Adjustable Deterministic Pseudonymisation of Speech: Idiap-NKI's submission to VoicePrivacy 2020 Challenge

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

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne (EPFL), Switzerland
[3]Netherlands Cancer Institute (NKI), Amsterdam, Netherlands

Abstract

While more and more speech resources become publicly available, the privacy of the speakers needs to be taken care of, in terms of anonymising the speaker information while preserving the linguistic content. In these terms, this paper proposes to use a recently developed deterministic and reversible *pseudonymisation* method that uses signal processing based on formant-shifting to hide the speaker identity. Evaluations on speaker verification and speech recognition on the VoicePrivacy challenge data indicate that the method is better than the McAdams coefficient based signal processing baseline given by the challenge. We show that the proposed method preserves formant tracks better than the McAdams method. We also show, from the intelligibility computed using phone posterior probabilities, that the proposed method preserves intelligibility comparably as the baseline.

Index Terms: speech processing, voice privacy, hand-crafted features.

1. Introduction

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 advancements of speaker verification technologies, it is possible that sensitive information related to vulnerable speakers be extracted from their speech and be misused. Speech anonymisation methods, thus, aim at decoupling the two parts of a given utterance – *what was spoken* and *who spoke it* – and preserve the former while destroying the latter. This work aims at contributing to such novel speech anonymisation approaches, and serves as a submission to the VoicePrivacy 2020 challenge [1, 2].

Anonymisation aims at removing the identity related information from speech. *Pseudonymisation* is a type of anonymisation that is reversible, i.e. the speakers can be re-identified using additional, e.g. private, information. A simple anonymiser recognises the sequence of words spoken in a given utterance and tries to automatically synthesise it. The primary baseline of the challenge uses such an approach. However, in doing so this may also destruct paralinguistic pieces of information, such as the expressed emotions, articulation changes depending on the speaking skills or pathological conditions, etc. Thus, such an anonymisation 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 pseudonymisation based prior knowledge. The Voice privacy challenge provided such an approach as secondary baseline, using McAdams method [3]. Its performance is inferior to that of the primary one in terms of automatic speech recognition (ASR) and automatic speaker verification (ASV) performances. However, signal processing based approaches have the advantage that the changes made and their effects observed could be explained. Thus, further research along this direction is required.

This paper starts with application of van Son's method for adjustable deterministic pseudonymisation of speech as presented by [4]. In their previous work, listening tests showed that the listeners had difficultly in identifying the true speakers from the pseudonymised versions. In the current paper, we further build on it and verify the findings in terms of automatic methods (ASR and ASV) through the VoicePrivacy challenge.

The rest of the paper is organised as follows. Section 2 summarises the methods used, Section 3 presents the experimental setup and results, Section 4 analyses formant measurement in pseudonymised speech and intelligibility measurement based on dynamic time warping (DTW) and Section 5 concludes the paper.

2. Methods

In this section, we first present the proposed pseudonymisation method and present briefly the baseline methods of Voice Privacy challenge.

2.1. Proposed Pseudonymisation method

We utilise the pseudonymisation method proposed by van Son [4]. The speaker pseudonymisation method consists of the following steps sequentially.

2.1.1. Simulate a different vocal tract for the speaker

To simulate a different vocal tract, mainly the formant frequency locations and their amplitudes are modified to match those of a *desired* speaker. To achieve this, first the playback speed of the audio is altered so that the formant frequencies are shifted by a linear factor. This was previously shown to simulate a different vocal tract length (VTL) [5]. Then, the individual formant frequencies are further shifted to the desired values as follows: for each formant, (i) centre two band-pass Hann filters, one at the current formant location and the other at the desired location and (ii) extract and swap the spectral contents of the two locations. This, precisely, creates a version of the source signal with formants at the desired locations. This approach also allows modifying the amplitude of the formant through the filter's gain. Since the formant locations and their amplitudes vary

e-mail: (see http://www.idiap.ch/en/people/directory.)

across time for each utterance, the desired values are typically set as offsets between the two speakers' median parameter values. In other words, the median of the parameters VTL, formant frequencies and their amplitudes are pre-computed per speaker by aggregating across several (a few hundred seconds) of their utterances. During pseudonymisation of a given speaker, the short-time parameter values computed at each small-segment level are added with the appropriate offsets (equal to the difference in the speakers' median values) to simulate the desired speaker's vocal tract.

2.1.2. Change the speaking rate and fundamental frequency

Further, the speaking rate is estimated by an existing method that automatically locates syllables from speech without using transcription [6]. The method uses peaks in the signal energy, that are preceded and succeeded by dips in energy, as cues for syllables. The fundamental frequency and the speaking rate are changed by using a pitch synchronous overlap and add method [7].

2.1.3. Additional processing to hide the speaker identity

The additional anonymising steps consist of (i) exchanging the F4 and F5 bands by using the Hann filter method described above and (ii) adding modulated pink noise at the speaker's F6-F9 bands to mask these formants.

Note that, except for the overlap-add synthesis step and noise insertion, all the steps in this process are deterministic and reversible.

In [4], the ability of human listeners to identify speakers after pseudonymisation using this method was investigated in a series of ABX listening experiments. Overall, the average correct identification of pseudonymised speakers dropped from over 90% in the original recordings to below 70% in pseudonymised speech (where 50% is random). This corresponds to a relative increase of entropy by 91%, from 0.46 of 0.88. (Entropy is computed in a two-class scenario, where 100% identifiability corresponds to 0 bits of entropy and 50% identifiability corresponds to 1 bit.) This indicates that, after pseudonymisation, the uncertainty in identifying the speaker increases considerably.

2.2. Baseline

The challenge provided two baseline systems. The first uses a neural source filtering (NSF) based approach that synthesised speech in a target speaker's characteristics. The system generates the profiles (neural embeddings known as x-vectors) of a pre-defined pool of target speakers which are not part of the enrollment and trials. During the ASV evaluation, each speaker is pseudonymised to the "farthest" speaker, among the target speaker pool, in terms of the PLDA affinity score. In the second baseline method, a signal processing method based on McAdams coefficient is used. In this method, each utterance is analysed using short-time processing, where the poles fit on a given segment using linear prediction are scaled by the McAdams coefficient, and the resultant signal is overlap-added across segments to reconstruct its corresponding pseudonymised utterance.

3. Experimental setup and results

The Praat-based script we used for pseudonymisation is available publicly [8]. We followed the protocol set by the challenge, and evaluated ASR and ASV performances by pseudonymising the given subsets of VCTK and LibriSpeech data sets. For pseudonymisation, target speaker profiles were created using `libri-other-500` set of the LibriSpeech corpus. In a given subset, each speaker is pseudonymised 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). If only the trials sets are pseudonymised, ASV may indicate a higher performance due to acoustic mismatch introduced by the pseudonymisation method. A higher equal error rate (EER) in ASV implies better pseudonymisation of the speakers, and a lower word error rate (WER) on ASR implies better preserving of intelligibility.

In the method we refer to as *F03-9*, we pseudonymised the F_0, F_3, F_4 and F_5 by selecting a random speaker's characteristics from the target speaker pool `libri-other-500`. We also switched the F_4 and F_5 bands, and replace bands F_{6-9} with intensity modulated pink noise.

Tables 1 and 2 compare the ASV and ASR results, respectively, of the baseline anonymisation methods using neural source-filtering (NSF) and McAdams, and the proposed pseudonymisation method. In ASR, the proposed method gave a lower WER than the McAdams baseline, indicating better intelligibility, in all the cases. In ASV, the EER in all the cases except one (vctk-different female) is higher, implying a better pseudonymisation, than the McAdams baseline. This is also indicated by a consistently higher or equal C_{llr}^{min} in all the cases.

Table 1: *ASV results for both development and test partitions (G-gender, E-experiment, o-original, p-pseudonymised(F03-9), b1-NSF. b2-McAdams).*

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

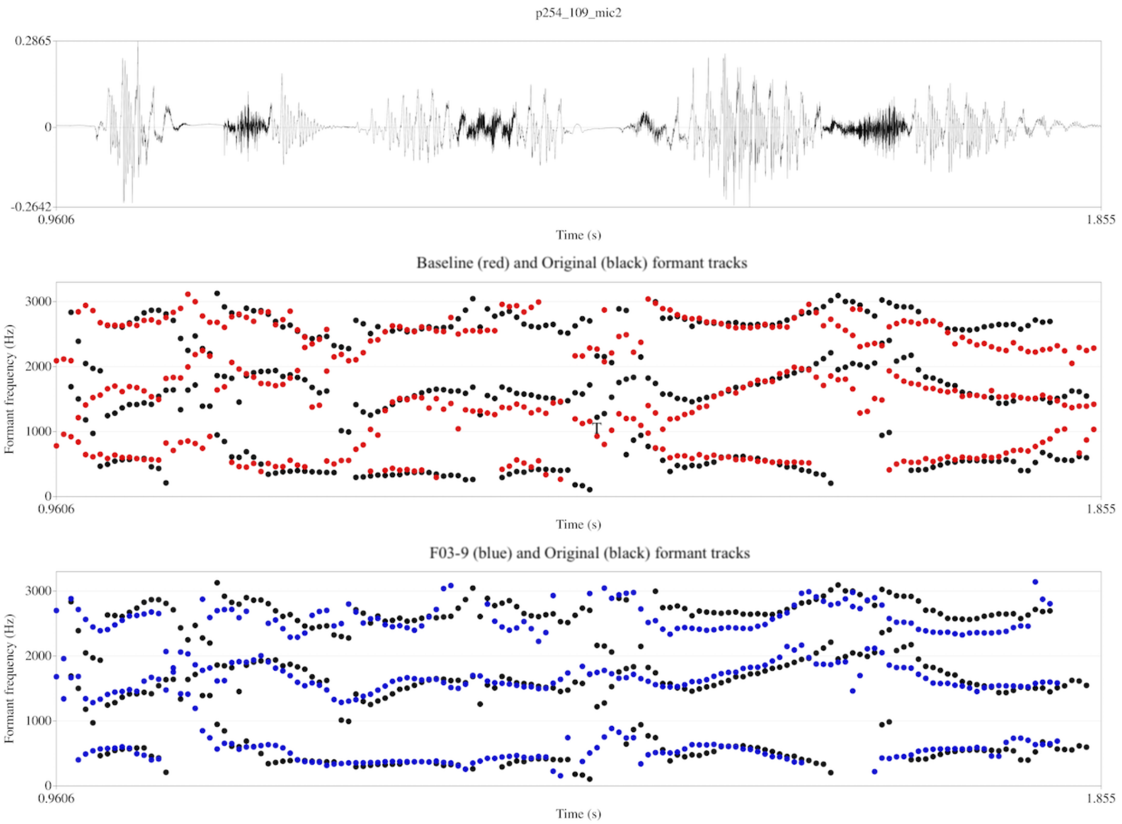

Figure 1: *Example formant tracks for correlating formant values between pseudonymised speech and the original recordings. Top: waveform of sentence [but it is a pleasure] from speaker p254, center: F_1-F_3 formant tracks for Baseline (red) and Original (black) speech, bottom: id. for F03-9 pseudonymisation (blue) and Original (black).*

Table 2: *ASR results in WER% for both development and test partitions (o-original, b1-NSF, b2-McAdams, p-pseudonymised(F03-9), s-LM_s, l-LM_l).*

E	**libri**				**vctk**			
	Dev. set		**Test set**		**Dev. set**		**Test set**	
	s	**l**	**s**	**l**	**s**	**l**	**s**	**l**
o	5.24	3.84	5.55	4.17	14.00	10.78	16.38	12.80
b1	8.76	6.39	9.15	6.73	18.92	15.38	18.88	15.23
b2	12.19	8.77	11.77	8.88	30.10	25.56	33.25	28.22
p	8.82	6.48	8.04	5.87	21.99	18.23	23.32	18.89

4. Analysis

4.1. Measuring formants in pseudonymised speech

The aim of the pseudonymisation method proposed by [4] is to protect the identity of the speaker while preserving selected linguistically and phonetically relevant aspects of speech. The relative performance of the pseudonymisation method is investigated by comparing the ability to compare formant values between pseudonymised and original recordings for the *baseline* and *F03-9* pseudonymisation procedures.

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 [9, 10, 11, 12]. Formant values are also related to the intelligibility of phonetic contrasts [13, 14, 15]. These relations are also relevant to the study of pathological speech, such as dysarthic speech [16] and Parkinson's disease [17].

For the comparison, the first three formant tracks of pseudonymised speech samples were correlated to those of the original recordings, using the *Robust* formant tracking in *Praat* [18]. The same recordings from 60 speakers (30F/30M from vctk_dev and vctk_test) were used for *Baseline* and *F03-9* pseudonymisation. A higher average correlation coefficient (*R*) indicates that the pseudonimised speech would be more useful to studying acoustic effects of differences in articulation.

The results of the comparison show that the average *R* of the pseudonymised formant values were consistently higher for the *F03-9* pseudonymisations than for the *Baseline* method for all three formants. *R* values were 0.1-0.3 higher on average for all speakers (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 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).

The outcomes indicate that the *F03-9* pseudonymisation better preserves F_{1-3} formant track movements than the *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.

4.2. Intelligibility measure based on DTW distance

The challenge proposed WER of ASR as a measure of intelligibility. Several components such as language model, pronunciation lexicon, etc. can affect the performance of such a system. Here we propose to utilise the reference speech to get an intelligibility score, by directly comparing its linguistic content to that of the pseudonymised speech using a DTW based approach. First demonstrated in the context of using synthetic speech for template-based ASR using posterior features [19] and then extended to speech intelligibility assessment by Ullmann *et al.* [20], the method consists of estimating phoneme posterior probabilities, typically from an artificial neural network (ANN), and then comparing such reference and test probability sequences using DTW based on Kullback Leibler (KL) divergence as the local score [21, 19].

The local score can be written as

$$d_{jt} = \mathbb{KL}(\mathbf{y}_j \parallel \mathbf{z}_t)\,, \tag{1}$$

and the cumulative score as

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

where the initial values $D_{00} = D_{(-1)0} = 0$. 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_{JT} normalised by the 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 ANN acoustic model provided with VoicePrivacy challenge and marginalising 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 pseudonymisation method (F03-9) and the baseline method (NSF) by averaging the scores of all the utterances in each method and then comparing them.

Results from Table 3 indicate that the intelligibility scores for the proposed pseudonymisation method are comparable to those of the NSF baseline.

Table 3: *Intelligibility in terms of DTW distances (b1-NSF, p-pseudonymised(F03-9)).*

E	**libri**		**vctk**	
	Dev.	**Test**	**Dev.**	**Test**
b1	0.006915	0.005584	0.007484	0.005751
p	0.007041	0.006047	0.007012	0.004604

5. Conclusions

We proposed to evaluate a deterministic and adjustable pseudonymisation method on the VoicePrivacy challenge and showed that the method pseudonymises utterances better than a signal processing based comparable method, that uses McAdams coefficient, provided by the baseline. A formant track analysis showed a better correlation of the formant tracks with the proposed method than with the baseline approach. DTW distance-based intelligibility computed from the phone posteriors indicate that the proposed method performs comparable to the NSF baseline.

6. Acknowledgements

This work was funded by Hasler Foundation under the project Flexible Linguistically-guided Objective Speech aSsessment (FLOSS). 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 (Malmö, Sweden), which contributes to the existing infrastructure for quality of life research.

7. References

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

[2] ——, “Introducing the VoicePrivacy initiative,” submitted to Interspeech 2020, https://arxiv.org/pdf/2005.01387, 2020, [Online; accessed 10th May 2020].

[3] J. Patino, M. Todisco, A. Nautsch, and N. Evans, “Speaker anonymisation using the McAdams coefficient,” Eurecom, Tech. Rep. EURECOM+6190, 02 2020. [Online]. Available: http://www.eurecom.fr/publication/6190

[4] R. J. J. H. van Son, “Adjustable deterministic pseudonymization of speech,” Report on Zenodo, 2020. [Online]. Available: http://doi.org/10.5281/zenodo.3773931

[5] A. C. Lammert and S. S. Narayanan, “On Short-Time Estimation of Vocal Tract Length from Formant Frequencies,” *PLOS ONE*, vol. 10, no. 7, p. e0132193, 2015.

[6] N. H. De Jong and T. Wempe, “Praat script to detect syllable nuclei and measure speech rate automatically,” *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[7] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[8] R. J. J. H. van Son, “Pseudonymize speech,” https://robvanson.github.io/PseudonymizeSpeech/, 2020, [Online; accessed 10th May 2020].

[9] C. Dromey, G.-O. Jang, and K. Hollis, “Assessing correlations between lingual movements and formants,” *Speech Communication*, vol. 55, no. 2, pp. 315–328, 2013.

[10] S.-H. Lee, J.-F. Yu, Y.-H. Hsieh, and G.-S. Lee, “Relationships between formant frequencies of sustained vowels and tongue contours measured by ultrasonography,” *American Journal of Speech-Language Pathology*, vol. 24, no. 4, pp. 739–749, 2015.

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

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

[13] R. Kent, J. Kent, G. Weismer, R. Martin, R. Sufit, B. Brooks, and J. Rosenbek, “Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects,” *Clinical Linguistics & Phonetics*, vol. 3, no. 4, pp. 347–358, 1989.

[14] S. Harper, L. Goldstein, and S. S. Narayanan, “Quantifying labial, palatal, and pharyngeal contributions to third formant lowering in american english /ɹ/,” *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2582–2582, 2017.

[15] K. Richardson and J. E. Sussman, “Discrimination and identification of a third formant frequency cue to place of articulation by young children and adults,” *Language and speech*, vol. 60, no. 1, pp. 27–47, 2017.

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

[17] S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, “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*, pp. 899–912, 2007.

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

[19] S. Soldo, M. Magimai.-Doss, and H. Bourlard, “Synthetic references for template-based asr using posterior features,” in *Proceedings of Interspeech*, 2012.

[20] R. Ullmann, M. Magimai.-Doss, and H. Bourlard, “Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences,” in *Proceedings of ICASSP*, 2015, pp. 4924–4928.

[21] S. Soldo, M. Magimai.-Doss, J. P. Pinto, and H. Bourlard, “Posterior features for template-based asr,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011.