A Data Perspective on Ethical Challenges in Voice Biometrics Research

Leschanowsky, Anna; Rusti, Casandra; Quinlan, Carolyn; Pnacek, Michaela; Gorce, Lauriane; Hutiri, Wiebke

DOI
10.1109/TBIOM.2024.3446846

Publication date
2025

Document Version
Final published version

Published in
IEEE Transactions on Biometrics, Behavior, and Identity Science

Citation (APA)
Leschanowsky, A., Rusti, C., Quinlan, C., Pnacek, M., Gorce, L., & Hutiri, W. (2025). A Data Perspective on Ethical Challenges in Voice Biometrics Research. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *7*(1), 118-131. https://doi.org/10.1109/TBIOM.2024.3446846

A Data Perspective on Ethical Challenges in Voice Biometrics Research

Anna Leschanowsky, *Associate Member, IEEE*, Casandra Rusti, *Graduate Student Member, IEEE*, Carolyn Quinlan, Michaela Pnacek(ova), Lauriane Gorce, and Wiebke Hutiri, *Student Member, IEEE*

***Abstract*—Speaker recognition technology, deployed in sectors like banking, education, recruitment, immigration, law enforcement, and healthcare, relies heavily on biometric data. However, the ethical implications and biases inherent in the datasets driving this technology have not been fully explored. Through a longitudinal study of close to 700 papers published at the ISCA Interspeech Conference in the years 2012 to 2021, we investigate how dataset use has evolved alongside the widespread adoption of deep neural networks. Our study identifies the most commonly used datasets in the field and examines their usage patterns. The analysis reveals significant shifts in data practices since the advent of deep learning: a small number of datasets dominate speaker recognition training and evaluation, and the majority of studies evaluate their systems on a single dataset. For four key datasets–Switchboard, Mixer, VoxCeleb, and ASVspoof–we conduct a detailed analysis of metadata and collection methods to assess ethical concerns and privacy risks. Our study highlights numerous challenges related to sampling bias, re-identification, consent, disclosure of sensitive information and security risks in speaker recognition datasets, and emphasizes the need for more representative, fair, and privacy-aware data collection in this domain.**

***Index Terms*—Ethical aspects, privacy, biometrics (access control), speaker recognition, human voice, data transparency, data handling.**

I. Introduction

SPEAKER recognition is widely used in voice biometrics in the private and public sectors, e.g., to verify the identity of banking clients [1], [2] or employees [3], and to secure an expanding network of voice assistants and voice-based Internet of Things devices through which people interact with digital services [4]. The large scale deployment of speaker recognition systems has been facilitated by the adoption of deep learning, which has greatly improved technology performance [5]. However,data-intensive, deep learning systems are prone to produce disparate speaker recognition error rates across demographic groups [6], [7], [8], [9], a phenomenon that we call bias.

Bias is well studied in machine learning and algorithmic fairness research [10], [11]. In its simplest form bias refers to a skewed or slanted perspective. Biased technologies can carry significant social consequences if they produce systematic errors in their outputs that disproportionately advantage or disadvantage certain people without reason [12]. Bias is frequently viewed as a source of *unfairness* which can arise in the machine learning development process, for example from unrepresentative training or evaluation data, or inappropriate data labelling choices [11]. *Fairness* is the aspirational antithesis to *unfairness*. A biased speaker recognition system can lead to discriminatory decision outcomes. In many countries *discrimination* is illegal [13], and (algorithmic) decision-making processes must treat individuals and groups of people equally with regards to protected personal attributes [14].

Despite rapid progress and widespread adoption of speaker recognition technology, bias, fairness and discrimination remain largely unexplored in voice biometrics research. In commercial systems, evaluating bias in biometrics is however gaining prominence. For example, the NIST Face Recognition Vendor Test now includes an evaluation of bias across demographic groups [15], and several studies have evaluated error rate disparities across groups in face recognition models [6], [8], [16]. However, model bias is only one of several sources of bias in deep learning systems [17]. Oftentimes it is caused by bias in training datasets, which then reflects downstream in the learned models [18]. Similarly, bias in evaluation datasets skews evaluation outcomes, channels future development efforts and makes it impossible to assess if models are biased [17]. Beyond evaluating bias in models, it is thus also important to interrogate training and evaluation datasets.

Motivated by prior research on dataset evaluations [19], [20], [21] and data collection [22], [23], this paper presents the first study of ethical concerns in speaker recognition datasets, and their impact on bias in voice biometrics. Grounded in a comprehensive literature review of papers published at the ISCA Interspeech conference between 2012 and 2021, we explore dataset usage dynamics to gain insights into community adoption of datasets and potential

Manuscript received 30 November 2023; revised 21 May 2024; accepted 12 August 2024. Date of publication 21 August 2024; date of current version 27 December 2024. This work was supported in part by the Fair EVA Project, and in part by the Mozilla Technology Fund through the Mozilla Foundation. This article was recommended for publication by Associate Editor N. Evans upon evaluation of the reviewers' comments. *(Corresponding author: Casandra Rusti.)*

Anna Leschanowsky is with the Fraunhofer-Institut für Integrierte Schaltungen IIS, 91058 Erlangen, Germany (e-mail: anna.leschanowsky@iis.fraunhofer.de).

Casandra Rusti is with the University of Southern California, Los Angeles, CA 90089 USA (e-mail: rusti@usc.edu).

Carolyn Quinlan is with the University of Toronto, Toronto, ON M5S 1A1, Canada.

Michaela Pnacek(ova) is with York University, Toronto, ON M3J 1P3, Canada.

Lauriane Gorce is with Mines Paris - PSL, 75272 Paris, France.

Wiebke Hutiri is with Technische Universiteit Delft, Delft, The Netherlands (e-mail: w.toussaint@tudelft.nl).

Digital Object Identifier 10.1109/TBIOM.2024.3446846

cultural shifts in data practices. In particular, our study aims to address the following questions:

1) Which datasets are used for training and evaluation in speaker recognition research?
2) How has dataset usage changed over the period from 2012 to 2021?
3) What are the attributes of the most used datasets?
4) What are the implications of the above questions for bias, fairness, privacy, and other ethical challenges in speaker recognition?

This paper expands our prior work [24] with an extensive metadata analysis, and a detailed review of the ASVspoof, VoxCeleb, Switchboard and Mixer Corpora. We start by reviewing related literature in Section II and describe our research approach in Section III. We present results on dataset dynamics in Section IV, analyse the metadata in Section V and examine ethical concerns in Section VI. We consolidate and reflect on our findings in Section VII, before concluding in Section VIII.

II. Related Work

This section explores the history of voice biometric systems and existing research on bias in biometric systems, with a specific emphasis on face recognition. It examines methodologies for evaluating demographic bias in biometric verification and the impact of data and dataset biases on machine learning models. This context underscores the need for our study, which expands these discussions to speaker recognition, emphasizing the importance of rigorous dataset evaluations in this specialized area of voice biometrics.

A. A Brief History of Voice Biometrics

The ability to recognize a person's voice is inherent in humans and forms the foundation of voice biometrics technology. Automatic voice biometric systems have emerged alongside human-based approaches, such as auditory comparison or visual spectrogram inspection, often conducted by forensic experts [25]. The first fully automated speaker recognition system built by Texas Instruments in the early 1970s [5], [26], [27]. Statistical models like Hidden Markov Models (HMMs) replaced rule-based speech recognition systems in the 1980s, followed by Gaussian-Mixture-Models (GMMs) in the mid-1990s to early 2000s [25], [28], [29]. The introduction of GMM supervectors and their ability to represent a single utterance by a fixed-dimensional vector made it possible to use machine-learning classifiers for speaker recognition tasks [25], [30], [31]. In particular, support vector machines (SVMs) and various combinations thereof used supervectors [30], before i-vectors became the state-of-the-art approach [32]. By the mid-2010s, deep neural networks (DNNs) became dominant for speaker recognition due to their overall better performance, and ability to learn from unlabeled data [9]. Unlike HMMs and GMMs, the performance of DNNs improves with larger training sets, provided that the target speaker is well represented [29], [33], [34]. Deep learning in speaker recognition has been utilized for feature extraction and has replaced the i-vector with the d-vector [5], [35] and x-vector [36], [37]. This has improved the classification and comparison of speaker embeddings. The new approaches rely on large datasets such as VoxCeleb and data augmentation [5], [37], [38]. Recent work has highlighted the vulnerabilities of anonymized speaker voices, revealing that such voices can be easy to imitate and difficult to recognize, thus posing new challenges for voice biometric security [39].

B. Bias in Biometrics

Existing bias literature on biometrics mainly focuses on measuring disparate error rates across demographic groups in face recognition systems. Various measures have been proposed for doing this, such as statistical methods [16] and the Fairness Discrepancy Rate [8]. Meanwhile, others have provided checklists for measuring racial bias in face recognition, emphasizing the need to consider data-driven factors and scenario modeling including accounting for subpopulation distributions, algorithm quality, the representation of and conditions captured by images, threshold selection and appropriate considerations around demographic pairing [6].

In the voice biometrics domain, an empirical and analytical examination of bias in the machine learning development workflow of speaker verification benchmarks identified various sources of bias during the data gathering stage, and when models are deployed [9]. A follow-up study showed that the pairing of trials in speaker recognition benchmarks can result in evaluation datasets of variable difficulty across demographic groups [40]. Similar effects have been shown to lead to bias in evaluation settings in face recognition systems [6].

C. Bias in Data and Datasets

Machine learning models are impacted by bias in data and datasets, including historical, representation, measurement, and evaluation bias [17], [22]. As datasets form the basis for training, evaluating, and benchmarking models, dataset evaluations are important to interrogate bias in machine learning systems [18]. Prior research has found that the dominant developer culture, which emphasizes rapid progress and ever-larger models, can lead to representation bias in datasets and inadequate dataset documentation [18]. Similarly, evaluation failures can result from implementation variations, errors in test set construction, overfitting, and inadequate baselines [23].

While prior speaker recognition dataset studies have been published [41], [42], they focused primarily on describing available corpora and did not interrogate how data practices impact ethical and societal outcomes. Some recent studies have analysed data documentation [43] and benchmark practices [44] in speech recognition, but their findings do not account for the nuances particular to voice biometrics. As the advancement of speaker recognition research shares many common attributes with that of face recognition, this study takes inspiration from dataset evaluations in the visual domain [19], [20], [21]. We particularly draw on Raji and Fried's study of evaluation datasets and benchmarks in face recognition research [19]. In contrast to their work which surveyed over 100 face datasets, we start our study by examining (the change in) dataset usage in the speaker recognition research community over a period of time.

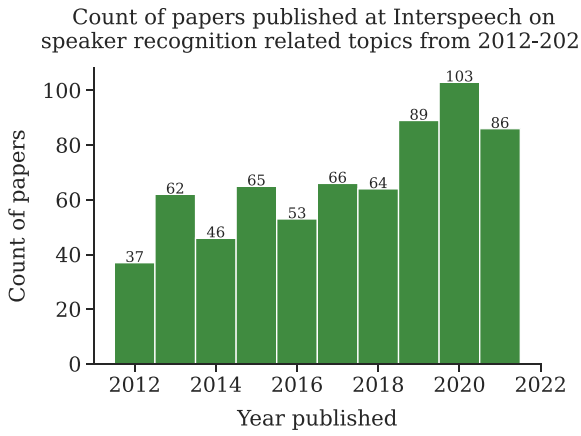

Fig. 1. Distribution of analyzed papers identified with the keywords in Section III, published at Interspeech over a decade from 2012-21.

III. Research Approach

Our study focuses on peer-reviewed research published over a ten year period from 2012 - 2021 at the Interspeech conference, one of two main international conference venues for academic and industrial speech research.

We included all papers from the International Speech Communication Association (ISCA) archive[1] that contained the search terms *speaker recognition*, *voice recognition*, *speaker verification*, *voice verification*, *speaker identification*, *voice identification*, *speaker authentication*, or *voice authentication* in their title or abstract. This query resulted in 702 papers, which we analyzed further. We excluded 29 papers that were overview papers, that studied speaker recognition by humans, applications rather than model development, or that did not explicitly mention which datasets were used to train and evaluate their models. Our final analysis thus includes 673 papers. Over the decade that we analyzed, the number of papers on speaker recognition published at Interspeech has doubled, as can be seen in Figure 1.

All papers were tagged with the training and evaluation datasets that they used. As datasets were not always named consistently, some assumptions were made. For example, studies that used datasets from the NIST Speaker Recognition Evaluations (SREs) often rely on several of these datasets for training and refer to them as a range (e.g., NIST 2004 - 2008). In these cases, we assumed that every dataset in the range was used for training or evaluation, as indicated by the authors. Overall, we encountered many naming inconsistencies. Especially evaluation datasets were sometimes only referred to on the highest level (e.g., VoxCeleb), without specifying which dataset version, subset or evaluation protocol was used. Whenever possible, we standardized dataset and subset names and otherwise tagged papers by their training and evaluation dataset *family*. The dataset family name was created by manually cleaning the dataset names, and then extracting the first word in the name as the family name. It is common in the speech processing domain to refer to training datasets as development data or corpora, and to evaluation datasets as test data. In this paper, we use the terms training and evaluation datasets, unless we refer to the names of specific datasets.

IV. Speaker Recognition Dataset Dynamics Over a Decade of Use

We now examine which datasets have been adopted over the past decade, and how the use of datasets for training and evaluation purposes has changed over time.

A. Adoption of Speaker Recognition Datasets

Over the past decade, a wide range of datasets has been used for training and evaluating speaker recognition systems. In total, the literature references 185 unique training and 164 unique evaluation dataset families. Despite this variety, a small number of dataset families has dominated speaker recognition research, as can be seen in Figure 2 which shows the frequency counts of the top 30 dataset families used for training. As papers can use more than one dataset to train and evaluate systems, the total count of use of dataset families exceeds the number of analyzed papers.

The NIST Speaker Recognition Evaluation (SRE) corpora dominate both training and evaluation. These corpora are not unique datasets in their own right, but rather collections and subsets of other datasets, predominantly Switchboard and Mixer. The NIST SREs were both users and drivers of these dataset collections, as annual evaluation challenges required new datasets to evaluate speaker recognition technology in ever more difficult settings [45]. We have kept the NIST SRE labels distinct from Switchboard and Mixer to stay true to the naming conventions used by researchers. Moreover, the NIST SREs typically required specific settings for training and evaluation that did not necessarily include the entire datasets. Switchboard occurs second most frequently for training, but surprisingly is only rarely used for evaluation. The reason for this is that speakers appear across multiple recordings in different dataset releases. As it is not possible to connect speakers between the various releases, using Switchboard for training and evaluation thus has potential for data leakage, which diminishes the quality of an evaluation.

VoxCeleb and ASVspoof are two further dataset families that have been popular for training and evaluation. ASVspoof datasets have been released by the Automatic Speaker Verification and Spoofing Countermeasures Challenge[2], which was launched in 2015 to address growing concerns of security breeches in speaker verification technology due to voice spoofing and deepfakes.

Moving from dataset families to individual datasets, Table I shows the top ten training datasets. For the NIST SREs we show whether they draw on the Switchboard or Mixer corpora. Overall, individual papers trained on a far greater number of datasets than what they evaluated on. While this in itself is not surprising, it is concerning that the majority of papers used only a single dataset for evaluation, as shown in Figure 3. Papers that evaluated on more datasets rarely used more than three. While speaker recognition development on limited

[1]https://www.isca-speech.org/archive/

[2]https://www.asvspoof.org/

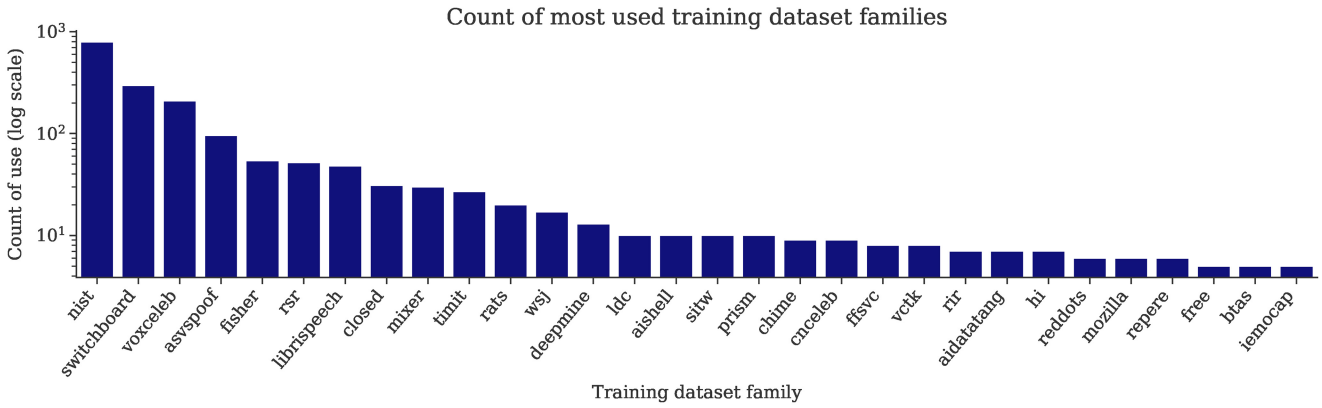

Fig. 2. Histogram showing the count of most used training dataset families from 2012 - 2021.

TABLE I
MOST FREQUENT TRAINING DATASETS

Dataset	times used
NIST SRE 04 (Switchboard, Mixer)	136
NIST SRE 05 (Mixer)	133
NIST SRE 06 (Mixer)	127
NIST SRE 08 (Mixer)	96
VoxCeleb 2	76
VoxCeleb 1	52
Switchboard (version not specified)	48
Switchboard Cellular 2	43
VoxCeleb 1 - dev	43
NIST SRE 10 (Switchboard, Mixer, etc.)	40

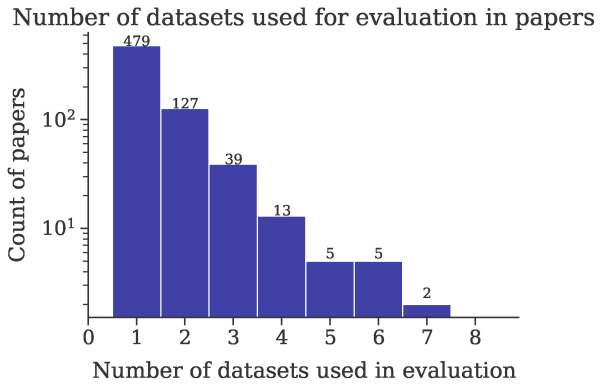

Fig. 3. Histogram of number of datasets used for evaluation by each paper that we reviewed. Most papers use a single dataset.

corpora that target specific and evolving tasks over an extended period of time may have been justified to advance the field prior to the adoption of deep neural networks [42], the same practices today will lead to overfitting. The limited diversity and use of evaluation datasets is reminiscent of evaluation failures in machine learning more broadly [23], and should lead to greater scrutiny of research claims.

B. Dataset Dynamics over a Decade of Use

Next, we examine changes in dataset usage over the past decade. Where the NIST SREs, Switchboard and Mixer datasets featured prominently when aggregating dataset use over the past decade, a finer grained year-on-year analysis reveals that their popularity has declined dramatically. In their stead, VoxCeleb now dominates speaker recognition training and evaluation. In addition, the ASVspoof datasets, notably the

TABLE II
EVALUATION DATASETS USED IN SPEAKER RECOGNITION STUDIES

Year	Published papers	1 evaluation dataset only	1 VoxCeleb dataset only	VoxCeleb1 - test only
2012	37	26 (70%)	-	-
2013	62	46 (74%)	-	-
2014	46	35 (76%)	-	-
2015	65	55 (85%)	-	-
2016	53	39 (74%)	-	-
2017	66	47 (71%)	1	-
2018	64	40 (63%)	4 (6%)	2 (3%)
2019	89	59 (66%)	15 (17%)	11 (12%)
2020	103	75 (73%)	22 (21%)	13 (13%)
2021	86	55 (64%)	10 (12%)	7 (8%)

ASVspoof 2017 dataset which focused on replay attack, have influenced speaker recognition training and evaluation within recent years. Figures 4 and 5 illustrate these dataset dynamics by visualizing the proportional use of datasets in training and evaluation. These figures show densities and should be considered together with Figure 1, which shows the growth in publications and consequently total dataset usage over the decade. Thus, since 2017 more papers have been published in speaker recognition, and a greater proportion of these studies uses VoxCeleb to train and evaluate their models.

Particularly striking is the extent to which VoxCeleb1 dominates speaker recognition evaluations. The dataset is disjoint from its successor, VoxCeleb2. A popular pairing is thus to use the larger VoxCeleb2 dataset for training, and VoxCeleb1 for evaluation. In 2020 and 2021, over half of all evaluations used VoxCeleb1. More so, VoxCeleb1-test, a small subset of 40 predominantly male, U.S. speakers whose name starts with an *E* is used in a significant proportion of evaluations. As mentioned previously, papers may use more than one dataset for evaluation. We thus investigate the number of papers relying solely on VoxCeleb, or on its even more limited subset VoxCeleb1-test for evaluation in Table II. This additional analysis reveals that VoxCeleb1 is not only popular for evaluation, but that a significant proportion of studies evaluated their methodological contributions on a single VoxCeleb dataset only.

V. DATASET ATTRIBUTES

The previous section highlighted that the Switchboard, Mixer, and VoxCeleb dataset families, together with the NIST

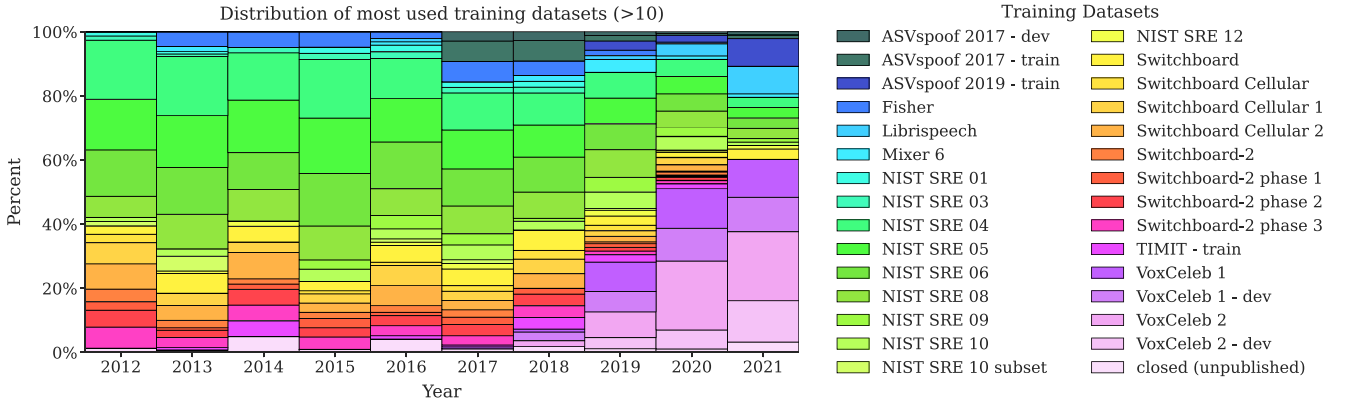

Fig. 4. Distribution (%) of dataset use for speaker recognition **training** (VoxCeleb datasets purple, NIST SREs turquoise & green). Datasets are included if they appeared in more than 10 papers. Over the past decade, the use of NIST SREs, Switchboard, and Mixer datasets has declined, with VoxCeleb datasets, and particularly VoxCeleb2, becoming the dominant datasets for training.

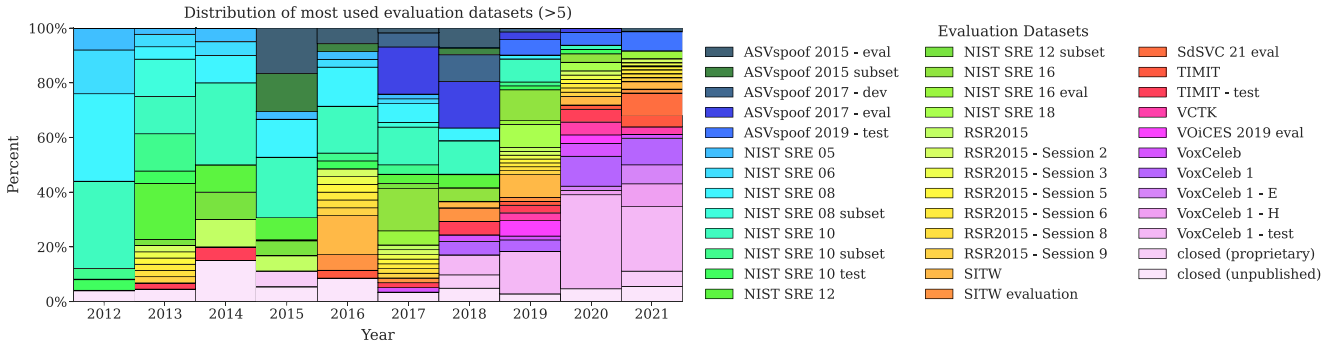

Fig. 5. Distribution (%) of dataset use for speaker recognition **evaluation** (VoxCeleb datasets purple, NIST SREs turquoise & green). Datasets are included if they appeared in more than 5 papers. From 2019 onwards, VoxCeleb1 has become the dominant dataset for evaluations. In 2020 and 2021, over half of all evaluations use VoxCeleb1, in particular the VoxCeleb1-test subset.

Speaker Recognition Evaluations (SRE) have dominated the development of speaker recognition technologies, each in its own era. Additionally, the ASVspoof datasets are increasingly adopted to address growing concerns of spoofing attacks. Therefore, we now investigate attributes of these dataset families and their influence on bias, fairness, and privacy of the technology today.

A. Background and Motivation for Corpora Collection

The collection and release of the **Switchboard** corpora started in the 1990s and continued through the early 2000s. In total, seven datasets of two-sided English language telephone conversations were released. The dataset collections were funded by the U.S. Defense Advanced Research Projects Agency (DARPA) and the U.S. Department of Defense. The Linguistic Data Consortium (LDC) was primarily responsible for data collection and management. According to the LDC, these datasets were intended for "research, development, and evaluation of automatic systems for speech-to-text conversion, talker identification, language identification and speech signal detection purposes" [46].

The Mixer and Transcript Reading (short **Mixer**) corpora succeeded Switchboard, as the collection protocol of the latter became complicated, time-consuming, and expensive. Moreover, telephone behavior of people changed as cellphones became popular [45]. The Mixer project aimed to support various speaker recognition tasks in multi-lingual and cross-channel settings.[3] Mixer was created by the LDC in collaboration with the Lincoln Laboratory, the U.S. NIST and the Speaker Identification research community.

The **VoxCeleb** datasets were a response to an increasing appetite in the speaker recognition community to test and develop their approaches in more challenging real-world (i.e., "in the wild") settings. VoxCeleb1 was released in 2017 by the Visual Geometry Group (VGG) at the University of Oxford, with the goal of creating a large scale, text-independent speaker recognition dataset that mimics unconstrained, real-world speech conditions [47]. A key driver for this was to explore the use of deep neural networks (DNNs), which had gained traction in computer vision, for speaker recognition tasks. A year later, VGG released VoxCeleb2 to expand the original data collection.

While the previously discussed datasets have been specifically constructed for the evaluation and development of speaker recognition systems, the Automatic Speaker Verification and Spoofing Countermeasures Challenge (**ASVspoof**) and its accompanying datasets were designed to encourage the development of anti-spoofing countermeasures [48]. Since 2013, this challenge has taken place every second year. The datasets released by ASVspoof are created

[3]Channel refers to the medium used for speech recordings, e.g., microphone types.

TABLE III
SPEAKER-DEPENDENT METADATA IN SWITCHBOARD DATASETS

Dataset	Switchboard Cellular Part 1	Switchboard Cellular Part 2	Switchboard 1 Release 2	Switchboard 2 Phase 1	Switchboard 2 Phase 2	Switchboard 2 Phase 3
Collection Time Frame	1999-2000	2000	1991-1992	1996	1997	1997-1998
Gender	✓	✓	Sex	Sex	Sex	✓
Age	✓	✓	Birthyear	✓	✓	✓
Years of Education	✓	✓	✓	✓	✓	✓
Country of Birth	✓	✓	Dialect Area	✓	✓	✓
City/State where raised	✓	✓	Dialect Area	✓	✓	✓
Phone Nr.	encoded	encoded	✓	✓	✓	✓
Topic	✓	✓	✓	✓	✓	✓
Calls Made	✓	✓	n.a.	✓	✓	✓
Calls Received	✓	✓	n.a.	✓	✓	✓
Date	✓	✓	✓	✓	✓	✓
Start Time	n.a.	n.a.	✓	✓	✓	✓
Record Length	✓	✓	✓	Call Duration	Call Duration	Call Duration
Talk Length	✓	✓	n.a.	n.a.	n.a.	n.a.
Connect Length	✓	✓	n.a.	n.a.	n.a.	n.a.

using existing speech datasets, e.g., RedDots corpus [49] or Voice Cloning Toolkit (VCTK) [50], [51]. Spoofed speech is then generated with voice conversion and speech synthesis algorithms as well as replayed versions. In 2021, ASVspoof extended its focus to include tasks independent of ASV systems to promote the detection of deep fakes more generally. In 2023, ASVspoof encouraged contributions for generating spoofed speech [52].

B. Attributes and Usage

Next, we analyse the metadata of the Switchboard, Mixer, VoxCeleb and ASVspoof datasets. The analysis on which this section is based is available as a Jupyter notebook.[4] We distinguish between *speaker-dependent* and *speaker-independent* attributes captured in the metadata. Speaker-dependent attributes are factors that are inherent to a particular speaker, such as their gender or accent, and that can influence a person's speaking style. Speaker-independent factors constitute environmental acoustics or instrument-related variabilities, such as background sounds or phone models.

1) Switchboard: Most Switchboard datasets typically exceed 2,000 recordings, totaling over 100 hours of speech. However, the Switchboard Credit Card dataset[5] is an exception, containing only 35 recordings and 227 minutes of speech data. These datasets were initially valued as rich sources for training speaker recognition models. An overview of speaker-dependent attributes captured by the different Switchboard datasets is shown in Table III. In addition to the speaker-dependent information, Switchboard metadata provides speaker-independent attributes such as environmental acoustic factors as shown in Table IV.

Speaker Dependent Attributes: The demographic metadata in Switchboard includes age, gender, years of education, country of birth, and location where the person was raised. These are speaker-dependent attributes that affect a speaker's style and that convey paralinguistic or extra-linguistic information. Additionally, call-specific details such as the recording date and start time are included in this category, recognizing that a speaker's style may fluctuate throughout the day. The frequency and duration of calls made or received reflect a participant's engagement in the data collection process. This has a critical effect on speaker representation in the dataset. For instance, the number of utterances per speaker can vary significantly based on their call activity. Importantly, the representation of individual speakers in a dataset can impact speaker verification systems trained or evaluated on the data [40]. The discussion topics during calls, linked to the linguistic content exchanged by the speakers, are another speaker-dependent element. Lastly, phone numbers are considered speaker-dependent due to their unique link to individual speakers within the dataset.

Demographic Representation: The gender distribution across males and females in Switchboard is generally balanced on a speaker level, but not reported for the number of conversation or the duration of recorded speech. Switchboard Cellular Part 2 Audio reported gender demographics across recordings and has an overall balanced split in male and female representation across the dataset versions (with most years overrepresenting females by 5-10%). The age distribution leans towards younger demographics with some variation based on gender and dataset version.

In Switchboard 1 Release 2, detailed dialect area annotations contrast with other dataset versions that only list speakers' birthplaces and locations where they were raised. "South Midland" is the most reported dialect in this dataset and regions like "North Midland" and "Northern" are predominantly represented by male speakers. In Switchboard Cellular Part 1 and Part 2, despite most speakers being U.S.-born, there is a noticeable difference in birthplaces, with the majority of

[4]https://github.com/wiebket/bt4vt/tree/metadata_analysis
[5]Excluded from extensive attribute analysis.

TABLE IV
SPEAKER-INDEPENDENT METADATA IN SWITCHBOARD DATASETS

Dataset	Switchboard Cellular Part 1	Switchboard Cellular Part 2	Switchboard 1 Release 2	Switchboard 2 Phase 1	Switchboard 2 Phase 2	Switchboard 2 Phase 3
Channel	✓	✓	n.a.	✓	✓	✓
SNR Estimates	n.a.	✓	n.a.	n.a.	n.a.	n.a.
Difficulty	n.a.	n.a.	labelled from 0 to 5	n.a.	n.a.	n.a.
Topicality	n.a.	n.a.	labelled from 0 to 5	n.a.	n.a.	n.a.
Naturalness	n.a.	n.a.	labelled from 0 to 5	n.a.	n.a.	n.a.
Echo	labelled from 0 to 3	labelled from 0 to 3	labelled from 0 to 5	Echo or Crosstalk labelled from 0 to 3	Echo or Crosstalk labelled from 0 to 3	Echo or Crosstalk labelled from 0 to 3
Background Noise	labelled from 0 to 2	labelled from 0 to 2	labelled from 0 to 5	labelled from 0 to 3	labelled from 0 to 3	labelled from 0 to 3
Distortion	labelled from 0 to 2	labelled from 0 to 2	Electrical Noise (Static) labelled from 0 to 5	labelled from 0 to 3	labelled from 0 to 3	labelled from 0 to 3
Location	1=indoors 2=outdoors 3=moving vehicle (reported by participants)	1=indoors 2=outdoors 3=moving vehicle (reported by participants)	n.a.	n.a.	n.a.	n.a.
Phone Manufacturer	✓	✓	n.a.	n.a.	n.a.	n.a.
Phone Model	✓	✓	n.a.	n.a.	n.a.	n.a.
Service Type	✓	✓	n.a.	n.a.	n.a.	n.a.
Service Provider	✓	✓	n.a.	n.a.	n.a.	n.a.

speakers citing "Philadelphia" as their place of birth. While the country of birth can serve as an approximate indicator of a person's dialect, it can potentially mislead evaluations due to the presence of diverse dialects within the same geographic origin.

We observed disparities in call frequency across individual speakers in the Switchboard datasets. Specifically, while Switchboard Cellular Part 1 and 2 have few outliers, Switchboard 2 Phases 1 and 2 displayed significant variations, with some individuals making or receiving over 30 calls compared to the average of 10 calls per caller. Such disparities can result in datasets that are imbalanced at the utterance level, despite appearing demographically balanced at the speaker level.

Speaker Independent Attributes: Switchboard recordings are annotated for channel quality (e.g., echo, crosstalk, static) and background noise. Annotations vary across datasets, with a majority indicating minimal to no acoustic environmental influence. Notably, the level of echo or background noise in calls does not significantly differ by speaker gender, suggesting a consistent quality of data collection across demographic subgroups.

In Table IV, Switchboard Cellular Part 2 includes annotations of the speaker's environment. The data is predominantly categorized as indoor, with relatively few outdoor or vehicle-based calls. Despite this skew towards indoor environments, the gender distribution remains balanced, with a slight overall tilt towards female speakers. Earlier versions of the Switchboard datasets record landline conversations with a variety of telephones and the later datasets contain cellphone conversations. Although most calls do not specify the phone model, Motorola, Ericsson and Nokia are the most frequently cited brands in Switchboard Cellular Part 1 and 2. The datasets aimed to encompass channel variability; however, our analysis of the metadata suggests that they are limited for robustness assessments, due to an uneven distribution of phone-related metadata. Additionally, acoustic quality can be affected by factors related to the environment and transmission which are not captured in the metadata, such as room acoustics, reverberation, recording quality and compression techniques [25].

2) Mixer: Mixer recorded significantly more data than Switchboard, with individual datasets capturing between 5 000 and 20 000 calls, resulting in tens of thousands of hours of speech. The early phases of the Mixer corpora focused on multi-lingual data collection, before shifting focus towards multi-channel set-ups. Similar to Switchboard, speaker-dependent and speaker-independent attributes were captured. At the time of writing this paper the authors had access to metadata capturing speaker-dependent information only. Publicly available information on speaker-independent attributes mostly contains details on the multi-channel set-up and recording devices.[6]

Speaker Dependent Attributes: Speaker-dependent attributes in the Mixer corpora go beyond those in Switchboard including sex, year of birth, education, occupation, ethnicity, height and weight, smoking status and information about the speaker's family. Given this detailed metadata, the Mixer collection has been used for age estimation [53], smoker identification [54], [55] and for predicting speaker demographics from word usage [56]. Annotations for these attributes are consistent across different versions of the Mixer corpora, making it possible to align and analyze metadata across different data releases.

Demographic Representation: The sex distribution of Mixer 3 (collected in 2006) is skewed towards female speakers on a speaker level, but is not reported on an utterance level. For Mixer versions 4, 5 and 6 (collected between 2007 - 2010), the sex distribution is more balanced. Similar to Switchboard, the age distribution leans towards younger demographics. In Mixer 3, the peak year of birth is around the 1980s (implying a speaker age around 26 years). For Mixer 4, 5 and 6 which were collected later, year of birth peaks also appear later, which implies speakers of a similar age participated in the collections.

[6]https://catalog.ldc.upenn.edu/docs/LDC2020S03/readme.txt; https://catalog.ldc.upenn.edu/docs/LDC2013S03/readme.txt

While age is a significant factor in speech, collecting the year of birth rather than age makes metadata analysis and use more cumbersome, as the collection date of the corpus needs to be known and considered.

Unsurprisingly, Mixer 3 contains a variety of speakers with different native languages, while later Mixer corpora are dominated by native English speakers. Regarding smoking status, there are notable differences between Mixer versions, with Mixer 3 leaning towards smokers and Mixer 6 towards non-smokers. Many metadata fields are empty. Over 80% of speakers did not report their smoking status and 99% did not report their education degree or family information. This underscores a tension between collecting rich metadata, which is only useful if it is complete, and collecting sensitive and personal information, which should only be done on a voluntary basis, and in a privacy-preserving manner.

3) VoxCeleb: The VoxCeleb datasets were scraped from celebrity YouTube videos to capture a large number of audio clips where people speak in unconstrained settings. VoxCeleb1 consists of 153,516 speech utterances from 1,251 speakers. VoxCeleb2 contains 1,128,246 utterances from 6,112 speakers. The creators of the dataset promote its use for speaker identification and verification, speech separation, talking face synthesis, cross-modal transfer between face and voice (i.e., making inferences about somebody's face based on their voice, and vice versa), emotion recognition and face generation.

The only metadata available for VoxCeleb1 are gender and nationality labels, while VoxCeleb2 only has gender annotations. Thus, all available metadata constitutes speaker-dependent factors only, making further analysis of speaker-independent attributes and their interplay with speaker-dependent attributes not possible. Future research could use signal-to-noise (SNR) estimators to evaluate the acoustic features of the recordings and evaluate their interaction with speaker-dependent and demographic characteristics. The dataset descriptions are not transparent about how gender labels were obtained, but it is likely that they came from VGGFace1 and 2 [57], [58], which provided the candidate list of speakers to include in VoxCeleb. The nationality labels were inferred from speakers' countries of citizenship, as obtained from Wikipedia. The motivation for doing this was to assign a label that is indicative of a speaker's accent [59]. The authors claim that the datasets are gender balanced, with 55% and 61% male speakers in VoxCeleb1 and 2 respectively. However, subsequent research has pointed out that VoxCeleb1, and in particular the VoxCeleb1-E, -H and -test subsets suffer from representation bias on a speaker and utterance level, across genders and nationalities [9].

An important difference between the VoxCeleb datasets and the other two corpora is that VoxCeleb used to be freely available for download.[7] By contrast, Switchboard and Mixer require a subscription to the LDC ($3 850 for universities, $27 500 for corporations) or must be purchased. Licensing costs for an individual dataset range between $100 to $300, however, not all datasets can be licensed without an LDC membership. This made the VoxCeleb datasets the first large scale, freely available datasets for speaker recognition. It is plausible that the free availability of VoxCeleb greatly contributed to its adoption in the research community.

4) ASVspoof: The ASVspoof datasets primarily sourced speech from existing collections to generate spoofed speech samples. Focusing on the generalizability of countermeasures, ASVspoof has employed various voice conversion and speech synthesis algorithms to construct spoofed speech. For instance, ASVspoof 2015 and 2019, use genuine English speech from 106 and 107 speakers respectively, with a gender distribution of 45/46 male and 61 female speakers [48] [60]. In 2017, the replay recordings of the RedDots corpora were collected in controlled and uncontrolled environments using various playback devices, but only for male speakers [61]. A later collection has focused on constructing replay spoofed speech corpora through different simulated acoustic configurations using a more gender-balanced speech sample [60].

Depending on the dataset release, metadata for ASVspoof is limited to gender information and information on the speech spoofing system or the replay configuration. With gender being the only demographic information, bias evaluation related to speaker-dependent factors becomes challenging. For speaker-independent factors, such as recording devices used for replay attacks, the distribution is slightly skewed towards high-quality devices, which may be more difficult to detect [62]. Moreover, in recent years, ASVspoof has emphasized the development of spoofing countermeasures that are robust to variabilities in codec and transmission channel [63]. This is similar to the collections of the Switchboard and Mixer corpora which have specifically focused on capturing channel and recording device variability. The latest version of the challenge introduced two new base datasets: the English-language subset of Multilingual Librispeech (MLS) [64] and an optional subset of the English Common Voice Corpus 11.0 [65].

VI. Ethical Concerns

Our metadata analysis has highlighted that the Switchboard and to a lesser exten the Mixer corpora are representative across some demographic attributes, but also contain an extensive amount of sensitive and personal information. By contrast, VoxCeleb and ASVSpoof contain almost no metadata, but their collection methods indicate that these datasets are not representative across demographic groups. In this section, we identify additional potential biases, privacy risks and ethical concerns associated with these datasets.

A. Influence of Collection Method on Bias

The four dataset families that we analyze span across four different data collection paradigms: direct data collection in a lab setting, crowd-sourcing, webscraping and finally synthetic data generation. This makes an examination of the impact of the data collection method on bias particularly interesting. Various sources of bias can arise in datasets and their collection processes [17]. For example, historical bias

[7]The public download link has now been replaced with a privacy note on the VGG website.

reflects the influence of existing societal biases on datasets, while representation bias arises when a dataset inadequately mirrors the target population. Measurement bias is linked to flawed or overly simplified data features or labels. We investigate these three sources of bias, and also discuss other types of biases that might have arisen, or that were accounted for in the data collection process.

1) Switchboard: For Switchboard and Mixer, participants were recruited to meet the language requirements of the study and received financial compensation for participating. While the Switchboard collection offered a free call with a fixed compensation (in 2005 it was $1/minute for a maximum of a 10 minute toll-free call to a friend), the Mixer collection changed compensation to a per-call incentive and completion bonuses [66]. The free phone call was a strong incentive to participate when the Switchboard data collection started. However, after the turn of the millennium it lost its appeal as phone calls and mobile phones were near ubiquitous [45]. Between the releases of the Switchboard and Mixer datasets, the team of data collectors had continuity. This has led to consistency in the data collection method. However, it also implies that cognitive biases of individuals who shaped the design of the dataset collections can have gone unnoticed.

Although the data collectors for the Switchboard corpora aimed to obtain gender balanced datasets, our analysis reveals imbalances in other speaker-dependent and independent factors. For instance, the datasets primarily feature native English speakers from the American South, with a significant proportion of college students. This skews the demographic representation in the corpora towards younger and highly educated individuals. Nonetheless, the extensive metadata available for Switchboard enables in-depth analysis of underrepresented subgroups, which is crucial for assessing the generalizability and robustness of speaker verification models.

Switchboard's annotation protocols and labelling taxonomy are not disclosed, which makes it unclear how categories like "country where raised" were determined. In particular, it remains unclear how data annotators were instructed and whether inter-annotator agreement was assessed. The lack of clarity on labels and labelling protocols can introduce measurement bias when using these datasets for bias evaluations. In addition, the different Switchboard releases make different labelling choices for speaker-independent factors such as echo, background noise or distortion. Most commonly these annotations are evaluated on scale of 0 to 3. Switchboard 1 however uses a scale of 0 to 5, and the Cellular datasets employ a 0 to 2 scale for background noise and distortion. This makes comparisons within the dataset family challenging and can lead to innacurate comparisons across datasets.

2) Mixer: English has always been a dominant language in speaker recognition. The first three phases of the Mixer project thus focused on collecting multilingual data from bilingual speakers. 16% of Mixer calls in Phases 1 and 2 are in Arabic, Mandarin, Russian or Spanish [67]. The Mixer 3 collection also aimed at supporting language recognition, and had more than 2900 participants making calls in 19 different languages [45]. The defense backing of the Mixer datasets is evident in the languages that were selected for the project, and their connection to U.S. national security and military interests. The different Mixer corpora include conversations in Arabic, Egyptian, Farsi, Bengali, Hindi, Urdu, Tamil, 4 dialects of Chinese (also Mandarin), Japanese, Korean, Tagalog, Thai, Vietnamese, German, Italian, Russian, 3 dialects of English (including American English), Spanish and Canadian French.

Mixer 4, 5 and 6 feature a wider variety of channels and recording scenarios than its predecessors. As a consequence, the variety of languages and accents decreased and the collections focused on native speakers of American English only [68]. The data collectors attempted to balance dialects by recruiting 25% of participants from Philadelphia, 25% from Berkeley, and specifically from Texas, Georgia, Illinois, and New York [68]. On-site recordings for Mixer 4 and 5 were carried out at two different locations, the LDC in Philadelphia, Pennsylvania, and at the International Computer Science Institute (ICSI) in Berkeley, California. Recruitment for Mixer 6 was done at the LDC [68], [69], thus decreasing the likelihood of collecting speech samples from speakers of various dialects and increasing the likelihood of representation bias.

3) VoxCeleb: The VoxCeleb datasets were constructed with a fully automated data processing pipeline from audio-visual media scraped from YouTube [47], [59]. Both data pipelines consist of the same processing steps: first select a list of candidate speakers, then download videos from YouTube, apply face tracking, identify active speakers, verify identities from faces, remove duplicates, and finally find associated nationality metadata on Wikipedia. The candidate speakers for the datasets were sourced from VGGFace1 [57] and VGGFace2 [58] respectively. In a previous study, a comprehensive analysis of historical, representation and measurement bias in VoxCeleb1 was carried out [9]. The study highlights that this automated processing pipeline reinforces popularity bias from search results in candidate selection, and directly translates bias in facial recognition systems into the speaker recognition domain. Moreover, celebrities, especially actors and singers, have a high degree of control over their voice and accent, and should not be assumed to represent ordinary conversational speech. VoxCeleb2 is likely to show similar biases as previously identified for VoxCeleb1 due to a similar data collection approach.

4) ASVspoof: The ASVspoof datasets deviate from traditional speech data collection methods, primarily employing voice conversion and speech synthesis algorithms. Their reliance on pre-existing speech datasets raises the possibility of transferring historical bias from these datasets into ASVspoof datasets. The selection of genuine speech samples can significantly influence the generalization capabilities of anti-spoofing countermeasures across speaker subgroups. This is particularly important as the early versions of ASVspoof datasets have featured only a limited number of speakers. A striking example of representation bias is observed in ASVspoof 2017, based on the RedDots replayed spoofing corpus, which features only male speakers [61]. This gender imbalance appears to be a deliberate choice by the dataset creators rather than a historical artifact, considering that the original RedDots corpus features both male and female speech samples. Even if the dataset creators can motivate this choice, dataset users may

be unaware of the male-centricity of the dataset. It is unclear whether the selection of speakers in that dataset sufficiently represents the various channel, session, and accent variations provided by the original RedDots corpus [49].

In contrast, the replay spoofing dataset for ASVspoof 2019 [60] shows a more balanced gender representation. Even though ASVspoof datasets typically dominate evaluations only in their challenge year, the potential reuse of prior releases still warrants careful scrutiny of representation bias. In addition to gender, language representation in ASVspoof is limited, as the datasets have exclusively concentrated on English language speech for both text-dependent and text-independent speaker recognition scenarios [70]. Finally, as the synthetic data creation method is highly reliant on algorithms and data-driven systems, which are trained on datasets that likely have their own challenges with representation bias, it is yet unclear how bias inherent in these systems influences the quality of spoofed speech datasets and thus, the development of countermeasures.

B. More Than Bias: Privacy Risks and Ethical Questions

1) Risk of Re-Identification: The Switchboard and Mixer dataset collections passed an institutional ethical review[8] and the LDC kept personal information like names and contact details separate from the recordings [67], [68]. Nonetheless, the two corpora would today be considered as posing significant privacy risks to study participants. The privacy risks stem from two sources, firstly the content of the conversations and secondly the rich metadata, which makes it possible with today's data processing techniques to retrospectively correlate personal attributes with voice characteristics. The amount of personal information stored in the metadata is quite extensive and makes it possible to use these datasets for various tasks that include the identification of personal information from speech data in future.

At the time of data collection, the ethical consequences and privacy concerns due to the extent of personal information contained in the voice may not have been clear to researchers. However, a decade of progress in speech science has changed that [71]. We are not aware that any efforts have been made to address the presence of personal identifiable information and sensitive attributes in the recordings, to assess risks of re-identification, and to examine the potential impact on data subjects. Moreover, it remains unclear whether a combination of factors classified as speaker-independent information can lead to re-identification or leak sensitive information. For example, inferences made from background noise can reveal context information or personal information [72]. In today's data-driven society, privacy and anonymity of data subject are vital concerns that require attention and proper measures. Therefore, future collections of speech corpora should consider a privacy-bias trade-off and consciously decide on their collection of speaker-dependent and independent factors.

2) Risk of Disclosing Sensitive Information: During data collection, participants were asked to discuss a specific topic with an automated operator on a phone call, but to withhold personal information. Yet, the topics provided for discussion included political, cultural, social and religious topics.[9] For instance, the topics annotated in the Switchboard Cellular Part 1 and 2 range from education and leisure activities to domestic politics and international news. While there are inconsistencies in topic labels and provided metadata (e.g., topics above 61 have been annotated but are not shown in the topics list), our analysis on the topic distribution shows that a majority of calls have discussed hypothetical situations or domestic politics. These categories include questions on preferences regarding smoking bans, minimum wages and personal wishes which can make participants disclose sensitive information. Another category of interviews took the form of informal conversations, adapted from sociolinguistic interview modules. Here subjects were encouraged to describe events of the past [69]. This form of interview creates the illusion of an informal setting, making it more likely that participants share personal information with interviewers [74]. While participants were not forced to discuss the topic provided, most of them followed the suggestion [68]. Moreover, participants might have shared sensitive information during phone calls or interviews, increasing the risk of re-identification. For instance, Mixer 5 interviews covered family and personal history, raising the likelihood of participants sharing personal stories with similarly sensitive information as collected in the metadata. To our knowledge calls were not redacted to exclude personal information.

3) Lack of Consent: The VoxCeleb datasets present different privacy concerns. As has been the case with other Web-scraped datasets, the dataset creators did not obtain consent from data subjects to use their biometric data for the purpose of technology development. Initially, the creation of the datasets was justified by the data being available on the public Internet. More recently, the authors have added a privacy notice to their website, calling on a data protection exemption of the University of Oxford based on Article 14(5)(b) of the U.K. GDPR, which allows data processing for scientific or historic purposes. Considering the sensitive nature of voice data, the military and security foundations of speaker recognition, and the wide-scale application in the surveillance industry, it seems prudent to interrogate whether this exemption ought to apply to voice data collected for speaker recognition purposes.

4) Security Risks: The collection of spoofed speech corpora comes with major ethical concerns which have been acknowledged in the evaluation plan of the ASVspoof 2023 [52]. While the challenge organizers ask contributors to responsibly report vulnerabilities, detailed ethics guidelines are missing. Meanwhile, the cybersecurity field has established standards and protocols for responsible vulnerability disclosure [75]. Along the same lines, codes of conduct and codes of ethics have been intensively discussed for ethical hacking [76], [77] and for computing professionals more generally. For instance, the Association for Computing Machinery

[8]Guidelines of the Institutional Review Board of the University of Pennsylvania.

[9]In the European Union the General Data Protection Regulation (GDPR) considers personal data revealing ethnic origin or religious beliefs as particularly sensitive and allows processing only on certain legal bases [73].

(ACM) has adopted a Code of Ethics and Professional Conduct to guide ethical usage and development of computing technology.[10] The speech domain, particularly speech spoofing, could benefit from established practices in other fields to foster ethical development and research. This is especially important as the data used for training spoofing attack algorithms stems from people who might have donated their voice sample with a specific purpose in mind, and would not have consented to other usage scenarios.

C. Implications of Dataset Reuse

Our analysis of ASVspoof shows that datasets for deepfake detection are typically sourced from existing datasets such as VCTK or Mozilla's Common Voice Corpus [52], [60]. Similarly, the Voice Privacy Initiative which focuses on advancing voice anonymization techniques, i.e., suppressing a speaker identity while preserving linguistic content and naturalness of the speech signal,[11] relies on several voice biometrics corpora, including VoxCeleb 1 and 2, for training, developing and evaluating voice anonymization systems [78]. While voice biometric dataset reuse is understandable given the time and cost of data collection, reuse tends to be limited to a few popular and openly available datasets. For instance, the NIST SRE datasets have not been used in any of the discussed challenges. This reliance on a few datasets increases the risk of biased models across various voice-processing tasks.

Recent trends in relaxing training data policies across challenges and allowing participants to train on external or pre-registered data can help to increase data diversity. However, increasing training dataset sizes may also pose challenges by restricting the choices for evaluation datasets. For instance, deepfake detection and voice anonymization systems often rely on similar datasets for evaluation, such as LibriSpeech or VCTK. This overlap, and the over-reliance on a small number of evaluation datasets across challenges, hinders reliable evaluations and the ability to detect bias [79]. There also exist feedback loops between datasets and voice biometric systems used for evaluation across tasks. For instance, in recent releases of ASVspoof, ASV systems for assessing countermeasures have been trained on VoxCeleb 1 and 2 [80]. Similarly, in the context of voice anonymization, privacy is assessed through an ASV system trained on subsets of LibriSpeech [81]. If ASV models reproduce bias in their training data, using these models for evaluation questions the reliability of voice anonymization and deepfake detection evaluations.

D. Voice Biometrics in the Era of Generative AI

Despite challenges like ASVspoof, the limited advances in speaker verification datasets suggest that voice biometrics will not be able to keep up with the rapid rise of voice cloning. In particular, generative AI breakthroughs are enabling speech generation systems that are capable of synthesizing very realistic voices. This has led to a sharp increase in voice cloning related attacks within the last years [82]. For example, an Australian journalist cloned their voice with just 4 minutes of audio data to break into their own self-service government accounts. While no real harm was done, it exposed the immense vulnerability of the Australian Bank and Tax Office [83]. Similarly, another journalist used voice cloning to trick voice biometrics systems of banks in the EU and U.S. [84].

While the statistics of successful voice cloning attacks on banks and government services evade us, we know that voice cloning "keeps Bruce Reed, chief AI strategist of the U.S., up at night" [85]. Yet, neither the attacks, nor the fragility of speaker verification systems should come as a surprise to the research community. For example, a study that critically assessed the VoxCeleb datasets [40] showed that the evaluation pairs that researchers construct from VoxCeleb1 to evaluate speaker verification systems are inadequate for modern applications of voice biometrics. While research has advanced the development of algorithms and models, only limited efforts and investments have been made to advance datasets and evaluation practices.

Voice cloning leaves civil society at risk, more so, if a person's unencrypted voice data is in the public domain. Would participants of the Switchboard and Mixer corpora still consent to their data being collected and processed, if they were well informed about the capabilities and risks of voice cloning today? In today's age, is Web-scraping of voice data not only a violation of privacy, but also a violation of personal security, similar to placing your credit card details in the public domain? Is it responsible to make voice data public given the limited capabilities of current voice anonymization techniques? The voice biometrics research community urgently needs to invest in the development of datasets that reflect modern applications and users. Simply collecting more data in the same way, however, is not the answer. Future datasets need to simultaneously address diversity and representation, privacy and security requirements, while being sufficiently challenging for real use cases.

VII. Discussion

This study highlights how the shift to deep neural networks in speaker recognition has led to changes in research and data practices. For over two decades, the NIST SREs, Switchboard, and Mixer datasets have significantly influenced speaker recognition research. An important focus of this period was to address audio processing challenges and reduce intra-speaker variability to ensure robustness of voice biometrics technology [42], [86], [87]. NIST's evaluation-driven research agenda evolved alongside technology advancements, considered varied task environments, collection devices, background noise, and room acoustics. However, inter-speaker differences related to demographics or other speaker-dependent attributes were considered secondary.

Since 2018, our analysis clearly illustrates the rise to dominance of the VoxCeleb datasets for training and evaluation in research contributions published at ISCA's Interspeech Conference. These datasets met the demand of researchers to develop speaker recognition systems with deep neural networks for unconstrained, "in-the-wild" settings. Has the prioritization of studying such in-the-wild settings potentially

[10] https://ethics.acm.org/

[11] https://www.voiceprivacychallenge.org/

come at the cost of developing systems that cater to diverse users? Building robust systems to address inter-speaker variability is crucial to avoid biased and discriminatory systems from being deployed in critical applications, such as financial systems and voice-activated emergency response. Yet, the shifts that we observed resemble observations made about data practices in machine learning research more generally [18].

A. Recommendations

Speaker recognition technology offers benefits but also poses potential risks and harms depending on its deployment and use. When used for voice-based authentication and access control, it is crucial to ensure the technology works for all users. To do this, representative evaluation datasets are necessary. To promote fairness and reduce bias when curating voice biometrics datasets, valuable insights can be drawn from prior work in facial recognition [19]. Diverse and representative datasets that accurately reflect the demographics of the population being served are needed. To create representative benchmarks, demographic factors should be considered alongside other speaker-dependent and independent characteristics. Factors such as age, gender, accent and language, and their intersections [88], should be considered when selecting people for dataset collection. Comparable recommendations regarding benchmarks have been discussed for speech recognition research [44] and could inform the development of diverse benchmarks for voice biometrics. Furthermore, representation should be ensured at the speaker and utterance level to ensure equitable evaluation across demographic groups [40]. Additionally, dataset collection procedures and dataset attributes should be documented carefully, for example by adopting datasheets [43], [89] for voice biometrics datasets. Further research is needed to understand application-specific requirements and how to incorporate these into evaluation protocols.

Beyond performance disparities, speaker recognition systems contribute to a hidden and pervasive surveillance infrastructure that enables governments and corporations to identify citizens and extract sensitive personal information from their voice. From a surveillance perspective, speaker recognition technology poses privacy risks to citizens. Striving for more representative datasets or detecting and mitigating bias can unintentionally increase harm to citizens rather than reduce it. Data collectors should pay attention to the privacy-bias tradeoff and critically examine the need for collecting sensitive information [90], [91]. Finally, continued research efforts are needed to enable private and privacy-preserving voice processing. In particular, given recent advances in voice cloning [92], anti-spoofing research should consider bias and fairness to ensure that all demographic groups are adequately protected.

B. Limitations

We acknowledge limitations in our study and research approach, focused solely on publications from the ISCA Interspeech Conference, which might miss broader dataset dynamics across the voice biometrics community. Including analysis from the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and biometrics venues could enrich this study. Nonetheless, we believe our findings to be broadly representative of general trends in speaker recognition research. By analyzing peer-reviewed research publications, we assume that dataset dynamics in the research domain are also indicative of adoption and attitudes towards datasets in the voice biometrics industry. Our assumption that research practices extend to industry where they may lead to bias in deployed applications is speculative. Our focus in this study was on dataset dynamics, and we did not consider evaluation protocols and metrics, which are also important in speaker recognition evaluations. Further studies and technology audits are necessary to evaluate bias in speaker recognition, and to enable accountability, transparency, and auditability of speaker recognition systems.

VIII. Conclusion

Our research provides a comprehensive overview of the evolution of speaker recognition datasets used for training and evaluation over the past decade. By analyzing the adoption, dynamics, and attributes of these datasets, we have identified issues related to bias, fairness, and ethical concerns in speaker recognition research. Importantly, these insights shed light on how data practices in research may influence downstream development of voice technologies, raising awareness about potential bias, privacy and security concerns. Our findings emphasize the importance of ongoing investigation into dataset attributes and usage, particularly in light of current research practices in a data-centric era. Finally, our study highlights the need for new datasets, that carefully trade-off challenging, modern deployment scenarios with considerations for ethics and fairness.

References

[1] "Voice print system privacy policy." TD Personal Banking. Accessed: Sept. 12, 2024. [Online]. Available: https://www.td.com/ca/products-services/investing/td-direct-investing/trading-platforms/voice-print-system-privacy-policy.jsp

[2] "Voice recognition banking: See how our voice ID works." BMO. Accessed: Sept. 12, 2024. [Online]. Available: https://www.bmo.com/main/personal/bank-accounts/voice-id/

[3] "This employee ID badge monitors and listens to you at work—Except in the bathroom," Thomas Heath. 2016. [Online]. Available: https://www.washingtonpost.com/news/business/wp/2016/09/07/this-employee-badge-knows-not-only-where-you-are-but-whether-you-are-talking-to-your-co-workers/

[4] K. Seaborn, N. P. Miyake, P. Pennefather, and M. Otake-Matsuura, "Voice in human–agent interaction: A survey," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–43, 2021. [Online]. Available: https://doi.org/10.1145/3386867

[5] Z. Bai and X. L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Netw.*, vol. 140, pp. 65–99, Aug. 2021. [Online]. Available: https://doi.org/10.1016/j.neunet.2021.03.004

[6] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?" *IEEE Trans. Biometr., Behav., Ident. Sci.*, vol. 3, no. 1, pp. 101–111, Jan. 2021. [Online]. Available: https://doi.org/10.1109/TBIOM.2020.3027269

[7] K. S. Krishnapriya, K. Vangara, M. C. King, V. Albiero, and K. Bowyer, "Characterizing the variability in face recognition accuracy relative to race," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 2278–2285. [Online]. Available:https://doi.org/10.1109/CVPRW.2019.00281

[8] T. De Freitas Pereira and S. Marcel, "Fairness in biometrics: A figure of merit to assess biometric verification systems," *IEEE Trans. Biometr., Behav., Ident. Sci.*, vol. 4, no. 1, pp. 19–29, Jan. 2022. [Online]. Available: https://doi.org/10.1109/TBIOM.2021.3102862

[9] W. Hutiri and A. Y. Ding, "Bias in automated speaker recognition," in *Proc. ACM Conf. Fairness, Accountabil., Transp. (FAccT)*, 2022, pp. 230–247. [Online]. Available: https://doi.org/10.1145/3531146.3533089

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2021. [Online]. Available: https://doi.org/10.1145/3457607

[11] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surv.*, vol. 55, no. 3, pp. 1–44, 2023. [Online]. Available: https://doi.org/10.1145/3494672

[12] B. Friedman and H. Nissenbaum, "Bias in computer systems," *Comput. Ethic.*, vol. 14, no. 3, pp. 215–232, 1996. [Online]. Available: https://doi.org/10.4324/9781315259697-23

[13] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data Soc.*, vol. 3, no. 2, pp. 1–21, 2016. [Online]. Available: https://doi.org/10.1177/2053951716679679

[14] S. Wachter, B. Mittelstadt, and C. Russell, "Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law," *West Virginia Law Rev., Forthcom.*, vol. 123, no. 3, pp. 1–51, 2021. [Online]. Available: https://doi.org/10.2139/ssrn.3792772

[15] (Nat. Inst. Stand. Technol., Gaithersburg, MD, USA). *NIST Evaluates Face Recognition Software'S Accuracy for Flight Boarding*. (2021). [Online]. Available: https://www.nist.gov/news-events/news/2021/07/nist-evaluates-face-recognition-softwares-accuracy-flight-boarding

[16] K. Kotwal and S. Marcel, "Fairness index measures to evaluate bias in biometric recognition," in *Proc. ICPR*, 2022, pp. 1–14. [Online]. Available: https://doi.org/10.1007/978-3-031-37660-3_34

[17] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Proc. Equity Access Algorithm., Mech., Optim.*, 2021, pp. 1–9. [Online]. Available: https://doi.org/10.1145/3465416.3483305

[18] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11, Nov. 2021, Art. no. 100336. [Online]. Available: https://doi.org/10.1016/j.patter.2021.100336

[19] I. D. Raji and G. Fried, "About face: A survey of facial recognition evaluation," in *Proc. AAAI*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2102.00813

[20] Y. Hirota, Y. Nakashima, and N. Garcia, "Gender and racial bias in visual question answering datasets," in *Proc. ACM Fairness Accountabil. Transp. (FAccT)*, 2022, pp. 1280–1292. [Online]. Available: https://doi.org/10.1145/3531146.3533184

[21] J. Pahl, I. Rieger, A. Möller, T. Wittenberg, and U. Schmid, "Female, white, 27? Bias evaluation on data and algorithms for affect recognition in faces," in *Proc. ACM Fairness Accountabil. Transp. (FAccT)*, 2022, p. 15. [Online]. Available: https://doi.org/10.1145/3531146.3533159

[22] N. Sambasivan, S. Kapania, and H. Highfll, "'Everyone wants to do the model work, not the data work': Data cascades in high-stakes AI," in *Proc. Conf. Hum. Fact. Comput. Syst.*, 2021, pp. 1–15. [Online]. Available: https://doi.org/10.1145/3411764.3445518

[23] T. I. Liao, R. Taori, I. D. Raji, and L. Schmidt, "Are we learning yet? A meta-review of evaluation failures across machine learning," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 1–19. [Online]. Available: https://openreview.net/forum?id=mPducS1MsEK

[24] C. Rusti, A. Leschanowsky, C. Quinlan, M. Pnacek, L. Gorce, and W. Hutiri, "Benchmark dataset dynamics, bias and privacy challenges in voice biometrics research," in *Proc. Int. Joint Conf. Biometr. (IJCB)*, 2023, pp. 1–10. [Online]. Available: https://doi.org/10.1109/IJCB57857.2023.10449225

[25] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015. [Online]. Available: https://doi.org/10.1109/MSP.2015.2462851

[26] S. Pruzansky and M. V. Mathews, "Talker-recognition procedure based on analysis of variance," *J. Acoust. Soc. Am.*, vol. 36, no. 11, pp. 2041–2047, 1964. [Online]. Available: https://doi.org/10.1121/1.1919320

[27] S. Furui, "50 years of progress in speech and speaker recognition research," *ECTI Trans. Comput. Inf. Technol.*, vol. 1, no. 2, pp. 64–74, 2016. [Online]. Available: https://doi.org/10.37936/ecti-cit.200512.51834

[28] J. M. Naik, L. P. Netsch, and G. R. Doddington, "Speaker verification over long distance telephone lines," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1989, pp. 524–527. [Online]. Available: https://doi.org/10.1109/ICASSP.1989.266479

[29] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, nos. 1–2, pp. 91–108, Aug. 1995. [Online]. Available: https://doi.org/10.1016/0167-vol. 6393,no. 9500009-D

[30] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006. [Online]. Available:https://doi.org/10.1109/LSP.2006.870086

[31] J. Markel, B. Oshika, and A. Gray, "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 4, pp. 330–337, 1977. [Online]. Available: https://doi.org/10.1109/TASSP.1977.1162961

[32] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011. [Online]. Available: https://doi.org/10.1109/TASL.2010.2064307

[33] J. S. Chung, J. Huh, and S. Mun, "Delving into VoxCeleb: Environment invariant speaker recognition," 2019, *arXiv:1910.11238*.

[34] E. Khoury et al., "The 2013 speaker recognition evaluation in mobile environment," in *Proc. Int. Conf. Biometr. (ICB)*, 2013, pp. 1–8. [Online]. Available: https://doi.org/10.1109/ICB.2013.6613025

[35] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2014, pp. 4052–4056. [Online]. Available: https://doi.org/10.1109/ICASSP.2014.6854363

[36] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. INTERSPEECH*, 2017, pp. 999–1003. [Online]. Available: https://doi.org/10.1109/SLT.2016.7846260

[37] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 5329–5333. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461375

[38] D. Sztahó, G. Szaszák, and A. Beke, "Deep learning methods in speaker recognition: A review," *Periodica Polytechnica Electr. Eng. Comput. Sci.*, vol. 65, no. 4, pp. 310–328, 2021. [Online]. Available: https://doi.org/10.3311/PPee.17024

[39] J. Williams, K. Pizzi, N. Tomashenko, and S. Das, "Anonymizing speaker voices: Easy to imitate, difficult to recognize?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 12491–12495. [Online]. Available: https://ieeexplore.ieee.org/document/10445935

[40] W. Hutiri, L. Gorce, and A. Y. Ding, "Design guidelines for inclusive speaker verification evaluation datasets," in *Proc. INTERSPEECH*, 2022, pp. 1293–1297. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-10799

[41] J. P. Campbell and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1999, pp. 829–832. [Online]. Available: https://doi.org/10.1109/icassp.1999.759799

[42] D. E. Sturim, P. A. Torres-Carrasquillo, and J. P. Campbell, "Corpora for the evaluation of robust speaker recognition systems," in *Proc. INTERSPEECH*, 2016, pp. 2776–2780. [Online]. Available: https://doi.org/10.21437/interspeech.2016-1609

[43] O. Papakyriakopoulos et al., "Augmented datasheets for speech datasets and ethical decision-making," in *Proc. ACM Conf. Fairness, Accountabil., Transp. (FAccT)*, 2023, pp. 1–24. [Online]. Available: https://doi.org/10.1145/3593013.3594049

[44] A. Aksënova, D. van Esch, J. Flynn, and P. Golik, "How might we create better benchmarks for speech recognition?" in *Proc. 1st Workshop Benchmark., Past, Present Future*, 2021, pp. 22–34. [Online]. Available: https://doi.org/10.18653/v1/2021.bppf-1.4

[45] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: The mixer 3, 4 and 5 corpora," in *Proc. INTERSPEECH*, 2007, pp. 2864–2867.

[46] D. Graff, K. Walker, and D. Miller. "Switchboard cellular part 2 audio." 2004. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2004S07

[47] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. INTERSPEECH*, 2017, pp. 2616–2620. [Online]. Available: https://doi.org/10.21437/Interspeech.2017-950

[48] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," in *Proc. INTERSPEECH*, 2015, pp. 130–153. [Online]. Available: http://dx.doi.org/10.7488/ds/298

[49] K. A. Lee et al., "The RedDots data collection for speaker recognition," in *Proc. INTERSPEECH*, 2015, pp. 1–6. [Online]. Available: https://doi.org/10.21437/Interspeech.2015-95

[50] Z. Wu et al., "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2015, pp. 4440–4444. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178810

[51] J. Yamagishi, C. Veaux, and K. MacDonald, *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92)*, University of Edinburgh, Edinburgh, U.K., Nov. 2019. [Online]. Available: https://doi.org/10.7488/ds/2645

[52] H. Delgado et al. "ASVspoof 5 evaluation plan." 2023. [Online]. Available: http://www.asvspoof.org/

[53] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2016, pp. 5040–5044. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7472637

[54] Z. Ma, Y. Qiu, F. Hou, R. Wang, J. T. W. Chu, and C. Bullen, "Determining the best acoustic features for smoker identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 8177–8181. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9747712

[55] Z. Ma et al., "Automatic speech-based smoking status identification," in *Proc. Comput. Conf.*, 2022, pp. 193–203. [Online]. Available: https://doi.org/10.1007/978-3-031-10467-1_11

[56] D. Gillick, "Can conversational word usage be used to predict speaker demographics?" in *Proc. INTERSPEECH*, 2010, pp. 1381–1384. [Online]. Available: https://doi.org/10.21437/interspeech.2010-421

[57] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015. [Online]. Available: https://ora.ox.ac.uk/objects/uuid:a5f2e93f-2768-45bb-8508-74747f85cad1

[58] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2018, p. 826. [Online]. Available: https://doi.org/10.1109/FG.2018.00020

[59] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Comput. Speech Lang.*, vol. 60, Mar. 2020, Art. no. 101027. [Online]. Available: https://doi.org/10.1016/j.csl.2019.101027

[60] J. Yamagishi et al., "ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," in *Proc. INTERSPEECH*, 2019, pp. 1–19. [Online]. Available: https://api.semanticscholar.org/CorpusID:2075599

[61] T. Kinnunen et al., "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2017, pp. 5395–5399. [Online]. Available: https://doi.org/10.1109/ICASSP.2017.7953187

[62] H. Delgado et al., "ASVspoof 2017 Version 2.0: Meta-data analysis and baseline enhancements," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, 2018, pp. 296–303. [Online]. Available: https://doi.org/10.21437/Odyssey.2018-42

[63] H. Delgado et al., "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," in *Proc. INTERSPEECH*, 2021, pp. 1–13. [Online]. Available: http://www.asvspoof.org/

[64] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. INTERSPEECH*, 2020, pp. 2757–2761. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-2826

[65] R. Ardila et al., "Common Voice: A massively-multilingual speech corpus," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[66] C. Cieri, J. P. Campbell, H. Nakasone, D. Miller, and K. Walker, "The mixer corpus of multilingual, multichannel speaker recognition data," in *Proc. 4th Int. Conf. Lang. Resour. Eval.*, 2004, pp. 627–630. [Online]. Available: https://aclanthology.org/L04-1502/

[67] C. Cieri et al., "The mixer and transcript reading corpora: Resources for multilingual, crosschannel speaker recognition research," in *Proc. 5th Int. Conf. Lang. Resour. Eval.*, 2006, pp. 117–120. [Online]. Available: https://aclanthology.org/L06-1318/

[68] L. Brandschain, C. Cieri, D. Graff, A. Neely, and K. Walker, "Speaker recognition: Building the mixer 4 and 5 corpora," in *Proc. 6th Int. Conf. Lang. Resour. Eval.*, 2008, pp. 1–4. [Online]. Available: https://aclanthology.org/L08-1104/

[69] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "The mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," 2010, pp. 1–4. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/792_Paper.pdf

[70] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Todisco, and H. Delgado, "ASVspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," in *Proc. INTERSPEECH*, 2018, pp. 1–6. [Online]. Available: http://www.asvspoof.org/index2017.html

[71] R. Singh, *Profiling Humans From Their Voice*. Singapore: Springer, 2019. [Online]. Available: https://doi.org/10.1007/978-981-13-8403-5

[72] J. L. Kröger, O. H.-M. Lutz, and P. Raschke, "Privacy implications of voice and speech analysis—Information disclosure by inference," *Privacy and Identity Management. Data for Better Living: AI and Privacy*. Cham, Switzerland: Springer, 2020, pp. 242–258. [Online]. Available: https://doi.org/10.1007/978-3-030-42504-3_16

[73] "Regulation (EU) 2016/679 of the European parliament and of the council." Protection Regulation. 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[74] C. Rodrigues and D. Simões, "How can sociolinguistic data be used?" *Revista Diacrítica*, vol. 27, pp. 287–308, Dec. 2012. [Online]. Available: https://scielo.pt/pdf/dia/v27n1/v27n1a12.pdf

[75] *Information Technology—Security Techniques—Vulnerability Disclosure*, ISO/IEC 29147:2018, 2018. [Online]. Available: https://www.iso.org/standard/72311.html

[76] D.-O. Jaquet-Chiffelle and M. Loi, "Ethical and unethical hacking," *The Ethics of Cybersecurity*. Cham, Switzerland: Springer, 2020, pp. 179–204. [Online]. Available: https://doi.org/10.1007/978-3-030-29053-5_9

[77] G. A. Thomas, "Issues of professionalism concerning the ethical hacking of law firms," M.S. thesis, School Comput. Math., Charles Sturt Univ., Bathurst, NSW, Australia, 2020. [Online]. Available: https://researchoutput.csu.edu.au/en/publications/issues-of-professionalism-concerning-the-ethical-hacking-of-law-f

[78] N. Tomashenko et al., "The VoicePrivacy 2020 challenge: Results and findings," *Comput. Speech Lang.*, vol. 74, Jul. 2022, Art. no. 101362. [Online]. Available: https://doi.org/10.1016/j.csl.2022.101362

[79] A. Leschanowsky, U. E. Gaznepoglu, and N. Peters, "Voice anonymization for all-bias evaluation of the voice privacy challenge baseline systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 4785–4789. [Online]. Available: https://doi.org/10.1109/icassp48485.2024.10447137

[80] J. Yamagishi et al., "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proc. Workshop-Autom. Speak. Verif. Spoof. Coutermeas. Chall.*, 2021, pp. 1–8. [Online]. Available: https://doi.org/10.48550/arxiv.2109.00535

[81] N. Tomashenko et al., "The VoicePrivacy 2024 challenge evaluation plan," 2024, *arXiv:2404.02677*. [Online]. Available: https://doi.org/10.48550/arXiv.2404.02677

[82] W. Hutiri, O. Papakyriakopoulos, and A. Xiang, "Not my voice! A taxonomy of ethical and safety harms of speech generators," 2024, *arXiv:2402.01708*. [Online]. Available: https://arxiv.org/abs/2402.01708

[83] N. Evershed and J. Taylor. "AI can fool voice recognition used to verify identity by centrelink and Australian tax office." 2023. [Online]. Available: https://www.theguardian.com/technology/2023/mar/16/voice-system-used-to-verify-identity-by-centrelink-can-be-fooled-by-ai

[84] J. Cox, "How I broke into a bank account with an AI-generated voice." 2023. [Online]. Available: https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice

[85] L. Varanasi. "Biden's AI chief says 'voice cloning' is what keeps him up at night." 2023. [Online]. Available: https://www.businessinsider.com/voice-cloning-technology-worries-biden-ai-bruce-reed-elevenlabs-scammers-2023-11

[86] T. F. Zheng and L. Li, *Robustness-Related Issues in Speaker Recognition*, vol. 2. Singapore: Springer, 2017. [Online]. Available: https://doi.org/10.1007/978-981-10-3238-7

[87] J. Le Roux and E. Vincent, "A categorization of robust speech processing datasets," Mitsubishi Electr. Res., Inc., Cambridge, MA, USA, Rep. TR2014-116, 2014. [Online]. Available: https://inria.hal.science/hal-01063805/

[88] S. Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA, USA: MIT Press, Mar. 2020. [Online]. Available: https://doi.org/10.7551/mitpress/12255.001.0001

[89] T. Gebru et al., "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, 2021. [Online]. Available: https://doi.org/10.1145/3458723

[90] J. King, D. Ho, A. Gupta, V. Wu, and H. Webley-Brown, "The privacy-bias tradeoff: Data minimization and racial disparity assessments in U.S. government," in *Proc. ACM Conf. Fairness, Account., Transpar.*, 2023, pp. 492–505. [Online]. Available: https://doi.org/10.1145/3593013.3594015

[91] S. Mittal et al., "On responsible machine learning datasets with fairness, privacy, and regulatory norms," 2023, *arXiv:2310.15848*. [Online]. Available: https://doi.org/10.48550/arxiv.2310.15848

[92] M. Le et al., "Voicebox: Text-guided multilingual universal speech generation at scale," 2023, *arXiv:2306.15687*. [Online]. Available: https://doi.org/10.48550/arxiv.2306.15687