Attackers. In *Proceedings of ICASSP*.

[Stalla-Bourdillon and Knight, 2017] Stalla-Bourdillon, S. and Knight, A. (2017). Anonymous Data v. Personal Data – A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data. *Wisconsin International Law Journal*, 34(2):39.

[Tomashenko et al., 2020a] Tomashenko, N. et al. (2020a). Introducing the VoicePrivacy initiative. In *Proceedings of Interspeech*, pages 1693–1697.

[Tomashenko et al., 2020b] Tomashenko, N. et al. (2020b). The VoicePrivacy 2020 Challenge. VoicePrivacy.

[Tomashenko et al., 2020c] Tomashenko, N. et al. (2020c). The voiceprivacy 2020 challenge evaluation plan. `https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_2.pdf`. [Online; accessed 1st April 2020].

[Ullmann et al., 2015] Ullmann, R., Magimai.-Doss, M., and Bourlard, H. (2015). Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences. In *Proceedings of ICASSP*, pages 4924–4928.

[van Son, 2020a] van Son, R. (2020a). Data set for: Adjustable Deterministic Pseudonymization of Speech Listening Experiment, Report of listening experiments.

[van Son, 2020b] van Son, R. (2020b). Listening experiment and Stimuli for: Adjustable Deterministic Pseudonymization of Speech.

[van Son, 2020c] van Son, R. (2020c). Pseudonymizespeech.praat.

[van Son, 2020d] van Son, R. J. J. H. (2020d). Pseudonymize speech. `https://robvanson.github.io/PseudonymizeSpeech/`. [Online; accessed 10th May 2020].

[Van Son et al., 2001] Van Son, R. J. J. H., Binnenpoorte, D., Heuvel, H. v. d., and Pols, L. (2001). The IFA corpus: a phonemically segmented Dutch "open source" speech database. In *Proceedings of EUROSPEECH 2001 Aalborg*, pages 2051–2054, Aalborg, Denmark.

[van Son et al., 2018] van Son, R. J. J. H., Middag, C., and Demuynck, K. (2018). Vowel space as a tool to evaluate articulation problems. In *Proceedings of Interspeech*, pages 357–361.

[Wang et al., 2020] Wang, X. et al. (2020). The voiceprivacy 2020 challenge subjective evaluation-1. Accessed on 25.05.2021.

[Yamagishi et al., 2019] Yamagishi, J. et al. (2019). Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database. [dataset]. University of Edinburgh. School of Informatics. `https://doi.org/10.7488/ds/2555`.

[Zhang et al., 2017] Zhang, Z., Cummins, N., and Schuller, B. (2017). Advanced data exploitation in speech analysis: An overview. *IEEE Signal Processing Magazine*, 34(4):107–129.