Speech Pseudonymisation Assessment Using Voice Similarity Matrices

Paul-Gauthier Noé, Jean-François Bonastre, Driss Matrouf, Natalia Tomashenko, Andreas Nautsch, Nicholas Evans

Speech Pseudonymisation Assessment Using Voice Similarity Matrices

Paul-Gauthier Noé[1], *Jean-François Bonastre*[1], *Driss Matrouf*[1], *Natalia Tomashenko*[1], *Andreas Nautsch*[2] *and Nicholas Evans*[2]

[1]Laboratoire Informatique d'Avignon (LIA), Avignon Université, France
[2]Digital Security Department, EURECOM, France

paul-gauthier.noe@univ-avignon.fr

Abstract

The proliferation of speech technologies and rising privacy legislation calls for the development of privacy preservation solutions for speech applications. These are essential since speech signals convey a wealth of rich, personal and potentially sensitive information. Anonymisation, the focus of the recent VoicePrivacy initiative, is one strategy to protect speaker identity information. Pseudonymisation solutions aim not only to mask the speaker identity and preserve the linguistic content, quality and naturalness, as is the goal of anonymisation, but also to preserve voice distinctiveness. Existing metrics for the assessment of anonymisation are ill-suited and those for the assessment of pseudonymisation are completely lacking. Based upon voice similarity matrices, this paper proposes the first intuitive visualisation of pseudonymisation performance for speech signals and two novel metrics for objective assessment. They reflect the two, key pseudonymisation requirements of de-identification and voice distinctiveness.

Index Terms: pseudonymisation, anonymisation, privacy preservation, VoicePrivacy

1. Introduction

The ubiquity and proliferation of speech technologies and the increase in data protection regulation such as the European General Data Protection Regulation (GDPR) [1] has fueled interests in privacy preservation solutions for speech data [2]. There are two general strategies: encryption and anonymisation. Encryption is applied to protect speech data from interception and eavesdropping. Anonymisation aims to ensure that the protected speech data cannot be linked to the original speaker.

With very few solutions having been proposed, and with the few existing solutions achieving only modest levels of anonymisation, the VoicePrivacy initiative[1] [3] was launched in 2019 to promote the consideration of privacy and to foster progress in privacy preservation. VoicePrivacy takes the form of a challenge in which participants are tasked with the development of anonymisation solutions to (i) suppress (as much as possible) the speaker identity from an utterance while nonetheless preserving voice distinctiveness and (ii) leave intact (as much as possible) the linguistic content, quality and naturalness. The requirement for voice distinctiveness implies that anonymised voices remain distinguishable and that all utterances from the same original speaker are anonymised with the same *pseudovoice*. Such a requirement avoids confusion between speakers during a dialogue session and thus allows speaker diarization. We hence refer to the process to meet all these requirements as *pseudonymisation*.

[1]https://voiceprivacychallenge.org

While VoicePrivacy stands to make substantial inroads, it is clear that the metrics used to assess pseudonymisation performance are far from being straightforward; they must reflect multifaceted criteria. The work in this paper is concerned with metrics that reflect criteria related exclusively to the speaker *identity*; it is not concerned with complementary metrics for assessing the preservation of linguistic content etc. Most of the prior work including VoicePrivacy, e.g. [3, 4, 5], measures privacy using trivial, generic metrics such as the Equal Error Rate (EER) estimated from Automatic Speaker Verification (ASV) experiments. The general idea is to gauge performance by comparing the EER using original speech data to that obtained using speech data after de-identification; the greater the difference, or the higher the EER, the better the de-identification and privacy.

Despite its simplicity and ease of interpretablity, the EER is ill-suited as a measure of privacy. Principally, this is because the EER reflects the perspectives of an evaluator and not those of a *privacy adversary*. While a framework to overcome these issues is proposed in [6], it addresses only one component of the pseudonymisation problem, namely that relating specifically to de-identification; it does not reflect *voice distinctiveness*. A solution to address both, i.e. a solution for the assessment of pseudonymisation, is the novel contribution in this paper.

We propose two pseudonymisation metrics for the assessment of de-identification and of voice distinctiveness. Voice similarity matrices, upon which the two objective metrics are inspired, provide easily-interpretable visualisations of any speaker-dependent pseudonymisation behaviour and performance. First, with a widely established privacy preservation terminology currently lacking, we provide definitions of de-identification and voice distinctiveness. We then present voice similarity matrices and show how the two metrics are derived from them. Finally, we present the results of pseudonymisation experiments performed using the VoicePrivacy 2020 data sets and baseline systems.

2. Pseudonymisation

Many of the terms used in privacy research are ill-defined or at least lack a shared understanding within the speech community. We define here more precisely the two requirements for pseudonymisation, how they relate to other terms referred to in the literature and how our work relates to them. The requirements are as follows.

- **De-identification:** a process to conceal in a speech utterance the true speaker identity [7, 8, 9], also referred to as speaker identity masking [10] or voice disguise [11, 12].
- **Voice distinctiveness:** de-identified voices should remain distinguishable within one session (e.g. a single teleconference) such that different speakers still have different, but consistent voices, i.e. protected utterances

produced by the same speaker should be mutually linkable within a session, but they should not be linkable to the original unprotected voice. Moreover, pseudovoices should not be linkable between pseudonymised sessions but this aspect is not assessed in this work. Voice distinctiveness is different to *voice-indistinguishability*, a term coined in [13]. The latter refers specifically and only to the unlinkability between original and protected voices.

Both anonymisation and pseudonymisation are defined within the European GDPR [1]. The GDPR specifies that anonymisation should be irreversible whereas pseudonymisation involves the replacement of an identity with a pseudo-identity. Thus in our speech pseudonymisation framework, the mapping between unprotected and protected voices should be injective (one-to-one mapping) in order to produce distinct pseudovoices. Hence, the pseudonymisation mapping may be reversible (at least in a single session) which is incompatible with the irreversibility requirement for anonymisation.

This paper proposes visualisations and metrics to assess the *level* of de-identification, i.e. the uncertainty in the linkability between a given utterance and the speaker identity, and the *level* to which voice distinctiveness is altered in the protected space. Each speaker should have their *own* protected voice. In terms of established speech research terms, speaker diarization should perform similarly in both unprotected and protected domains.

3. Voice Similarity Matrices, a Visualisation

This section describes voice similarity matrices for the assessment of pseudonymisation according to the two requirements of de-identification and voice distinctiveness. These matrices are similar to conventional confusion matrices except that they are formed with classification *scores* resulting from the exhaustive comparison of utterances collected from a set of speakers. Scores take the form of posterior probabilities, a visualisation of which is provided in the form of a heatmap.

3.1. Voice Similarity Matrix

Let $lr(x, y)$ denote the likelihood-ratio score from the comparison of two speech segments x and y. Assuming equal priors, it is expressed in terms of the posterior probabilities:

$$lr(x, y) = \frac{P(H_{\text{tar}}|x, y)}{P(H_{\text{imp}}|x, y)} \tag{1}$$

where H_{tar} is the target proposition (x and y were uttered by the same speaker) and where H_{imp} is the complementary impostor proposition. Scores, usually in the form of the log-likelihood-ratio (llr), can be calibrated [14] in order to produce so-called *oracle* scores. The latter are used to compute the *voice similarity* which, for two speakers i and j, we define as:

$$S(i, j) = \text{sigmoid}\left(\sum_{\substack{1\le k\le n_i \\ 1\le l\le n_j}} \frac{llr(x_k^{(i)}, x_l^{(j)})}{n_i n_j}\right) \tag{2}$$

which represents the posterior of the averaged llr, where $x_q^{(p)}$ is the q-th segment of the p-th speaker, n_p is the number of segments from the p-th speaker and $\text{sigmoid}(y) = (1 + \exp^{-y})^{-1}$. For $i = j$, scores for which $k = l$ are removed from the average in (2) in order to avoid the consideration of identical speech segments which could lead to an overestimated similarity. While the average in (2) operates upon log-likelihood-ratios, use of the sigmoid function yields voice similarity scores in posterior probability space. The voice similarity matrix M is then given by $M = (S(i, j))_{1\le i\le N, 1\le j\le N}$ where N is the number of speakers. An example voice similarity matrix is illustrated in the **top-left quadrant** of Fig. 1a. The horizontal (left-to-right) and vertical axes (top-to-bottom) indicate the speaker indices i and j for $N = 10$ speakers. In this example, the diagonal elements depict the dominant average similarity between same-speaker utterances, i.e. target trial comparisons that result in a higher $S(i, j)$. The off-diagonal elements depict lower average similarity between different speakers, i.e. impostor trial comparisons that result in a lower $S(i, j)$.

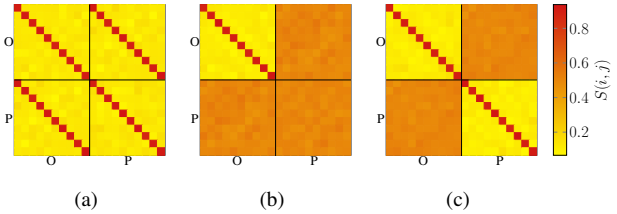


Figure 1: *Three artificial similarity matrices. The upper-left matrix is M_{OO}, the upper-right and lower-left are M_{OP} whereas the lower-right is M_{PP}.*

3.2. Visualisation of pseudonymisation performance

We now explain how voice similarity matrices are used to assess pseudonymisation systems. Pseudonymisation is applied to transform a set of original speech segments (O) to a set of protected, pseudonymised speech segments (P). We then define four voice similarity matrices. M_{OO} and M_{PP} reflect voice similarity *within* the original and pseudonymised speech segment sets. The other two, M_{OP} and M_{PO}, reflect the voice similarity *between* the original and the pseudonymised sets. $M_{\text{OP}} = (M_{\text{PO}})^\top$ where $(\cdot)^\top$ denotes the transpose operator. However, as the voice similarity S is assumed to be symmetric, all matrices are symmetric and $M_{\text{OP}} = M_{\text{PO}}$. In the remainder of this paper, we hence refer only to M_{OO}, M_{OP} and M_{PP}.

Fig. 1 shows three example similarity matrices. In each case, M_{OO} is in the upper-left quadrant, M_{OP} is in the upper-right quadrant and M_{PP} is in the lower-right quadrant. Fig. 1a illustrates the impact upon voice similarity of a poor pseudonymisation system; voice similarities between original, pseudonymised and original-pseudonymised segments are more-or-less identical. Pseudomynisation achieves nothing, even if voice distinctiveness is preserved (M_{PP} still exhibits a dominant diagonal). Fig. 1b illustrates the behaviour of a different pseudonymisation system for which voice distinctiveness is lost (there is no dominant diagonal in M_{PP}). This system, however, is more successful in de-identification (M_{OP} also has no dominant diagonal). Fig. 1c visualises the performance of an ideal case in which both de-identification and voice distinctiveness criteria are met: M_{OP} is uniform without a dominant diagonal; M_{PP} does exhibit a dominant diagonal.

The three M matrices serve to visualise any differences in pseudonymisation performance at the speaker level. While these are not apparent in the artificial examples in Fig. 1, since ASV performance typically varies across different speakers [15, 16], they are expected in practice for real speech data (see Section 5.2). We show next how the visualisations shown in Fig. 1 can be used to derive objective measures of both de-identification and voice distinctiveness.

4. Proposed Metrics

M_{OO} and M_{PP} show voice distinctiveness in original and pseudonymised space respectively, while M_{OP} shows the ease with which speakers in original space can be linked to speakers in pseudonymised space (and vice versa). This information is most easily visualised by the presence or absence of a dominant diagonal. Accordingly, a measure of de-identification and voice distinctiveness can be captured by quantification of diagonal dominance: the key idea behind both proposed metrics.

For any of the three M matrices, the diagonal dominance $D_{\text{diag}}(M)$ is defined as the absolute difference between the averages of the diagonal and the off-diagonal elements:

$$D_{\text{diag}}(M)=\Big|\Big(\sum_{1\leq i\leq N}\frac{S(i,i)}{N}\Big)-\Big(\sum_{\substack{1\leq j\leq N\\1\leq k\leq N\\j\neq k}}\frac{S(j,k)}{N(N-1)}\Big)\Big| \quad (3)$$

D_{diag} will be 0 for a constant/uniform matrix and 1 for an identity matrix as well as for a matrix where all diagonal elements are 0 and all off-diagonal elements are 1.

4.1. De-identification

A measure of de-identification performance is obtained from the comparison of D_{diag} in the original space to that *between* the original and pseudonymised space. Assuming that $D_{\text{diag}}(M_{\text{OO}})$ is strictly positive and that $D_{\text{diag}}(M_{\text{OP}}) \leq D_{\text{diag}}(M_{\text{OO}})$ (de-identification should always reduce the diagonal dominance in M_{OP}), then de-identification performance is measured according to:

$$\text{DeID} = 1 - \frac{D_{\text{diag}}(M_{\text{OP}})}{D_{\text{diag}}(M_{\text{OO}})} \quad (4)$$

The denominator acts to normalise $D_{\text{diag}}(M_{\text{OP}})$ so that a pseudonymisation solution that does nothing will yield a DeID of 0%. Conversely, if the de-identification is optimal, i.e. $D_{\text{diag}}(M_{\text{OP}}) = 0$, then $\text{DeID} = 100\%$.

4.2. Voice distinctiveness

Pseudonymisation can both degrade or improve voice distinctiveness. Motivated from electrical engineering and signal processing, we report a gain value on a decibel (dB) scale as follows:

$$G_{\text{VD}} = 10\log_{10}\Big(\frac{D_{\text{diag}}(M_{\text{PP}})}{D_{\text{diag}}(M_{\text{OO}})}\Big) \quad (5)$$

Gains above 0 dB indicate an increase in voice distinctiveness. Gains below 0 dB indicate a degradation whereas a value of exactly 0 dB indicates that voice distinctiveness in original space is preserved in pseudonymised space.

5. Pseudonymisation, a Case Study

In this section we present an analysis of pseudonymisation performance using the proposed matrices and the de-identification and voice distinctiveness metrics.[2]

5.1. Data sets, protocols and baselines

This work was performed using the VoicePrivacy Challenge 2020 [3] data sets and the two associated baseline systems. The primary baseline, inspired from [17], is based on a x-vector pooling and neural waveform model resynthesis approach. The secondary baseline is based on vocal tract filter transformations using McAdams coefficients [18].

[2]The matrices and metrics are integrated in the VoicePrivacy Challenge: `https://github.com/Voice-Privacy-Challenge`.

#	Official name	Short name
1	libri_dev_trials_f	ldtf
2	libri_dev_trials_m	ldtm
3	vctk_dev_trials_f	vdtf
4	vctk_dev_trials_m	vdtm
5	vctk_dev_trials_f_common	vdtfc
6	vctk_dev_trials_m_common	vdtmc

Table 1: *Renaming of the development sets presented in [21].*

Results are reported for the VoicePrivacy 2020 development data sets. They are drawn from *LibriSpeech-dev-clean* [19] and *VCTK-dev* [20]. We use the trial parts of the challenge development data sets. Details of both data sets and baselines can be found in [21]. For brevity, the sets are renamed as shown in Tab. 1.

As per challenge conditions, scores are obtained from the comparison of x-vectors [22] using probabilistic linear discriminant analysis [23]. Each of the three score sets used for the computation of the three similarity matrices are oracle calibrated [14].

5.2. Pseudonymisation assessment results

Fig. 2 provides separate visualisations of pseudonymisation performance for three of the data sets. They show that the primary baseline (left column) delivers better de-identification performance than the secondary baseline (right column). For the primary baseline, entries in M_{OP} (upper-right and lower-left quadrants) have values close to 0.5, indicating strong de-identification. In contrast, M_{OP} matrices for the secondary baseline show perceptible diagonals, indicating weaker de-identification. The same visualisations show that the secondary baseline better preserves voice distinctiveness; diagonals in M_{PP} matrices (lower-right quadrants) are more perceptible for the secondary than the primary baseline.

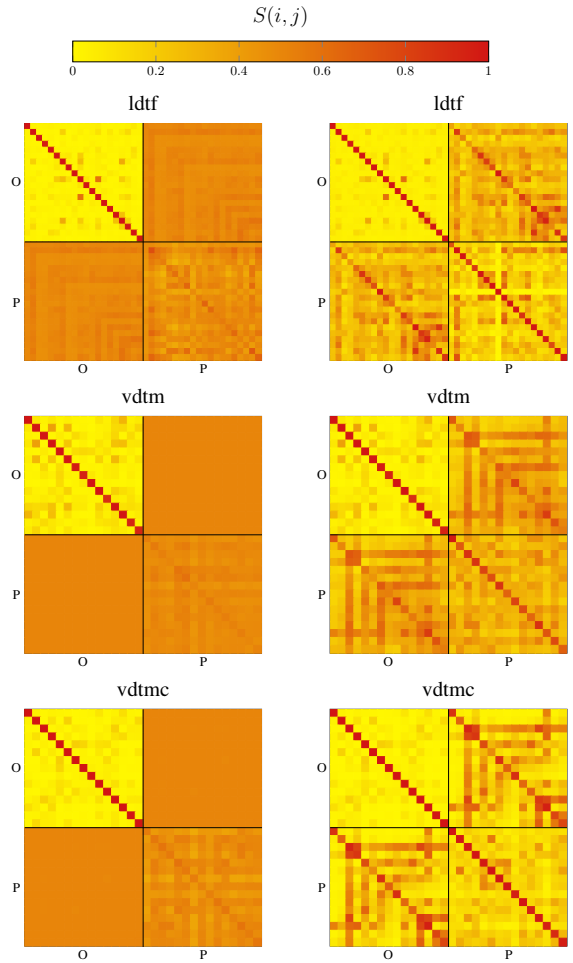
Some differences in performance across speakers and data sets are also visible in Fig. 2. For the primary baseline, voice distinctiveness seems to be slightly better for ldtf and vdtmc data than for vdtm data (more distinctive diagonals in M_{PP} matrices). While de-identification appears to be consistent for the primary baseline, the secondary baseline appears to perform slightly better for vdtm data than for ldft and vdtmc data, albeit it still poorly. De-identification performance is also seen to depend on the speaker, e.g. there are visible striations in M_{OP} for the secondary baseline and vdtm and vdtmc data and, to a much lesser extent, for the primary baseline and ldtf data. The resulting voice distinctiveness also depends on the speaker, e.g. for the secondary baseline, the pseudovoice of one speaker in ldtf is notably more distinctive than others (pure yellow row and column apart the diagonal element in M_{PP}).

DeID and G_{VD} results for each data set and for the primary (blue triangles) and secondary (red squares) baselines are shown in Fig. 3 and Tab. 2. DeID rates for the primary baseline are consistently close to 100%, whereas those for the secondary baseline vary between approximately 44% and 93%. G_{VD} rates for the secondary baseline are consistently close to zero, whereas those for the primary baseline vary between approximately -7.5 and -13 dB. Such substantial degradations to voice distinctiveness in comparison to the secondary baseline are no surprise since the primary baseline is

	primary baseline		secondary baseline	
set	DeID [%]	G_{VD} [dB]	DeID [%]	G_{VD} [dB]
ldtf	99.54	-9.19	55.41	-1.06
ldtm	100	-8.66	41.23	-1.19
vdtf	99.61	-8.71	93.26	-3.60
vdtm	100	-12.66	71.19	-2.98
vdtfc	99.51	-7.46	84.09	-1.39
vdtmc	99.96	-10.30	43.87	-0.81

Table 2: *Results of De-Identification and Gain of Voice Distinctiveness for both baselines and each set.*

based upon x-vector averaging over a subset of speakers. These objective results confirm observations from the visualisations in Fig. 2.



(a) *Primary baseline* (b) *Secondary baseline*

Figure 2: *Voice similarity matrices for the two baselines on ldtf, vdtm and vdtmc.*

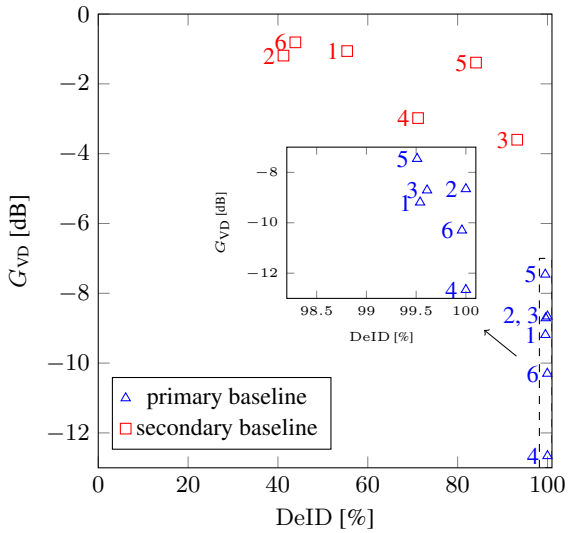


Figure 3: *Scatter plot of Gain of Voice Distinctiveness (G_{VD}) vs. De-Identification* (DeID) *for both baselines and each set.*

6. Conclusions

This paper describes an approach to visualise the de-identification and voice distinctiveness delivered by pseudonymisation solutions and defines objective metrics. Voice similarity matrices, upon which the visualisations and metrics are based, provide revealing, snapshot insights into pseudonymisation performance. They expose differences in performance across different data and speakers. For the latter, visualisations show that, while a particular pseudonymisation solution might perform well on average, it might leave some subjects with relatively weak protection, a finding which is not evident from results derived from objective metrics alone. Other findings point towards a possible trade-off or compromise between de-identification and voice distinctiveness. One pseudonymisation solution delivers near-to-perfect de-identification, whereas the other better preserves voice distinctiveness. Solutions based upon the pooling or averaging of speaker characteristics, as it is the case for the primary baseline, may lead to losses in voice distinctiveness.

Future work should hence investigate injective voice mapping techniques to preserve distinctiveness. However, they will require careful design since they may jeopardise irreversibility (voices cannot be re-identified through an inverse transformation), a key requirement for anonymisation. A compromise solution might be to insure the injectivity for preserving the voice distinctiveness within a diaglogue session whereas adding non-injectivity or randomness between the dialogue sessions. Thus the intra-session mapping could be reversible while the inter-session mapping could be irreversible. In this case a privacy adversary would not be able to use data across sessions. Hence, metrics to the assessment of voice de-identification versus voice distinctiveness in multi-session must be elaborated in future research.

7. Acknowledgements

This work was supported by the JST-ANR Japanese-French project VoicePersonae.

8. References

[1] European Council, "Regulation 2016/679 of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," 4 2016.

[2] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441 – 480, 2019.

[3] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy initiative," in *Interspeech*, 2020.

[4] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Speech Synthesis Workshop*, 2019, pp. 155–160.

[5] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[6] A. Nautsch, J. Patino, N. Tomashenko, J. Yamagishi, P.-G. Noé, J.-F. Bonastre, M. Todisco, and N. Evans, "The privacy ZEBRA: Zero evidence biometric recognition assessment, a speaker recognition perspective," in *Interspeech*, 2020.

[7] T. Justin, V. Štruc, S. Dobrišek, B. Vesnicer, I. Ipšić, and F. Mihelič, "Speaker de-identification using diphone recognition and speech synthesis," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 04, 2015, pp. 1–7.

[8] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," in *2009 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2009, pp. 529–533.

[9] F. Bahmaninezhad, C. Zhang, and J. Hansen, "Convolutional neural network based speaker de-identification," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 255–260. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-36

[10] M. Pobar and I. Ipšić, "Online speaker de-identification using voice transformation," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2014, pp. 1264–1267.

[11] R. G. Hautamäki, A. Kanervisto, V. Hautamaki, and T. Kinnunen, "Perceptual evaluation of the effectiveness of voice disguise by age modification," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 320–326. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-45

[12] C. Zhang, "Acoustic analysis of disguised voices with raised and lowered pitch," in *2012 8th International Symposium on Chinese Spoken Language Processing*, 2012, pp. 353–357.

[13] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, "Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release," *arXiv preprint arXiv:2004.07442*, 2020.

[14] N. Brummer and J. Preez, "The PAV algorithm optimizes binary proper scoring rules," 04 2013.

[15] G. R. Doddington, W. Liggett, A. F. Martin, M. A. Przybocki, and D. A. Reynolds, "Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," in *ICSLP*, 1998.

[16] J. Kahn, S. Rossato, and J.-F. Bonastre, "Beyond doddington menagerie, a first step towards," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 04 2010, pp. 4534 – 4537.

[17] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," 09 2019, pp. 155–160.

[18] 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

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[20] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[21] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "The VoicePrivacy 2020 Challenge evaluation plan," 2020. [Online]. Available: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[23] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 531–542.