Anonymizing Speaker Voices: Easy to Imitate, Difficult to Recognize?

Jennifer Williams, Karla Pizzi, Natalia Tomashenko, Sneha Das

ANONYMIZING SPEAKER VOICES: EASY TO IMITATE, DIFFICULT TO RECOGNIZE?

Jennifer Williams[1], *Karla Pizzi*[2,3], *Natalia Tomashenko*[4], *Sneha Das*[5]

[1] University of Southampton, UK; [2] Fraunhofer AISEC; [3]TU Munich, Germany
[4] University of Avignon, France; [5] Technical University Denmark, Denmark

ABSTRACT

A vastly under-explored area in speech anonymization involves characterizing how different speakers perform in voice privacy tasks. In this paper, we present a deeper analysis by creating and analyzing groups of challenging speakers categorized based on their performance in two related facets of voice anonymization evaluation: (1) speaker similarity using automatic speaker verification (ASV) and (2) human perception using a large-scale A/B listening test. We group speakers into four categories (sheep, goats, lambs, and wolves) based on their anonymization properties. We present an extension of voice anonymization evaluation by identifying speakers who are easy to imitate or difficult to recognize. This knowledge is important for trustworthy anonymization evaluation, and it has the potential to influence how evaluation datasets are created from a pool of speakers. We provide further insights on speaker influence on anonymized speech between human perception and automatic speaker similarity scoring.

Index Terms— voice anonymization, speaker characterization, anonymization perception

1. INTRODUCTION

A prominent open question in speech technology is understanding how individual speakers impact downstream tasks, spanning fields such as text-to-speech (TTS) synthesis [1], voice conversion [2], speech synthesis evaluation [3], automatic speaker verification (ASV) [4], automatic speech recognition (ASR) [5], and voice anonymization [6, 7]. Anonymization techniques are primarily evaluated based on their performance in ASV [8]. Our work in this paper performs a more granular analysis and is a stepping-stone towards uncovering the influence of individual speakers in anonymization evaluations. Through this work, we aim to inform how evaluation techniques and datasets may be designed in the future. Furthermore, understanding the limitations of voice anonymization systems with respect to speakers may be used to improve their trustworthiness and lead to safer deployment.

As [9] describes, there are multiple sources of variability among speakers, ranging from the data collection apparatus and nuisance factors (often referred to as channel [10]), the Lombard effect [11], to the situation of the task (often found in read vs. spontaneous speech [12]). We adopt the definition of anonymization that is used in the VoicePrivacy initiative [13] where *anonymization* refers to the goal of: suppressing personally identifiable information in the speech signal, leaving other attributes intact. This may be detected by automatic speaker recognition algorithms or human listeners or a combination of both. For example, the field of forensics still employs human experts for speaker recognition who may use various automatic tools as additional information sources. Such tools must be trustworthy and reliable in order to be useful. This motivates our cross-sectional study of comparing automatic scoring with human judgements.

In this work, we study how to characterize speakers based on their performance in a set of voice anonymization systems from the 2020 VoicePrivacy Challenge [8]. The first work to examine speaker categories in the context of anonymization was [7], which proposed that distinct speaker categories could be re-purposed from speaker recognition and applied to voice anonymization. Our study takes this idea one step further by assessing multiple anonymization systems based on the per-speaker performance. We also compare these results with a large-scale human subject evaluation. For both setups, we consider four speaker categories from Doddington et al. [4], which are animal-named classes characterized by their performance in speaker recognition: sheep (average), lambs (easy to imitate), goats (difficult to recognize), and wolves (often misdetected), described in more detail in Section 3.1. We include a human perception component in order to expand from previous evaluations from the VoicePrivacy Challenge and as an additional step to further our understanding of how human perception of anonymized voice compares to automatic recognition.

We are primarily interested in identifying speakers who are easy to imitate[1] versus difficult to recognize. We focus this work on one type of attack model in voice privacy called the *lazy-informed* attack model. In this scenario, for each speaker of interest, an adversary has access to an anonymized trial utterance and several anonymized enrollment utterances, ultimately compromising the trustworthiness of the system [8]. There are only few protection mechanisms against this type of attack. The contributions of our paper are the following:

1. Compare categories of speakers based on human and ASV scoring in a *lazy-informed* attack model with respect to their anonymization behavior.
2. Present vulnerabilities of voice anonymization systems based on speaker-centric performance and analysis.
3. Investigate potential differences between human perception and ASV evaluation.

This work is supported by the UKRI TAS Hub (EP/V00784X/1); the Bavarian Ministry of Economic Affairs, Regional Development, and Energy; the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection; and the VoicePersonae Project, the SELMA project (grant No 957017), the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska Curie ESPERANTO project (No 101007666). We thank Poppy Welch for her support with Qualtrics and Farida Yusuf for thoughtful discussions.

[1]We use the term *imitate* w.r.t. speaker verification and we do not mean to convey that we are explicitly studying deepfakes.

2. BACKGROUND AND RELATED WORK

Voice anonymization: The recent VoicePrivacy Challenges of 2020 and 2022 have accelerated the field of voice anonymization work with successful evaluation campaigns and ongoing research [8, 14]. These evaluation campaigns have explored multiple dimensions of voice anonymization including objective and subjective measures as well as anonymized voice distinctiveness. While submissions to the challenge are evaluated at the system-level and performance is often reported separately for male and female speakers, the contribution of individual speakers on system-level performance still remains under-investigated. Also our paper answers the call for more exploration of privacy and utility evaluation metrics [8]. More importantly, they observed some degree of speaker-dependent performance, as more than half of their test speakers did not show improved privacy. Similarly, recent work from [7] presents the idea that the amount of anonymization achieved for each speaker can be characterized by several distinct categories such as those first proposed by [4], according to the types of errors made during ASV evaluation. In our work, we adopt this idea of speaker categories for our analysis and describe the categories in more detail in Section 3.1.

Speech synthesis: Voice anonymization employs a variety of methods and techniques, ranging from low-level signal manipulations to neural models of speech synthesis [8, 15]. Because of this, we include a brief discussion of the impacts of speaker influences in speech synthesis. One of the reasons why it is difficult to assess how suitable different speakers are to speech synthesis is because synthesis often requires very large datasets in order to create robust acoustic models. Selecting smaller sets of speakers who meet certain criteria can help to achieve higher quality text-to-speech (TTS) synthesis [1, 16]. The complexity is compounded by how speech synthesis is typically evaluated, often relying on human judgements which naturally vary from listener to listener and are difficult to replicate over decades of research [3, 17, 18]. Voice anonymization can also be framed as a special type of voice conversion task [6] – one where the original speaker shall not be revealed. Still, voice privacy is evaluated along different metrics or based on different assumptions than voice conversion. The biennial Voice Conversion Challenge (VCC) [19, 20] has not yet evaluated how individual speakers influence voice conversion systems.

Speaker recognition: The set of speakers used for training systems can impact the quality and generalizability of the ASV system at inference. One of the most notable reflections of this comes from the universal background model technique for speaker recognition, which pre-dates neural methods [21]. A further review and brief history of human and machine speaker recognition is provided by [9], including a discussion of the role of forensic scientists for human-based speaker recognition in the legal and law enforcement domains.

3. METHODOLOGY

3.1. Speaker categories based on animals

Our work adopts a set of categories inspired by Doddington et al. [4], originally used to characterize speakers according to their performance in speaker recognition, as measured by true positives (TP), false positives (FP) and false negatives (FN). We present and define the four categories below. This technique calls attention to speakers who, e.g., are particularly easy to imitate or difficult to recognize. Consider a target speaker s and non-target speaker n, and classification decisions on a per-speaker basis:

- **Sheep** – speakers who can be thought of as the average type of speaker that represents the majority of the population, and for which speaker recognition systems are usually able to recognize accurately and consistently: $\textit{sheep}_s = \text{TP}_s$.
- **Goats** – speakers who are difficult to recognize and who have a very high rate of false-negatives: $\textit{goat}_s = \text{FN}_s$.
- **Lambs** – speakers who are particularly easy to imitate, meaning that other non-target speakers (n) are often being detected as this speaker. This is the usual definition of false-positive: $\textit{lamb}_s = \text{FP}_n$.
- **Wolves** – speakers whose voice is inaccurately detected as a target speaker: $\textit{wolf}_s = \text{FP}_s$.

For every target speaker s, the values for **sheep**, **goat**, **lamb**, and **wolf**, are calculated and the category is given by the maximum score among the four categories for that speaker: $\textit{category}_s = \max\{\textit{sheep}_s, \textit{goat}_s, \textit{lamb}_s, \textit{wolf}_s\}$. While these categories have been discussed in previous voice anonymization work [7], they have not been evaluated extensively across multiple voice anonymization systems or compared with human perception of anonymized voices. We adopt these categories in our methodology and analysis, which we describe in greater details in Section 4.

3.2. Data

To investigate categories of speakers, we use anonymization systems from the 2020 VoicePrivacy enrollment and test trial data [8]. It is based on speech from the Voice Cloning Toolkit (VCTK v0.92) [22] corpus. In this study, we use the VCTK *common* test set. Our set contains audio samples from nine different anonymization systems corresponding to the challenge primary system submissions (**A**, **D**, **I**, **K**, **M**, **O**, **S**) and baselines (**B1**, **B2**). The test set contains 30 held-out speakers (15 female and 15 male). The primary baseline **B1** uses an x-vector model [23, 24] and neural speech synthesis as a vocoder. The secondary baseline **B2** uses the McAdams coefficient [25, 26]. The system **D** [27] is derived from baseline **B2**. Systems based on x-vectors include: **A** [28], **M** [29], **O** [30], and **S** [31]. While system **I** [2] uses modifications to formants, system **K** [3] combines x-vectors, speaker similarity models, and a voice indistinguishability metric.

3.3. Human scoring

We obtained human judgement scores through an A/B similarity test with 600 listeners[4], administered via the Qualtrics platform. Participants were recruited using the Prolific platform. Male and female listeners (balanced across gender) were recruited from the United Kingdom, and each was paid £ 4.50 to rate 75 different enrolment/trial pairs. They were presented with two audio samples and asked if the two samples were from the same or different speakers. We used a large subset of speech samples from the 2020 VoicePrivacy Challenge held-out test set to create 45,000 A/B pairs (enroll/trial) including all 15 female and 15 male speakers. We evaluated male and female speakers in separate listening tests. For each of the 30 enrollment speakers, we conducted 150 trials (10 target utterances, and 10 non-target utterances), across the seven system submissions, two baselines, and original speech (ten sources in total). All enrollment and test samples were matched with each of the ten systems as described. For example, in a trial for system **A**

[2] https://www.voiceprivacychallenge.org/docs/Idiap-NKI.pdf

[3] https://www.voiceprivacychallenge.org/vp2020/docs/Kyoto.pdf

[4] Ethics approval Ref: 77895, University of Southampton

both the enrollment and test came from that system. We calculated the percent of trials that listeners marked correctly for the A/B pairs, aggregated by system: **A**: 53%, **D**: 44%, **I**: 69%, **K**: 64%, **M**: 48%, **O**: 47%, **S**: 35%, **B1**: 49%, **B2**: 65%, **Orig**: 76%.

3.4. ASV scoring

We utilize the ASV scores from the VoicePrivacy 2020 evaluation. With anonymization, EER values are expected to increase as privacy increases, signifying that speakers become less recognisable.

Table 1: Comparison of EER,% on the VCTK-*common* test set, for two attack models: *ignorant* using original enrollment with anonymized trial (O-A), and *lazy-informed* using anonymized enrollment with anonymized trial (A-A).

	Male		Female	
System	O-A	A-A	O-A	A-A
A	55.7	20.3	48.6	28.6
D	27.7	12.9	29.8	17.1
I	29.1	13.8	32.7	26.0
K	57.0	5.7	50.3	2.9
M	53.7	35.6	51.5	34.4
O	46.3	42.4	45.7	38.8
S	45.7	38.7	46.5	39.0
B1	53.4	31.0	48.3	31.2
B2	24.3	12.2	30.6	14.2

In Table 1, we show anonymization system performance as measured by the EER for two attack models: *ignorant* and *lazy-informed*. For original data (without anonymization), the performance on the original speech was 1.1% EER for male and 2.9% EER for female speakers. The case of *O-A* (original enrollment / anonymized trial) represents the *ignorant* attack model because the adversary does not have access to the speaker's anonymized data and so uses original enrollments. The *A-A* experiment (anonymized enrollment / anonymized trial) represents the *lazy-informed* attack model where an adversary has access to the anonymized enrollment data of speakers. In this scenario, it is assumed that the trial and enrollment data are anonymized using the same anonymization method, but with different pseudo-speaker voices, because the attacker does not know the exact parameters of anonymization system and pseudo-speaker chosen by each user. We see the lowest EER values for the *A-A* attack model for systems **K**, **D**, **I** and **B2**, which indicates that these systems offer the least protection. This is especially interesting for **K** because it achieves the most privacy protection in the *ignorant* attack model (O-A), which further highlights that these privacy systems have a particular vulnerability under the *lazy-informed* scenario (A-A). For the remainder, we focus on the *A-A* or *lazy-informed* scenario.

In our analysis, we used log-likelihoods based on a speaker-level x-vector representation compared with an utterance-level x-vector representation. The number of trials per system was 5185, and from this we randomly selected ten target and ten non-target trials (150 trials per speaker), similar to the subset selection in Section 3.3. To calculate the predicted speaker versus the true speaker, we apply isotonic regression [32] to the raw log-likelihood scores and target/non-target labels to obtain binary decisions. The binary decisions allow us to then calculate the confusion matrix values per speaker (TP, FP, FN), which we use for our analysis of speaker groupings.

4. ANALYSIS OF SPEAKER CATEGORIES

4.1. Calculating animals

From the human A/B scoring and the ASV scores, we calculated speaker animal categories for each system. The ASV scoring resulted in consistently extremely high TP values which biased all speakers towards *sheep*. To produce a more balanced view for ASV, we sampled 10 matched and 10 unmatched pairs per speaker and recalculated the animal categories. The variation among human A/B scores did not exhibit such an extreme skew toward *sheep*. For both human A/B scoring and ASV scoring, we noticed once or twice for a given system that a speaker could be equally ranked in two (or more) categories (e.g., equally likely to be a *lamb* and a *goat*). In those cases, the category was chosen randomly among the top-ranking categories for that speaker.

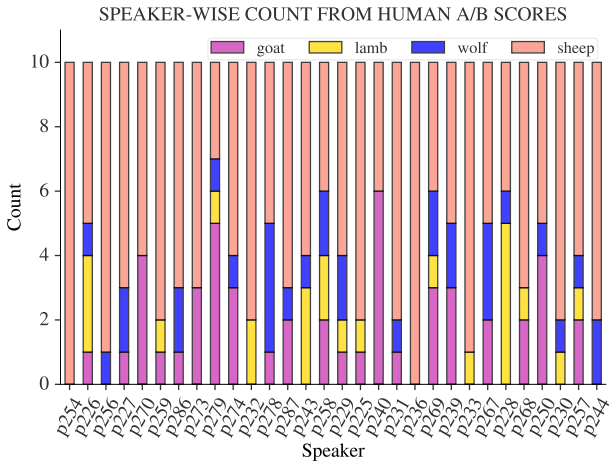
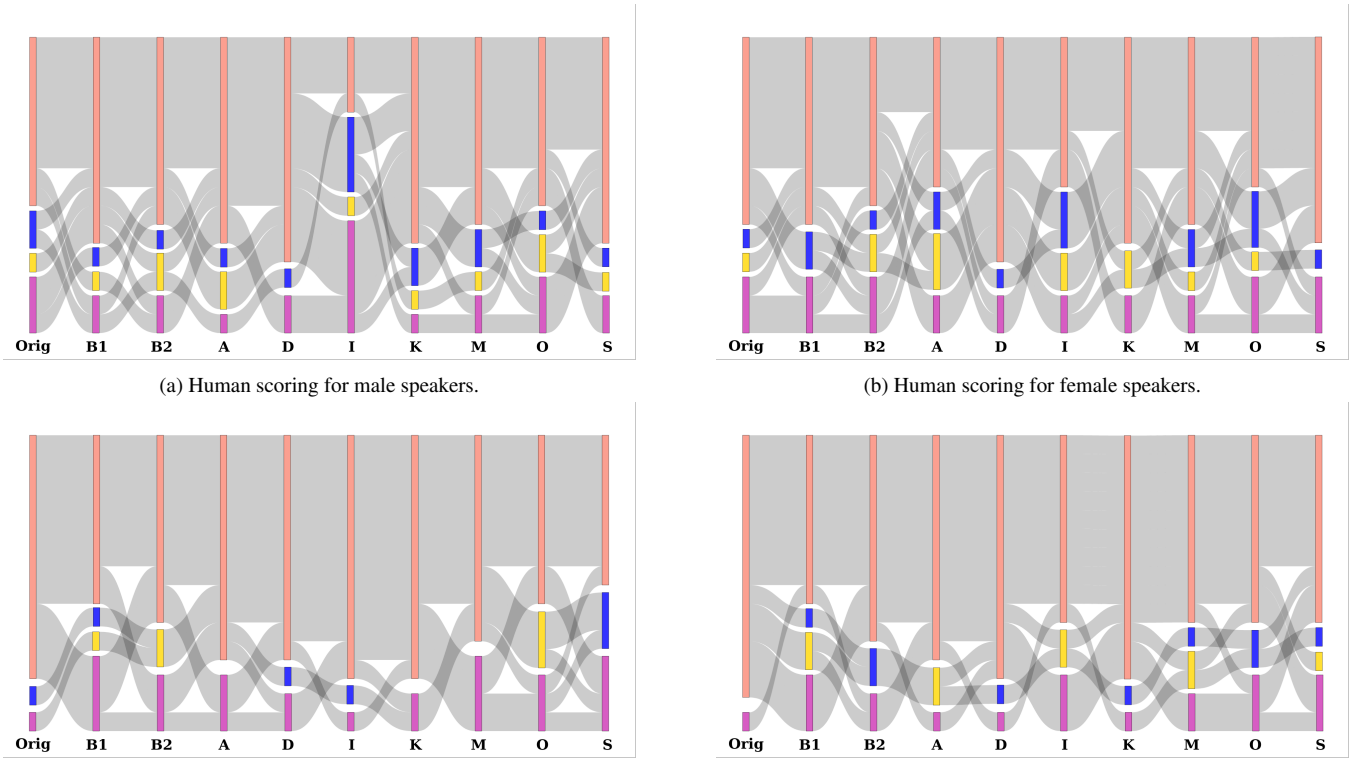


Fig. 1: Per-speaker animal categories from human A/B listening test.

4.2. Animal category analysis

In Figure 1, we show the distribution of animal categories per speaker based on the human A/B listening test, and aggregated across all systems and original speech. We see that some speakers, like *p254* are primarily *sheep* regardless of anonymization system, whereas other speakers, like *p279* are primarily *goats*. Since we computed the animal categories for each anonymization system, we are able to track how speaker groups behave with respect to each system. In Figures 2a–2d we show Sankey diagrams [33] which represent flows from one set to another set. The width of the gray connections between bars is proportional to the flow rate. The colored bars represent the distribution of animal categories for each system. **Orig** denotes original unmodified speech.

The most common animal class for both human and ASV scoring was the *sheep* class, especially for original speech with ASV scoring. This is an expected result and corresponds with the definitions provided in Section 3.1 as this class makes up the majority of speakers, though we are the first to confirm it for anonymized voices. The second most-common category is *goats* (difficult to recognize) indicating that there are some speakers from both genders who are difficult to recognize across systems and scoring techniques. The *lambs* (easy to imitate) and *wolves* are fewer overall. In general, it is more difficult for humans to correctly recognize the speakers compared to ASV, which also implies that humans are perceiving more anonymization across the voices.



(a) Human scoring for male speakers.

(b) Human scoring for female speakers.

(c) ASV scoring for male speakers.

(d) ASV scoring for female speakers.

Fig. 2: Sankey diagrams showing set flow between systems for speaker animal categories with colored bars indicating the proportions of each animal category: ■ – sheep; ■ – lamb; ■ – wolf; ■ – goat.

From the human scoring (Figures 2a–2b), the skew toward *sheep* is slightly less pronounced than with ASV scoring, likely due to the higher variability in human perception. Among the signal-processing based systems (**B2**, **D**, **I**), we notice that system **I** has elevated *goats* and *wolves* for male speakers, but slightly less so for female speakers. System **I** had the highest naturalness and intelligibility in the VoicePrivacy challenge, which might explain why speakers are more easily confused [8]. In contrast, systems **B1** and **D** both have similar performance across genders. Among the neural-based systems (**B1**, **A**, **K**, **M**, **O**, and **S**), there is more consistency of category distributions for male speakers than for female speakers.

From the ASV scoring (Figures 2c–2d), we see more *goats* for male speakers than for the female speakers, meaning that male speakers are more difficult to recognize than female speakers for ASV. However, system **K** has very few *goats*, *lambs*, and *wolves* among both male and female speakers. System **K** also has the lowest EER from Table 1 for the lazy-informed attack model, indicating low false-positives and false-negatives. For trustworthy privacy, however, this is an undesirable performance. Among the neural-based models several similar categories for male speakers (**A**, **K**, and **M**), while others (**B1**, **M**, and **S**) are similar for female speakers.

5. SUMMARY AND FUTURE WORK

The ability to characterize speakers with respect to a downstream task, such as voice anonymization, has far-reaching implications. The differences among speakers is potentially a driving factor in how well anonymization systems perform. Our work using the *lazy-informed* attack model and systems from the 2020 VoicePrivacy Challenge shows it is possible to adapt the speaker categories for anonymization. We propose that the design of datasets for voice anonymization and speaker verification consider the acoustical properties of speakers in these datasets, where such speakers may differentially contribute to overall system performance, and evaluation campaigns that highlight performance on particularly challenging speakers. Our characterization of challenging speakers has broad relevance to a variety of speech processing fields. While we focused on a voice anonymization task, our paper brings attention to a wider issue for evaluation and future work should explore what specifically makes these speakers challenging.

6. REFERENCES

[1] Pilar Oplustil Gallegos, Jennifer Williams, Joanna Rownicka, and Simon King, "An unsupervised method to select a speaker subset from large multi-speaker speech synthesis datasets.," in *Interspeech*, 2020, pp. 1758–1762.

[2] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[3] Erica Cooper and Junichi Yamagishi, "How do Voices from Past Speech Synthesis Challenges Compare Today?," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 183–188.

[4] George R Doddington, Walter Liggett, Alvin F Martin, Mark A

Przybocki, and Douglas A Reynolds, "SHEEP, GOATS, LAMBS and WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *ICSLP*, 1998, vol. 98, pp. 1351–1354.

[5] Xuedong Huang and Kai-Fu Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 150–157, 1993.

[6] Pierre Champion, Denis Jouvet, and Anthony Larcher, "Evaluating x-vector-based speaker anonymization under white-box assessment," in *Speech and Computer: 23rd International Conference, SPECOM*, 2021, pp. 100–111.

[7] Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Jose Patino, Jean-François Bonastre, Natalia Tomashenko, and Driss Matrouf, "Towards a unified assessment framework of speech pseudonymisation," *Computer Speech & Language*, vol. 72, pp. 101299, 2022.

[8] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, et al., "The VoicePrivacy 2020 Challenge: Results and Findings," *Computer Speech & Language*, vol. 74, pp. 101362, 2022.

[9] John HL Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[10] Douglas A Reynolds, "An overview of automatic speaker recognition technology," in *IEEE ICASSP*, 2002, vol. 4, pp. IV–4072.

[11] John HL Hansen and Vaishnevi Varadarajan, "Analysis and compensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 366–378, 2009.

[12] Sadaoki Furui, "Recent advances in spontaneous speech recognition and understanding," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[13] Natalia Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, et al., "Introducing the VoicePrivacy Initiative," in *Interspeech*, 2020.

[14] Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, et al., "The VoicepPivacy 2022 challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[15] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, et al., "Supplementary material to the paper The VoicePrivacy 2020 Challenge: Results and findings," 2021.

[16] Massa Baali, Tomoki Hayashi, Hamdy Mubarak, Soumi Maiti, Shinji Watanabe, Wassim El-Hajj, and Ahmed Ali, "Unsupervised data selection for tts: Using arabic broadcast news as a case study," *arXiv preprint arXiv:2301.09099*, 2023.

[17] Jennifer Williams, Joanna Rownicka, Pilar Oplustil Gallegos, and Simon King, "Comparison of speech representations for automatic quality estimation in multi-speaker text-to-speech synthesis," in *Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 222–229.

[18] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi, "Generalization ability of mos prediction networks," in *IEEE ICASSP*, 2022, pp. 8442–8446.

[19] Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, et al., "Voice conversion challenge 2020—intra-lingual semi-parallel and cross-lingual voice conversion—," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020, vol. 2020, pp. 80–98.

[20] Rohan Kumar Das, Tomi Kinnunen, Wen-Chin Huang, Zhen-Hua Ling, Junichi Yamagishi, et al., "Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020.

[21] Douglas A Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[22] Christophe Veaux, Junichi Yamagishi, et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh, CSTR*, 2017.

[23] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019.

[24] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE ICASSP*, 2018, pp. 5329–5333.

[25] S Mcadams, "Spectral fusion, spectral parsing and the formation of the auditory image," *Ph. D. Thesis, Stanford*, 1984.

[26] Jose Patino, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans, "Speaker anonymisation using the mcadams coefficient," in *Interspeech*, 2021, pp. 1099–1103.

[27] Priyanka Gupta, Gauri P Prajapati, Shrishti Singh, Madhu R Kamble, and Hemant A Patil, "Design of voice privacy system using linear prediction," in *APSIPA ASC*, 2020, pp. 543–549.

[28] Candy Olivia Mawalim, Kasorn Galajit, Jessada Karnjana, and Masashi Unoki, "X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system.," in *Interspeech*, 2020, pp. 1703–1707.

[29] Pierre Champion, Denis Jouvet, and Anthony Larcher, *Speaker information modification in the VoicePrivacy 2020 toolchain*, Ph.D. thesis, INRIA Nancy, équipe Multispeech; LIUM-Laboratoire d'Informatique, 2020.

[30] Henry Turner, Giulio Lovisotto, and Ivan Martinovic, "Speaker anonymization with distribution-preserving x-vector generation for the voiceprivacy challenge 2020," *arXiv preprint arXiv:2010.13457*, 2020.

[31] Fernando M Espinoza-Cuadros, Juan M Perero-Codosero, Javier Antón-Martín, and Luis A Hernández-Gómez, "Speaker de-identification system using autoencoders and adversarial training," *arXiv preprint arXiv:2011.04696*, 2020.

[32] Jan De Leeuw, Kurt Hornik, and Patrick Mair, "Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods," *Journal of Statistical Software*, vol. 32, no. 5, pp. 1–24, 2009.

[33] Richard C Lupton and Julian M Allwood, "Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use," *Resources, Conservation and Recycling*, vol. 124, pp. 141–151, 2017.