Privacy in Speech Technology

Tom Bäckström, *Senior Member, IEEE*

***Abstract*—Speech technology for accessing information and services has rapidly improved in quality. It is convenient and appealing because speech is the primary mode of communication for humans. Such technology however also presents proven threats to privacy. Speech is a tool for communication and it will thus inherently contain private information. Importantly, it however also contains a wealth of side information, such as information related to health, emotions, affiliations, and relationships, all of which are private. Exposing such private information can lead to serious threats such as price gouging, harassment, extortion, and stalking. This paper is a tutorial on privacy issues related to speech technology, modeling their threats, approaches for protecting users' privacy, measuring the performance of privacy-protecting methods, perception of privacy as well as societal and legal consequences. In addition to a tutorial overview, it also presents lines for further development where improvements are most urgently needed.**

***Index Terms*—speech technology, privacy and security, machine learning**

TABLE I
A SELECTION OF CATEGORIES OF PRIVATE INFORMATION POTENTIALLY IDENTIFIABLE FROM SPEECH SIGNALS, AND THE EXTENT TO WHICH THEY ARE SUSTAINED OVER TIME AND CAN BE WILLFULLY CONTROLLED.

Category	**Examples**	**Permanence**	**Control**
Biological	Body characteristics	Sustained	No
	State of health	Variable	No
Psychological	Emotions	Variable	Partly
	Intelligence	Sustained	No
	Education, skill	Sustained	Yes
	Gender identity	Sustained	No
Message	Text, emphasis, style, expression	Variable	Yes
	Mannerisms, context	Variable	Partly
	Language choice & skills	Partly	Partly
Affiliation	Ethnic, national, cultural, religious, political, etc.	Sustained	No
Relationship character	Hierarchy, familiarity, attraction, intimacy	Sustained	Partly
	Has met with person	Variable	Yes
Physical environment	Background sounds, distance to sensor, transmission distance, reverberation, location	Variable	Yes
Hardware used	Sensor type & manufacturer	Variable	Yes

TABLE II
PRACTICAL EXAMPLES OF THREATS RELATED TO PRIVATE AND SENSITIVE INFORMATION THAT IS CONVEYED BY PEOPLE'S VOICES.

Exploit	**Example**
Price gouging	Signs of depression or other health problems in users' voices could be misused to trigger an increase in their insurance premiums. Signs of users' emotions could be exploited to offer them products at higher prices.
Tracking, stalking	Voice re-identification could link users across platforms, i.e., from work-related social media to online support groups and dating apps, making it possible to follow them anywhere.
Extortion, public humiliation	Private health problems and romantic affairs could be detected in the voice and used for blackmail or made public against a user's wishes.
Algorithmic stereotyping	Recommender systems based on voice can become biased with respect to age, identity, religion or ethnicity, in ways that are nearly impossible to monitor.
Harassment, inappropriate advances	Users in chat rooms or virtual reality could be automatically singled out by gender or opinions, making them a target for unwanted attention and harassment.

I. Introduction

SPEECH is a mode of communication and thus inherently contains a wealth of information (see table I). Communicating information already known by the receiver is pointless, and efficient communication will thus mainly contain information that is *not* widely known or which is private. In addition, speech contains also a wide range of side information like the state of health and emotions, as well as physical, psychological, and social identity, most of which is private information. Finally, speaking is a dynamic interaction between two or more speakers. Dialogues thus also contain information about the relationship between participants such as their level of familiarity, affiliation, intimacy, relative hierarchy, shared interests, and history. From an information content perspective, we can thus expect that *privacy is a multifaceted issue in all areas of speech technology*.

Studying and improving privacy in speech technology is important because breaches in privacy can have serious consequences. Table II lists examples of threats and exploits. The list is incomplete and we can expect new threats to be discovered. The generic solution template is however typically always the same: *Minimize transmission and storage of as well as access to sensitive information which is irrelevant to the service that the users want.* It is then "merely" a question of *how* such minimization is achieved, *what* information is relevant, and *how* to determine what the user wants (provide information and enable control of services and threats).

With respect to threats, media attention is often focused on *breaches which have large economic consequences* [1–3]. While such breaches are important by themselves, the media attention introduces unfortunate biases in two ways. First, a breach is a worst-case event whereas threats, which have not yet led to a breach event, can already have a large impact. For example, users may choose not to use systems that threaten their privacy; known weaknesses and even a lax attitude of the service provider toward privacy can therefore have an effect on the adoption, retention, and sales of products and services. Importantly, users can avoid systems that are

T. Bäckström is with the Department Information and Communications Engineering, Aalto University, Finland e-mail:tom.backstrom@aalto.fi.

Manuscript received Month Day, Year; revised Month Day, Year.

perceived threatening, even when there is no actual threat. To maintain users' trust, it is therefore important to both uphold the users' actual privacy, but also design systems to actively and frequently communicate the level of privacy, including known threats, their current status, and measures taken to protect against them.

Second, while a single breach in a large service can have large economic, psychological, societal, and legal consequences, small crimes are so common that their joint effect is comparable or larger in size [4]. With speech technology, such ”small” crimes include stalking, extortion, harassment, humiliation, and inappropriate advances (see table II). While the economic damage of a single such incident can be small, when combined, their joint psychological and societal effect is potentially large and their prevalence makes them a considerable threat also economically.

There are two primary uses of speech technology, telecommunication, and human-computer interfaces. With respect to telecommunication, approaches to ethical, legal, and technological questions related to privacy in telecommunication over landlines are well-established, and open discussions are primarily related to how the existing regulation and oversight should be extended to cover also mobile telecommunication and voice-over-Internet protocols (VoIP) [5–7]. The scope of this paper can thus be limited to human-computer interfaces where a computer processes speech signals (see fig. 3). Another limitation is that focus is here limited to the acoustic speech signal also known as the *voice* since natural language processing is a distinct and largely independent field that warrants its own treatment (e.g. [8]).

Privacy is closely related to challenges in security and it is often difficult to distinguish between them. Here we strike the balance by considering *privacy scenarios where an agent has legitimate access to some private speech data of the user, but uses it for purposes contradicting with, or gains access beyond, the users' expectations or preferences*. For example, a voice interface can be used to control home automation, but if the service provider shares that data also to advertisers against the users' preferences, then it is a violation of privacy (see fig. 4). Consequently, security concerns such as identity spoofing, deep fakes as well as attacks on devices or networks are also excluded here, as these threats are more related to security rather than privacy.

This paper presents a tutorial overview of privacy in speech technology that covers a wide range of threats, methodologies, and algorithms. Analysis of threats however demonstrates that while attack surfaces have great variety, the threat models are similar (see section II). Protection against those threats have four distinct categories of approaches, removing side information, improving overall system performance, limiting access to private message and limiting access to reproduced audio (see section III). To quantify of extent of protections we further need methods for objective evaluation (see section IV). While objective evaluation quantifies the actual level of privacy, users' impressions of privacy do not necessarily follow the objective level. For the best user experience, therefore, we need to quantify and understand users' perceptions, experiences, and preferences related to privacy, as well as design systems accordingly (see section V). The range of applications and content types where privacy-preserving processing methods are needed is vast and section VI presents a brief overview thereof. Finally, threats to and breaches in privacy have considerable consequences on both societal, economic, and individual levels, which has motivated governments to increasingly regulate the use of technology (see section VII).

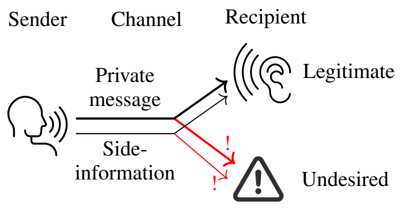

Fig. 1. The high-level *threat model* of speech interaction, where a private message is sent through a channel to the legitimate recipient, but consequential side-information is bundled to that message. It is a threat to privacy when an undesired recipient gains access to that private message or side information (marked by red arrows and exclamation marks).

To the author's knowledge, this is the first wider tutorial overview of privacy in speech technology. Recent works in the field however include an technical overview [9], as well as a popular-science review [10] of this area, and several doctoral theses have their respective summaries, e.g. [11–14]. Privacy has been extensively discussed in other areas of science, such as the neighboring field of natural language processing [8], in statistical theory of privacy [15], in social sciences [16] and psychology [17]. There is even an excellent and thorough meta-study of all areas of science which discuss privacy [18].

II. Threats

A. Exposure

Speech information can be *exposed* in two forms (see fig. 1): First, when a *private message is transmitted*, it is a threat to privacy when contrary to the preferences of the user, it is used in an unexpected way or transmitted to a third party. Second, since speech contains a wide range of private information (see table I) which is bundled in the acoustic signal in complicated ways, it is challenging to extract only the desired message. Typically speech information thus always contains consequential side-information, bundled into the private message. It is a threat to privacy when such *private side-information is transmitted* to an undesired recipient alongside the private message. Per definition, we here assume that the legitimate recipient can be trusted with the side information and that any undesirable use of it is labeled as an undesirable service.

The difference between the two forms of information makes an important difference in approaches to mitigation. When a private message is exposed, then our only available solutions are to either reduce the accuracy or to limit access to the private message for example through cryptography (see sections III-C2, III-D and IV-B). However when side-information is exposed, as additional methods we can also use signal processing to better remove, replace or distort such side-information (see section III-A).

Observe that the threat exposure does not depend on the *route of the information*, as long as the content arriving to the undesired recipient is the same. The attack surface or channel which leaks information to the undesired recipient however does have a large impact on the choice of mitigation (see section II-E). The *type of undesired recipient*, be it a service, device, or person, has an effect mainly on its potential *ability to extract and use* private information in a nefarious way. The main difference in the *type of sender*, person, device, or service, is their different *ability to remove* side-information from the speech signal prior to transmission.

B. Inference of Attributes and Identity

Threats to the speakers' privacy can be categorized into two varieties; *property/attribute inference* and *re-identification*. The difference is that given a single known speaker, through property inference we can associate new private information or attributes to that speaker. In comparison, by analyzing (anonymized) speech data of an unknown speaker, we can assign the identity to the most likely speaker within a database of a multitude of speakers. The only difference is thus that in property inference we assign attributes to a single speaker, whereas in re-identification we look at many speakers and assign attributes to one (or a subset) of them. However, if we treat the physical identity (like the name on the passport) as a property or attribute of the speaker, then re-identification means that we have been able to infer a property of the speaker. Re-identification is, in this sense, one particular type of property inference.

C. Attacker Scenarios

Attacks can further be classified according to the amount of information available for the attacker, such as information about the speakers and about the trained models. Access to the training data as well as speech samples from the specific user (known as *enrollment data*) are particularly useful for the attacker. For example, in the anonymization task of the VoicePrivacy 2022 challenge, the objective was to replace (pseudonymize) speaker identity but retain all other speech characteristics such as linguistic information. The anonymized sentences are known as *trial* utterances. The attack scenarios were then classified as [19]:

1) *Unprotected*: no anonymization is performed; attackers have access to original trial and enrollment data.
2) *Ignorant attacker*: Trial data is anonymized, but attackers are unaware of it, hence they use original data for enrollment.
3) *Lazy-informed*: Both trial and enrollment data are available to attackers, but anonymized with different pseudo-speaker.
4) *Semi-informed*: As in lazy-informed, attackers can also train their model with the same anonymization system but different pseudo-speakers.

With increasing information, obviously, the attacker's task becomes easier and the accuracy of the attacker's models improves. Further, such attack scenarios can then be devised according to the specific use case. For example, if the task is to anonymize emotional or health status, we can define different attack scenarios depending on the extent to which the attacker has information about categories of emotions and health status used in anonymization.

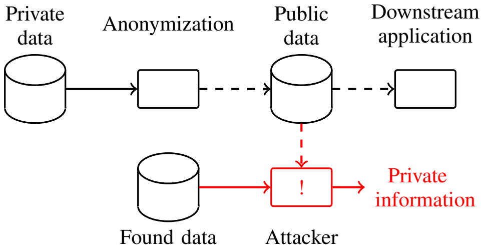

Fig. 2. An *attack model* for the evaluation of privacy-preserving anonymization, where private data is anonymized to remove private information, and the anonymized data is shared publicly. An attacker uses any available (found) data and anonymized public data to infer private information contrary to the users' preferences. Anonymized data flow is indicated by dashed lines and the attack by red lines and an exclamation mark.

D. Attack Model

Figure 2 illustrates an attack model for measuring the extent of privacy [20]. We assume that there is some private data, which is anonymized such that it can be used in a trusted task. The attacker has access to the anonymized data, which is here called ”public data”. Observe that the public data is not necessarily openly available for anyone to see, but it only indicates that the data is sufficiently freely available that the attacker has access to it. The attacker has also access to some other data about speakers, found from some other source (found data), which helps in extracting private information. Again, the term ”found data” is a loose term and denotes any data that is available to the attacker.

E. Attack Surfaces

We categorize scenarios according to the attack surface from where information is extracted (see fig. 3); a channel between cloud services (section II-E1), a user interface to the edge device (section II-E2) or to the cloud service (section II-E3), from the local network (section II-E4), the acoustic pathway (section II-E5), or through the shared user interface of the local device (section II-E6). We consider only cases where a piece of technology is receiving or transmitting information and exclude human-to-human communication without devices. We also assume that devices and network connections are secured such that only authorized services can communicate with them. The threats we focus on are thus illustrated with solid lines in fig. 3.

1) Cloud Leak: Suppose a user accesses a remote cloud service through an edge device (see fig. 4). The cloud service thus has legitimate access to the private message of the user. The cloud server however can then use that private message or bundled side information for some other purpose than that requested by the user, extract more information than anticipated, combine it with other information, or share

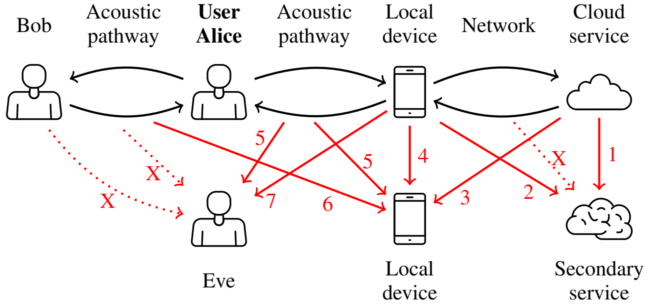

Fig. 3. Threats to the privacy of a user Alice when speaking with another person Bob or a local device, but where information is leaked or shared through the acoustic pathways, the edge device, the network or the cloud service, to another person Eve, device or service, contrary to the preferences of Alice (red arrows). Each threat is marked with the number of the corresponding part of section II-E. Dotted lines marked with ”*X*” indicate threats outside the current scope and are not considered here.

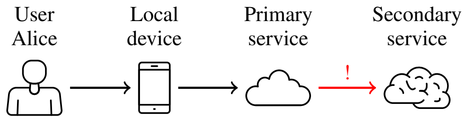

Fig. 4. Threat scenario ”*Cloud Leak*”, where a user Alice accesses a (primary) cloud service using an edge device, but the information is shared to a secondary service contrary to preferences (red arrow and exclamation mark).

information with a third party. For example, the cloud server of a voice assistant could inappropriately share information with an advertiser.

2) False Activation: Devices with speech interfaces can hear all conversations in the same acoustic space as the device and therefore need mechanisms to determine which utterances are intended for the speech interface. A popular approach is to use a specific utterance, known as a wake word, to start all interactions with the interface [21]. The wake word is then like a rudimentary password, which prevents the interface from activating when speech is not directed to the device. Unfortunately, designing wake word detectors is non-trivial, and they will occasionally make mistakes. They might sometimes miss a wake word when it is spoken (false negative) or mistake some other unrelated sound as a wake word (false positive). While false negatives are annoying for the user when the service does not activate, false positives potentially present serious threats to privacy. In some famous cases, speech interfaces have activated from sounds on the television to buy unwanted items at the users' behest, and users' private conversations have been leaked to third parties [22, 23].

3) Cloud Access: Figure 6 illustrates the threat where a user Alice accesses a cloud service through an edge device, and where some private information of Alice is stored. A second user Eve can then access the same cloud service through another device and potentially gain access to the stored private information, contrary to Alice's preferences. This attack surface is similar to the Cloud Leak scenario, with the main difference being the recipient, which is here a person, whereas, in the Cloud Leak scenario, it is an automated agent.

An example of this threat is medical records, which can be collected from patients to a centralized database. A researcher with authorization to access the database can then potentially extract private information beyond the expected. [24]

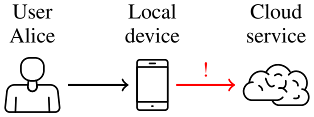

Fig. 5. Threat scenario ”*False Activation*”, where a user Alice accesses an edge device, but the information is shared to a cloud service contrary to preferences (red arrow and exclamation mark).

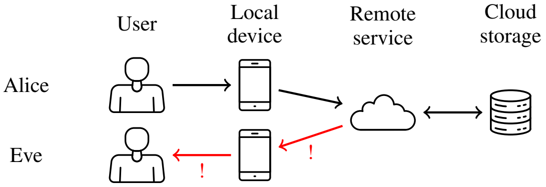

Fig. 6. Threat scenario ”*Cloud Access*”, where user Alice uses a cloud service through an edge device, where it is potentially stored and shared with user Eve, contrary to Alice's preferences (red arrow and exclamation mark).

4) WASN Authentication: The audio quality of speech pickup as well as the usability of voice interfaces can be improved by using all available connected devices with microphones that reside in the same room or acoustic space [25–27]. Such collaboration can be realized with acoustic sensor networks, where several independent devices simultaneously pick up speech and where several channels are combined to obtain a high-quality signal (see fig. 7) [11, 13]. A sensitive question is however authorization; Which devices are allowed to share information? One approach is to assume that those devices which are in the same room or acoustic space can hear the same signal [13, 28]. Presence in the same space is thus already an implicit authorization to participate in a joint signal pickup. To protect privacy, it is then necessary to determine which devices reside in the same acoustic space. Devices in a different room can belong to the same company or family, and they can be connected to the same network, but still, they are outside the sphere of the current discussion. Without proper authorization mechanisms, devices outside the room could then gain access to private speech inside the room.

5) Speech Interface and Discussion Leaks: Figure 8 illustrates a scenario where a device or person overhears a discussion between a user Alice and a local device. Similarly, fig. 9 illustrates a discussion between two persons, overheard by a local device. The defining property of both scenarios is the leak in the acoustic pathway, which necessitates that the eavesdropper is physically present in the same acoustic space as the private communication. There are four distinct cases to consider depending on whose speech is heard, Alice (or Bob) or the local device, and who is the eavesdropper, Eve, or the other local device.

In the case where Eve overhears Alice's speech, we assume that people have a learned awareness of which other people are present in the same room and adjust their speech accordingly; Then either Alice does not mind that Eve overhears her speech

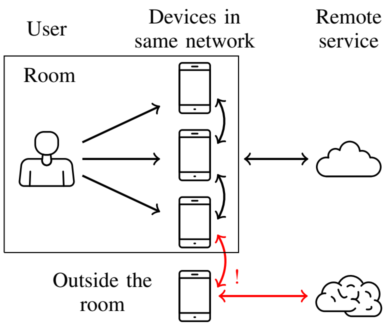

Fig. 7. Threat scenario "*WASN Authentication*", where a user uses a service through a distributed sensor network inside a room, but another device, outside the room in the same network, joins the distributed sensor network and shares information contrary to preferences (red arrow and exclamation mark).

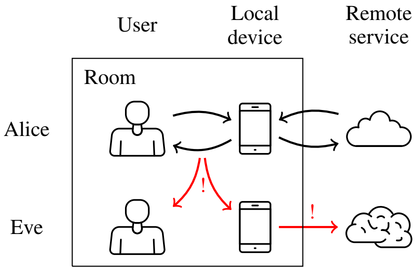

Fig. 8. Threat scenario "*Speech Interface Leak*", where a user Alice accesses an edge device (potentially including also a remote service), but another local person Eve, or device in the same room or acoustic space overhears the speech interaction contrary to preferences (red arrow and exclamation mark).

(inconsequential information), Alice can change her speech style to a whisper, or change content, such that Eve does not hear anything private (reduced information transmission) or she can go to a different room to continue the interaction in private (modified acoustic channel).

6) Shared Device: Many practical scenarios include one or several devices shared by multiple users. For example, a family can share a smart speaker or television, an office meeting room can have smart devices and customer service points can have a phone shared across duty officers. Each of these devices can collect private information about the user(s) over time and can potentially share it with other users (see fig. 10). Notably, this scenario highlights that leak happens over a distance in *time*, whereas threats occurring over a *spatial* distance often receive the majority of the attention.

This threat model clearly demonstrates that devices and services used by multiple users need to employ access control and authorization management if they store any private information. It is also obviously related to many security threats – unauthorized access to devices should be prevented – but those are outside the scope of this work.

7) Levels of Trust: In the Cloud Leak-scenario in fig. 4, observe that it represents a sequence of automated agents – edge, primary and secondary cloud services – where it would be desirable that with each step in the sequence, the amount of information shared is reduced (see fig. 11). Such minimization of the potential attack surface is sensible from a risk management perspective. However, at the same time, we can interpret this as *levels of trust*. The user can exert direct control over the local acoustic space and the edge device, and thus correspondingly we can expect that the user has the highest level of trust in the edge device. With each step further in the sequence, the level of control will diminish and similarly, the level of trust decreases.

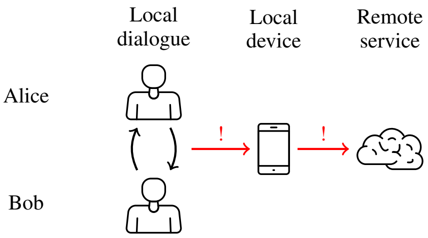

Fig. 9. Threat scenario "*Discussion Leak*", where a dialogue between two users, Alice and Bob, is picked up by a local device and potentially shared to a remote service contrary to preferences (red arrow and exclamation mark).

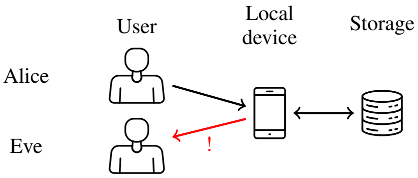

Fig. 10. Threat scenario "*Shared Device*", where user Alice uses an edge device, where information is stored, and shared with another user Eve, contrary to Alice's preferences (red arrow and exclamation mark).

Trust is clearly a concept that depends on at least the psychological, social, and cultural context, making it hard to define. Here we however define trust as follows. If user Alice has observed an agent for some time, they can form a prediction of how the agent will behave. If the probability distribution of the prediction is narrow, then Alice has high confidence in that prediction. We then say that Alice has trust in the agent. We thus define that *a prediction of high confidence is equal to trust*. If the prediction is also that the agent behaves in a beneficial way for Alice, then they *perceive* the agent *as trustworthy*. This is however dependent on Alice's ability to predict actions. If the agent actually behaves beneficially, then the agent is *objectively trustworthy*.

III. Protections

A. Reducing Side Information

The primary approach for reducing private side information in speech signals is signal processing. It has two opposing objectives corresponding respectively to *utility* and *privacy*; 1) the *trusted task* of processing or analysis, where some category of information is extracted for a legitimate purpose and 2) protection against the *threat task*, where a nefarious operator tries to extract private information beyond the legitimate objective. Quality of the trusted task typically follows classic speech processing methodology, while protection against the threat task can be achieved by removing, replacing, or distorting

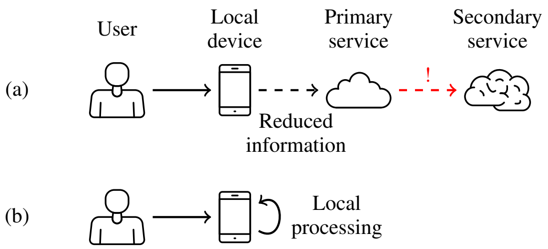

Fig. 11. Two alternative protections against the "*Cloud Leak*" scenario, where the transmission of private information to the cloud is either (a) reduced or (b) prevented.

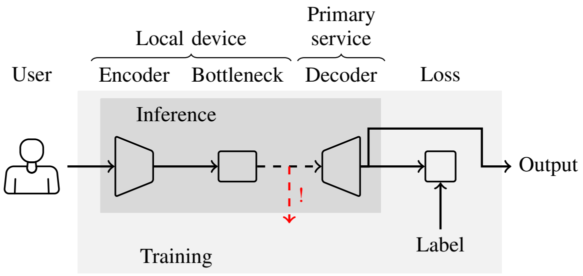

Fig. 12. Training of a privacy-preserving speech analysis method with the *information bottleneck* principle. The point of attack is indicated by the red arrow and exclamation mark.

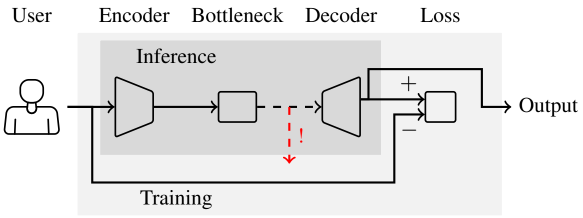

Fig. 13. Training of a privacy-preserving an *autoencoder* structure as an example of the information bottleneck principle. The point of attack is indicated by the red arrow and exclamation mark.

private information. The central question is how information for the trusted task is separated from other private information.

The most common approach for protecting against this scenario is to limit information flow from the local edge device to the cloud server (see fig. 11). Such limitation or reduction in information flow can be implemented by removing, replacing, or distorting private content as in fig. 11(a). At the extreme, the edge device can be entirely disconnected from the cloud or network, and process information only locally, as in fig. 11(b) [29]. In doing so, we then need to assume that the local device has sufficient capacity and software to complete the requested tasks, that it is not compromised, and that the restriction of information flow is sufficient to ensure that private information is not leaked.

The two main approaches for removing private information are based on an *information bottleneck* or an *adversarial model*. Information bottlenecks are based on squeezing the desired information through a metaphorical bottleneck, such that there is no room for side-information to pass through [30]. Often such models follow an autoencoder structure, which is trained to reconstruct the input signal from the transmitted signal (see section III-A1).

Adversarial models (which should not be confused with generative adversarial models) are in turn used in the training of machine learning models to make sure that no private information can be deduced from the transmitted information. It is based on modeling both the trusted and threat tasks in parallel, but the transmitted message is optimized so that the trusted task succeeds and the threat task fails (see section III-A2). A generalization of the information bottleneck idea is to *disentangle* speech into multiple independent data streams. This allows applications to cherry-pick the private information that should be transmitted case by case (see section III-A3).

1) Information Bottleneck: To distill only the private message and discard any side-information, with the information bottleneck approach, speech information is passed through a bottleneck so tight, that only the legitimate message can pass through [31, 32] (see fig. 12). This approach thus provides protection against an attacker who has access to the output of the bottleneck. The challenges are to design a bottleneck that is sufficiently tight such that private side information is discarded and a model structure and training methodology which optimizes the quality of the legitimate message.

Information content at the bottleneck can be reduced by either reducing the signal rank (reducing the number of units of data passed through the bottleneck, e.g. [33]) or by quantizing and coding the signal at a low bitrate (e.g. [9, 34]). The trade-off between the accuracy of reconstruction and bottleneck entropy then determines the extent of privacy (see section IV).

In parallel, when using a machine learning model, the model has to be trained for the best trade-off between utility and privacy. Typically this entails minimizing the loss in the accuracy of the private message. For example, if the task is to extract text content with an automatic speech recognizer, then the error rate of that recognizer should be minimized.

A frequently used approach to implementing a bottleneck is an autoencoder structure, consisting of an encoder, bottleneck, and a decoder, where the objective is to reconstruct the input signal from the bottleneck output [33, 35, 36] (see fig. 13). However, to be privacy-preserving, the bottleneck should be sufficiently tight that the original speech signal cannot be perfectly reconstructed to resemble the original in their waveform. This makes it challenging to design a loss function because the output does not sufficiently resemble the input. One solution akin to representation learning is to feed the output of the autoencoder again to the encoder and compare the bottleneck features [37, 38].

2) Adversarial Approach: An alternative to reducing the size of the information bottleneck is to use an adversarial model during training, such that side-information in the bottleneck is minimized [39, 40]. Figure 14 illustrates the model structure, where the trusted and threat tasks correspond respectively to utility and privacy, which in turn respectively correspond to the extraction of the private message and side-information. The threat task is independently optimized to extract private side information from the bottleneck output. While keeping the threat task fixed, the encoder and the trusted

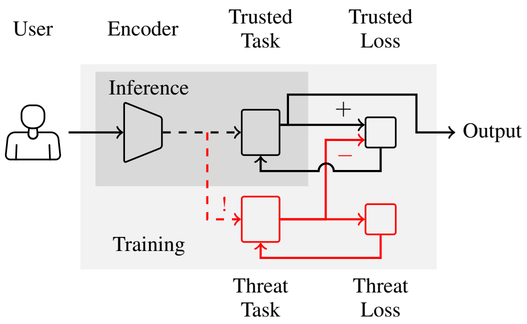

Fig. 14. Training of a privacy-preserving speech analysis method with an *adversarial* approach. The trusted task is authorized to extract some information, while the threat task (drawn in red) is extracting some other private information. During training, the trusted task is competing with the adversarial threat task, such that in the encoder block, all private information is removed. During inference, the threat task can be ignored.

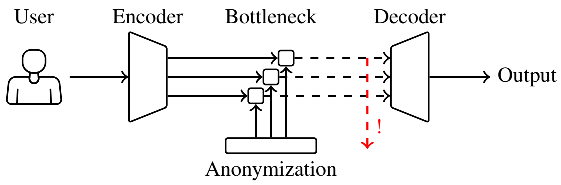

Fig. 15. Privacy-preserving speech processing through *disentanglement*, where speech is decomposed into independent streams of information, each representing a distinct category of information and where the level of anonymization can be individually chosen for each category. The point of attack is indicated by the red arrow and exclamation mark.

tasks are jointly optimized to maximize the accuracy of the private message *and* to minimize the accuracy of private side information.

Adversarial training thus has a mixture of maximizing and minimizing the accuracy of private side information. This forces the encoder to minimize private side information passed through the bottleneck, since the threat task tries to maximize the extraction of private side information.

An advantage of an adversarial configuration is that the designer can specifically choose which type of threat and information category is removed from the data. The system can also be optimized end-to-end, such that all components are jointly optimized for the best performance. This can lead to an efficient model in terms of the trade-off between computational cost and output quality.

The principal issue with adversarial training is that it does not give any theoretical guarantees or measures of privacy. The best it can do is to demonstrate the extent to which the chosen adversarial model was unable to extract private side information from the chosen category of private information. We can thus expect that some other attacker with a better model could still extract private side-information *and* that other categories of private information are potentially present in the bottleneck output.

3) Disentanglement: Information bottlenecks can be generalized to encompass multiple bottleneck channels, where each bottleneck represents a *disentangled* representation (see fig. 15). That is, each channel represents a distinct and independent attribute of the signal such that the extent of anonymization can be cherry-picked per channel as per use-case e.g. [41–43].

The main challenge with disentanglement is to define the constraints with which information is funneled to the corresponding channel. For example, the model can be trained to match specific channels with labels in the dataset (e.g. [42, 43]). Alternatively, using representation learning, we can constrain channels based on objective criteria such as time-scope, i.e. the length of time over which data is integrated (e.g. [37, 44, 45]).

B. Improving Performance

In many use cases, like False Activation, the actual culprit is the inadequate performance of the service. For example, if a wake word detector is incorrectly triggered (false positive), then the system will start to listen to a conversation when it was not supposed to, thus breaching privacy. Since the inadequate wake word detector is thus causing a privacy breach, then the best solution is to improve the wake word detector. The treatment thus addresses the cause rather than the symptoms thereof.

This approach has however two principal challenges. First, improving performance often requires an increase in computational power and other resource consumption. This is not only financially costly but has also an environmental penalty [46]. In particular, balancing the computational load is easier on a cloud server, and thus existing resources can be more efficiently used. Second, even with improved performance, speech interfaces will always have occasional errors. That is, the design of privacy-preserving speech technology must include multiple layers of protection, especially when it comes to operations with large consequences [47]. Say, when the wake word detector is activated, then before any actions which would potentially breach privacy, the system could require that the speaker is in the same room, or the speaker identity is verified, or an extra confirmation step "*Are you sure?*" or similar. The intrusiveness of such additional protections should then reflect the severity of the potential breach such that the protections are not perceived as overly obtrusive and decrease the utility of the service.

Improving performance can also mean that the system requires additional functionalities. For example, in the case a second local device overhears Alice's interaction with Bob (Discussion Leak) or the primary device (Speech Interface Leak), the required privacy protection is that the secondary device either 1) is aware that it is not the intended recipient of speech (speech analysis), or 2) notifies the user of its presence (user-interface design). This highlights the importance of trusting the devices present in the acoustic space. The secondary local device *can hear* everything spoken in the local space. The user thus has to have a high level of trust in the device and service provider that it is sufficiently *competent* to know when it is part of a conversation, that it is sufficiently *benevolent* to protect the users' privacy as well as notify the user of potentially privacy-infringing activity (cf. [48, 49]).

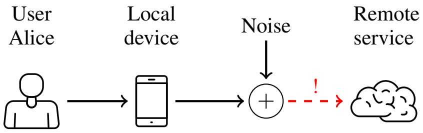

Fig. 16. Protecting privacy by adding noise to private data, following the idea of *differential privacy*.

C. Limiting Access to Private Messages

Every speech application involves the transmission of the necessary information, which is likely to be private. Since transmission of such private information is thus unavoidable, it exposes the user to threats to their privacy. Insofar we need the speech application, threats can be mitigated by 1) *reducing* the accuracy of information by adding noise, 2) making information inaccessible by e.g. *encryption*, or 3) choosing not to send anything but apply the processing only locally, known as *edge processing* (see section III-A).

1) Reducing Accuracy and Differential Privacy: Many applications require only population averages rather than information about specific individuals. For example, a call center could plausible need to know the distribution of genders, but not need to know the gender identity of any specific customer. This information can be extracted with a simple stochastic scheme known from the area of differential privacy [50]. Namely, the customer would first flip a coin to determine whether they should lie or tell the truth. When lying, the customer then flips a coin again to determine which gender to report (here we use only binary gender categories for brevity). This gives a 75 % likelihood that the true identity was reported and 25 % likelihood that it was false [51]. In other words, this approach corresponds to distorting the signal, or to adding noise to the signal to protect privacy (see fig. 16).

The customer can then always claim that they were lying, such that this system gives *plausible deniability*. Simultaneously, given a sufficiently large sample, we can then always estimate the true distribution from the noisy sample. Observe that while this approach gives plausible deniability, information is still statistically correlated with the true information. Given multiple noisy pieces of information, it may then still be possible to recover accurate, private information about the speaker. It is thus paramount to quantify the extent of protection with measures of differential privacy (see section IV-B).

The above example readily generalizes to any categorical information like political or religious affiliation, but also to continuous parameters like the age of the speaker or other biometric characterizations. Heuristically, reducing the accuracy of continuous parameters thus becomes similar to additive noise, which can be measured by the signal-to-noise ratio.

2) Cryptography and Secure Computing Systems: If the transmitted information is encrypted, then it cannot be used for malicious purposes without access to the key. Surprisingly, however, it is possible to process encrypted information when using *homomorphic encryption*, such that the encrypted result of the computation can be returned and opened [24, 52, 53] (see fig. 17). This can be applied for example in the extraction of spectral features of speech or speech recognition in the cloud, without plain text access to the speech signal [54, 55]. The edge device would then encrypt the speech signal, and send the encrypted data to the cloud which extracts encrypted information, and returns it to the edge device, which can open the encrypted result. While homomorphic encryption in principle provides a beautiful solution to privacy, it comes at a prohibitively great cost in computational cost. The number of encrypted operations performed and bits transmitted increase exponentially with the complexity of the original problem.

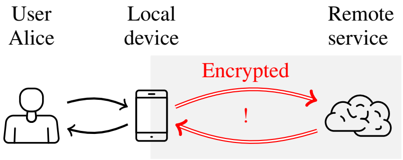

Fig. 17. Protecting privacy by *encrypting* communication and processing. The gray area indicates domain which is encrypted and red, double arrows marked with exclamation mark indicate the protected stream of data.

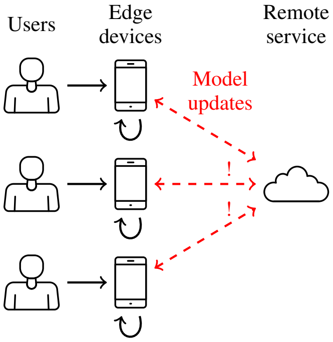

Fig. 18. *Federated learning* as an example of distributed learning, where devices train models locally and transmit only model updates to the cloud. Red dashed arrows marked with exclamation marks indicate the reduced flow of information.

Another approach is *secure multiparty computation* (MPC), where multiple parties can compute a joint function without revealing anything to each other [56]. MPC provides a much smaller overhead in computations and communication, while it can simultaneously be shown that the unlinkability, irreversibility, and renewability of biometric information are granted [57]. It can be applied for example in speaker recognition in the cloud, such that the users' speaker model is not revealed to the cloud and the recognition model is not revealed to the user [58].

3) Distributed learning: A majority of advanced speech processing today uses machine learning, which has to be trained using large databases of speech. The best quality data correspond closely to scenarios where the services are used. Recording users' interactions with their devices is then attractive for training improved models since it corresponds exactly to the use case. This presents a considerable threat to privacy because such unrestricted recording could capture a wide range of private information, including all interactions with the device but potentially also any and all speech in the vicinity of the device.

Distributed learning is an approach to training models

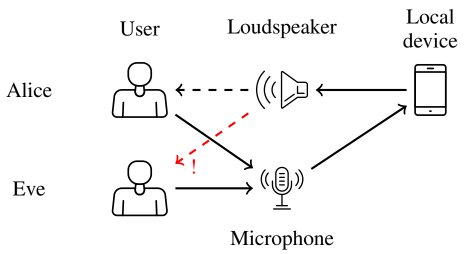

Fig. 19. Protection against a reproduction leak by *authorization tracking* of users, where the device keeps track of people present in the room such that private information is not shared with unauthorized listeners (red dashed line and exclamation mark).

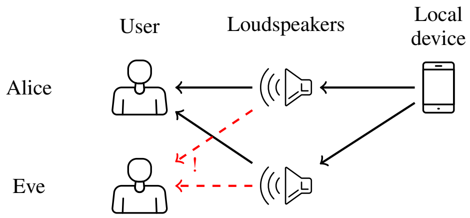

Fig. 20. Protection against a reproduction leak by using *sound zones*, where constructive interference between loudspeakers is used to retain intelligibility for user Alice, and destructive interference distorts it for user Eve (red dashed lines and exclamation mark).

without the need for centralized data collection [59]. Models are trained on local edge devices and only model updates are shared between the nodes and/or with a central server (see fig. 18). Since raw data then never leaves the edge device, this corresponds to a privacy protection approach where information flow is restricted (cf. section III-C1).

Federated learning is one of the flavors of distributed learning, where a central, cloud server collects, merges, and redistributes model updates. It has been used for example in speaker and emotion recognition, language modeling, as well as for unsupervised estimation of microphone clusters in sensor networks [60–67].

Overall, distributed learning is a promising approach, even if it has several challenges. First, constructing, training, and testing systems architectures is much more complicated than regular machine learning. Second, model updates are model-specific and cannot easily be reused if the model structure is updated. The learning accumulated during training is then effectively lost every time the model is updated, which also jeopardizes fair comparison of competing approaches. Third, even if distributed learning does improve privacy, it is not a guarantee for privacy, since model updates can also contain private information [68].

D. Limiting Access to Reproduced Audio

When a user is interacting with a speech interface, then the spoken answer of the speech interface can contain private information. This private information can be overheard by other users in the same acoustic space and that presents a threat to privacy (see fig. 8 and section II-E5). In theory, it would be possible to identify and track users in the same acoustic space and communicate private information only when it does not pose a threat to privacy (see fig. 19). This places great trust in the local device and in its ability to track and identify authorization levels of the people who are present in the acoustic space. Such systems have however not yet been widely published.

Figure 20 presents another solution, which uses constructive and destructive interference between loudspeakers to create *sound zones* where the private information is, respectively, intelligible or distorted (see e.g. [69–71]). By choosing the spatial location where speech is intelligible and assuming we know where the target user is located, we can thus limit access to private information only to the target speaker. Note that sound zones with destructive interference do receive a partial observation of the private message, but it is (hopefully) distorted to the extent that it is unintelligible. This approach is thus another method that uses distortion of the private message to preserve privacy (cf. section III-A). The central challenges of this approach are to make the constructive sound zone large enough that it allows for small head movements, and to make the destructive interference uniform everywhere else, such that there are no isolated points with constructive interference outside the desired sound zone. A benefit of this approach is however that it requires tracking of only the target listener, which is, while difficult, still much simpler than tracking all the people in the room.

IV. Evaluating Privacy

To evaluate the performance of any privacy-preserving methods, we need performance measures for both utility and privacy, corresponding respectively to the trusted and threat tasks (see fig. 1). The performance measures of a trusted task are defined by the application, they follow the typical procedures of conventional speech processing methodology [72, 73] and thus need not be discussed further here. Only need to note that there is often a trade-off between measures of utility and privacy, such that it is pointless to evaluate only one, but proper experiments should always evaluate the trade-off. The objective is then to define performance measures for measuring the extent of privacy provided by the method.

A. Objective Metrics

Metrics applicable to the above attack model include:

- The *equal error rate (EER)* considers an attacker which makes decisions by applying a threshold to a scoring function, where the threshold is chosen such that false positives and false negatives are equal [20, 74]. An increase in EER means that the attacker has made more errors and privacy has improved.
- The *application-independent log-likelihood-ratio cost function* $C_{llr}^{\min}$ generalizes the EER by considering optimal thresholds over all possible prior probabilities and all possible error cost functions [20, 75].
- *Linkability* is defined as the (log-)likelihood that two datasets are instances from the same or different origin [76].

- *Mutual information* can be used to quantify how much new information about a speaker we gain from a new dataset [77].

Note that all above metrics share two notable weaknesses and these are shared with the differential privacy approach (below, section IV-B). First, they do not protect against future attacks. This is particularly evident for the EER, which measures the performance of one implementation of an attacker. It is clear that a larger, more advanced model could make a more powerful attack. This applies also to the likelihood-based measures, where the definitions of metrics are model-independent, but rely which rely on case-specific models of probability distributions. When those statistical models are improved, then new weaknesses may be discovered.

Second, each of the above metrics is related to single observations. In contrast, speech is a continuous flow of information that gives a series of observations. With each observation, we can reduce the confidence intervals as long as different instances have different probability distributions. Repeated observations will thus, with all certainty, breach privacy for all distinguishable characteristics, when the number of observations is sufficiently high.

A more comprehensive evaluation of objective privacy metrics is provided in [78] and the VoicePrivacy challenge gives a practical example of how to apply the metrics [19].

B. *Differential privacy*

Anonymization and privacy are never absolute. Even if current methods and currently available datasets do not allow us to infer anything private from an anonymized dataset, there is no guarantee that the user would be protected also when the novel methodology is introduced in the future or when new associated datasets become available. We can therefore only make claims about *how well* users are protected. *Differential privacy* is such a theory and methodology for characterizing the extent of protection to privacy for the users [50, 79].

Differential privacy operates with a database with private information, such as the age of n users, from which we calculate the average age m_n. Such population statistics can usually be treated as anonymized when the population size is sufficiently large. However, suppose a new user Alice is enrolled in the database, and the average age is updated to m_{n+1}. If an attacker then gains access to the two averages, m_n and m_{n+1} as well as the number of users before Alice n, then we can trivially find Alice's age as $m_{n+1}(n+1)-m_n n$. In other words, by tracking changes in anonymized information, we were still able to reveal private information!

To protect privacy, we can however add noise to individual measurements (see section III-C1), and this gives some protection against the above demonstrated differential attack. Calculation of the population statistics thus becomes a randomized algorithm $\mathcal{M}$. Formally, we say that the randomized algorithm $\mathcal{M}$ gives *ϵ-differential privacy* if for all data sets D and D', differing by one user, and subset of the output $S \subseteq \mathrm{Range}\,(\mathcal{M})$,

$$\ln \Pr\left[\mathcal{M}\left(D\right) \in S\right] \leq \epsilon + \ln \Pr\left[\mathcal{M}\left(D'\right) \in S\right]. \tag{1}$$

Here $\Pr[\cdot]$ refers to the probability, and ϵ is the loss of privacy. The smaller ϵ the better privacy.

An interpretation of this definition is that *it constrains the effect of any single user on the overall log-likelihood of the output to be smaller than ϵ*. Since log-likelihoods characterize entropy, ϵ thus corresponds to the amount of private information available to the attacker. While differential privacy is defined using membership in a dataset as its basis, it can be applied to any attribute of the speaker. In other words, if an attacker is interested in any particular attribute of a speaker, then an algorithm with ϵ-differential privacy will give at most ϵ nats of information about that attribute. Here the unit nat corresponds to natural units of information, defined as the natural logarithm of likelihood, whereas bits correspond to the base-2 logarithm of the likelihood.

The benefit of differential privacy is that it provides an exact mathematical framework for analysis of the extent of privacy. It is however necessarily always based on a statistical model which approximates the underlying system. Even if differential privacy thus gives exact answers, their reliability then still always depends on the accuracy of the underlying statistical models. Nonetheless, with differential privacy, we can for example analyze the effect of having two parallel sources of information, with differential privacy of ϵ_1 and ϵ_2, respectively. Since the ϵ_k values represent the loss in privacy, then the combined loss of privacy from two sources of information is at most their sum $\epsilon_1 + \epsilon_2$ [79].

Information bottlenecks (see section III-A1, [31]) have a notable parallel with differential privacy. Where information bottlenecks limit the amount of information that is allowed to pass through, eq. (1) also quantifies the amount of information leaked. The difference is that the bottleneck contains both the intended private message and leaked information, whereas eq. (1) contains only leaked information. If the bitrate of the intended private message is known, then obviously the leak size of a bottleneck can be quantified. This interpretation is however based on two implicit assumptions. First, the bottleneck has to be quantized, like in the vector-quantized variational autoencoder (VQ-VAE) approach [80], since a continuous-valued bottleneck can, *in theory*, hold an unlimited amount of information because any N-dimensional space can be mapped to a 1-dimensional scalar using space-filling curves [81]. Second, eq. (1) holds only for a single observation, whereas a time series gives repeated observations. With every new observation, we receive new information. We can expect such repeated observations to be correlated, but nevertheless, with a sufficient amount of observations, we can differentiate between any distinct distributions. In its plain form, differential privacy thus does not give any protection when we can observe a time series for a sufficiently long time.

Methodology of differential privacy has been only recently introduced within speech technology, among others privacy-preserving speech recognition [64], emotion recognition [60] and speaker anonymization [82]. As the mathematical rigor of differential privacy has obvious advantages, it is likely that the adoption of this methodology will increase.

V. Perception and Psychology of Privacy

People tend to have a deep sense of *ownership* of some things, both material and immaterial. In particular, people have a *feeling of ownership* toward information about themselves [83, 84]. Such feelings are to some extent detached from the material consequences of breaches in privacy. For example, even if publishing an audio recording revealing a secret intimate relationship would not have direct economic consequences, it can cause psychological damage and harm personal relationships. The effect of threats to privacy then necessarily becomes a question of psychology and social sciences. To qualitatively or quantitatively measures such effects, we then also need subjective tests with users.

Concurrently, people have well-established social rules regarding human-to-human privacy [16]. Moreover, users also have a tendency to anthropomorphize technology, that is, treat devices and services as if they were human [85, 86], such that they will likely *assume* that devices apply human-like social rules to privacy. Observe that such social rules are related to human-like behavior and performance. That does not reveal whether they are applicable to the super-human performance that computers can possess, such as the ability to integrate information over massive databases and to retain accurate records of events long past. How machines should then behave especially with respect to super-human capabilities, is then not only a question of user-interface design but of moral psychology [87]: We need discussions on a societal level of how automated services *should behave* and how they are allowed to behave (see also section VII).

A. Perception and Experience of Privacy

Irrespective of whether a system *actually* preserves privacy or not, some users can *perceive* the system as threatening and others can be oblivious to the threats it poses [88]. Clearly, users do not like using services they perceive as threatening. Understanding how people perceive privacy is interesting in its own right but such understanding is essential in the design of effective user interfaces.

Perception and experience of privacy can be approached in two alternative ways. We can make user studies where people interact with either machines or each other. The distinction is important in the sense that human-to-human interaction relies on social rules which are well-established and developed over a long time. We can thus expect them to be stable over time, whereas human-computer interaction is continuously evolving as people learn more. Moreover, by using studies of human-to-human interaction, we can learn effects related to human-like performance but can probably not rely on observing effects related to super-human performance. Studies of human-to-human interaction are however always a proxy if the actual target is to design human-to-computer interfaces.

Studies of human-computer interaction are thus characterizing the desired phenomenon directly. The compromise is however that people's understanding of privacy with devices might not reflect the true level of privacy, and conclusions made based on the users' opinions might then not reflect their true preferences *and* those preferences are moving targets. As people have more experiences with and as they learn more about technology, their attitudes and perception of it change.

Despite these shortcomings, both types of experiments are essential for improving our understanding of privacy and for improving technology. For example, human-to-human studies have revealed that people experience privacy differently in different acoustic environments; a noisy cafeteria can be better at masking sounds and defending against eavesdropping than a reverberant hallway [12, 13, 89–91] and multiple-bed patient rooms in hospitals have privacy-concerns [92]. Similarly, human-machine studies have revealed that when a chatbot actively communicates choices related to privacy it improves users' experience of privacy [93], voice interfaces with unknown features cause fear in users, and reduce retention of services [94–96] and breaches in privacy are highly detrimental to the trust in services and reduce users' willingness to use the services, but such trust can be rebuilt [97].

B. User Interface Design

For usable privacy, it is important that information about privacy is readily provided, changes in the extent of privacy are promptly notified, and that users have control of the level of privacy [93, 95, 97–99]. In comparison, visual interfaces can use lights or icons for monitoring and tactile interfaces for controlling the level of privacy. While sound can be used to monitor system status with *sonification* [100, 101], this is not in widespread use within signaling of privacy. It is a compromise between filling the acoustic space with information and the ability and tendency of the human auditory system to block out monotonous sounds. To be effective, the sound should be perceivable when the user consciously wants to check the privacy level and changes in privacy level should evoke correspondingly and appropriately large changes in the soundscape that users consciously register such changes. Furthermore, user interfaces should give correct information about and allow controlling the privacy level [10, 95].

In any case, it is imperative that services are designed such that they *reflect the true extent of privacy*. Observe that it can well be possible to design systems that communicate an advanced level of privacy even when the system does not respect user privacy. In fact, service providers have short-term incentives to follow such approaches to design as long as they improve overall user satisfaction. However, such approaches to design are known as *dark patterns* or *deceptive design patterns* and they are considered unethical [102]. Through deception, such dark patterns lull users to believe they are safe when in fact they are abused. Good design of privacy should therefore actively communicate and enable control of the true extent of privacy. Such design practice is not only ethical but also rewards service providers in the long term by improving retention of that services [97].

VI. Content Categories and Applications

To give a complete picture of the implications of privacy in speech technology, this section provides a brief discussion both on categories of private information, complementing table I as well as the application of such information. Observe that this (nor table I) is not a complete list of the content nor of application categories. The purpose is merely to provide a characterization of work done and challenges in the research field, with accompanying references.

First, note that while all presented categories are (potentially) private information, all *sustained* information can potentially be used to identify a speaker. For example, while the current emotional state cannot alone identify a speaker, the tendency to display emotional states can aid in identification. However, where we want to *verify* that a speaker is who they claim to be, we can use only information which cannot be willfully changed. That is, speech style is a particular example of a property that a good voice actor can freely choose, making it ”easy” to change also for fraudulent purposes. Also note that table I should not be interpreted as a complete nor unambiguous categorization of information, but merely provided as an illustrative example. Second, all information in table I can be considered to be *biometric* information as it can be used to identify a person [74, 78, 103, 104], though some limit the term biometric to refer only physical and behavioral characteristics.

Recognizing the speaker’s identity then becomes the natural starting point for studies in privacy. We can attempt to recognize who is speaking (speaker recognition), verify whether a speaker has the claimed identity (speaker verification), clustering audio to segments with a single speaker (speaker diarization) and we can develop methods for deceiving identity (spoofing), e.g. [33, 39, 40, 57, 58, 61, 62, 105–113]. Similarly, by voice conversion we can anonymize a speaker identity by replacing it with a random identity (anonymization) or a specific one (pseudonymization), e.g. [20, 82, 111, 114–119]. Speaker characterization is the natural complement of speaker idenfication [78, 120]. Such methods related to speaker identity can be applied for example to recognize the user of a device, verify a customer at a bank, or as a voice avatar in online gaming. Similarly, anonymization and pseudonymization can be used to hide the identity from the public media and gaming.

Speech recognition, as in speech-to-text, is probably the largest sub-area of speech research. With respect to privacy, it contains two obvious challenges for privacy. We can try to limit to side-information, such as eliminating all non-text information from the data stream, e.g. [64, 66, 121–125], or we can use natural language processing to anonymize the text content, e.g. [8, 126–130].

Privacy is even more important in always-on applications like wake-word detection, i.e. when the interface is triggered by a specific keyword like ”Computer” in ”Computer, lights-off.”. This application is more sensitive exactly because it is always on as users cannot choose when data is processed [28, 131–133]. The always-on characteristic is also prominent in assisted-living applications, which can ”monitor people’s daily exercises, consumption of calories and sleeping patterns, and to provide coaching interventions to foster positive behaviour” [134]. Such ambient voice interfaces are often implemented through acoustic sensor networks which pose their own challenges [26, 28, 35, 63, 106, 135].

Speech enhancement refers to removing background noises and distortions from the desired speech signal (private message) [72]. This task can have an impact on privacy in two ways. First, the recording environment can reveal private information about the speaker, and second, background noises, like a competing talker, can contain private information [136, 137]. Attenuating background noises and competing speakers as well as removing reverberation can thus improve privacy. In addition, while using a sensor network to capture speech can improve utility, it introduces novel threats as well [13, 63].

Speech is typically considered to be biometric information in its entirety [104, 138], though a tight definition would include only the physical and behavioral properties of the speakers. Such properties include health [139], emotions [60] and gender identity [39, 42, 132, 140], each warranting their own treatment.

VII. Legal and Societal Landscape

Privacy in speech technology has a great impact on both the individual and societal levels, as already discussed in section I. The magnitude of societal impact can be appreciated by recalling the Cambridge Analytica privacy scandal [141, 142] where private information was extracted from social media and used for targeted political advertising. In both speech technology and social media, services operate on massive user bases and involve interaction between multiple users. This exposes both areas to the same magnitude of risks. In the case of Cambridge Analytica, the most famous consequence was that it influenced election results in a large democratic country. Another prominent scandal from biometrics is the case where supposedly anonymized patient data for 10 % of Australians was released to the public, only to later be shown that individuals were readily identifiable [143]. All health data of those patients was thus made public contrary to user preferences and with unknown long-term consequences.

While we have not (yet) seen breaches related to speech technology with consequences of comparable magnitude, these parallels highlight the *potential* effect of breaches. Scandals directly related to speech technology include eavesdropping on private persons by employees and contractors of service providers [1, 2, 144].

Section V-B also makes the point that service and technology providers have clear short-term incentives which conflict with users’ preferences for privacy. We can easily find examples where users do not have real choices in protecting their privacy. For example, suppose all friends and family of a user use a particular platform for social and voice interaction even if its privacy configuration is inadequate. The individual user faces then the choice of either disconnecting from their social network or compromising privacy. This applies to all users with a sufficiently large portion of their communities participating in that platform. This is a version of the prisoner’s dilemma, where no single user has any incentive to transition to a solution that would be best for everyone.

On the individual level, the examples of potential exploits in table II cause clear damage to individuals, including psychological harm in particular (e.g. [139]). Additional dangers include stalkers [145–147]. While such individual damage is

”small” on the societal level, their prevalence makes their joint impact significant [4].

These examples demonstrate the inherent need for society-level regulation of speech technology with respect to privacy. Governments have already responded to this need, with the European Union spearheading the process with the General Data Protection Regulation (GDPR) [148, 149], with the State of California following soon thereafter with the California Consumer Privacy Act (CCPA) of 2018 [150]. While these laws cover only a small percentage of the global population, as cloud services typically operate globally, they need the capability to follow local laws. It can be, in many cases, to apply the strictest laws on all users. Service providers have therefore widely adopted the requirements of GDPR and CCPA and that has likely had a large impact also on users outside the scope of these regulations.

With respect to regulation, an important consequence of the objective measures of privacy in section IV is that our tools and measurements will give as an output only *statistical characterizations* of privacy, but they can never give absolute confidence. This is in stark contrast with the concept of *unique* identifiability used in legal documents, such as the General Data Protection Regulation (GDPR) by the European Union [148], which does not explicitly leave room for statistical uncertainty. This is reflected for example in the Guidelines for virtual voice assistants by the European Data Protection Board, which states that: [138, page 13, §31]

> ... voice data is inherently biometric personal data. As a result, when such data is processed for the purpose of uniquely identifying a natural person ... the processing must have a valid legal basis ...

This leaves the interpretation open. It is possible to argue that it is never possible to obtain absolute confidence in speaker identification such that the GDPR is never triggered. It is also possible to argue that all voice data contains personal information which can be used to uniquely identify a person, such that all processing must have a valid legal basis. Both interpretations lead to absurdity, which suggests that the truth must lie somewhere in the middle. In fact, the GDPR in practice requires (see [138, page 4]) that the design process of voice assistants includes a *data protection impact assessment*, where the risks and consequences are evaluated such that the designer can take appropriate precautions to preserve privacy. Authors of the GDPR are thus clearly aware that it is impossible to give absolute guarantees of privacy, but that the impact assessment (i.e. objective measures of privacy) must necessarily be based on statistical measures, even if such measures have not been defined.

While governments are in the process of regulating privacy, corporations and non-governmental organizations have also realized that proper privacy is an opportunity. For example, the Open Voice Network seeks to develop and standardize open technical standards and ethical guidelines for voice assistance [151] and the MyData Global seeks to help people and organizations to benefit from personal data in a human-centric way [152–154]. Within the research community, the author of this paper has been involved in establishing a special interest group within the International Speech Communication Association (ISCA) devoted to ”Security and Privacy in Speech Communication” [155]. It is as far as we know, the world-wide largest community focused on this topic.

VIII. Discussion and Conclusions

The quality and use of speech interfaces have increased rapidly in recent years. As with any new technology, the rapid progress has also revealed the dangers and in particular the threats to privacy it demonstrably poses. Unprotected users are exposed to threats like stalking, algorithmic stereotyping, harassment, and price gouging. Researchers, service providers, and governments thus have the impetus to protect the users, not only because it is ethical, but also because it makes for better products and long-term business.

This paper is a tutorial on privacy for speech technology. Its most notable contribution is an exhaustive categorization of threats (see fig. 3 and section II). Protections against those threats are further categorized according to whether they relate to the private message or side information. The pertinent difference is that transmitting a private message is the whole purpose of communication and there is not very much we can do to protect it other than encryption. With side-information, that is, all the other information that is bundled into a speech like health status and gender identity, we have a much larger arsenal of protections. The primary approach is however to remove as much of the side information as possible as early as possible. As the private message is all that communication that we need, all side-information should be removed to the extent it is possible. Such removal rapidly however demonstrates that paralinguistic information like speech style is often very useful in conveying the intended message. It is thus not always clear what constitutes the legitimate private message.

The first conclusion from this paper is that the range of possible threats to privacy is vast. Each agent – be it human or device – participating in an interaction as well as the acoustic pathways and network connection through which they are connected, is a potential attack surface. Any actor which can interact with the other agent or listen to the connection is a potential eavesdropper. Since we define privacy as a scenario where an agent is authorized with some access, but over-exceeds that authorization (intentionally or inadvertently), we cannot just cut connections but need more refined designs and methodologies. We thus need to dynamically adjust access according to need. Conversely, systems need to actively monitor the privacy status to determine appropriate actions.

Second, we find that privacy and ethics are largely overlapping challenges. Our ethical values govern our preferences for privacy. Most potential breaches of ethics in speech technology are based on breaching privacy. That means that we need a society-wide ethical discussion about what is allowed with respect to user privacy. Such discussions are needed to prevent an Cambridge Analytica-style scandal for speech technology [142].

A third implication of this paper is that, while research in this field has picked up only very recently, there is already a substantial body of research available. The research is not however mature but in a phase of rapid development, and there are important sub-areas that have not yet seen much work. This makes it a fruitful area for research as we can expect important understanding to be discovered in the coming years.

Particular research questions where the author sees an urgent need for and expects to see new results include:

a) Consent: While management of acquiring informed consent has established traditions and best practices for most interface types [156], speech, audio, and ambient systems are notably unique. Namely, acoustic information is a time-varying stream. Reading out a pages-long consent form before an interaction can start is clearly much too obtrusive and unnecessarily detailed. Privacy requirements also vary over time. Consent should thus be acquired per actual need basis. In addition to being more usable, it would also make choices better connected to the actual needs, since consent is acquired only once it is actually needed.

b) Metrics for Streaming: The available theoretical metrics reflect privacy with respect to a finite dataset, whereas speech is an open-ended stream of data. The consequence is that, in theory, we can resolve any private attribute or identity, provided that it has a unique probability distribution and we have a sufficiently long observation. We would thus need methodologies for characterizing the effect that the length of observation has on privacy.

c) Metrics for Out-of-category Information: The metrics discussed in section IV are all related to specific categories of private information and in particular, we can provide protection only to identified threats. For example, we can measure the threat to privacy related to health information, but that does not tell anything about the threat related to information about ethnic background. We thus need methods for evaluating privacy jointly with respect to *all categories* of private information except the private message.

d) Future-proof Metrics: Metrics are generally based on a model of the signal or the attacker. The metrics are thus subject to change when those models are improved in the future, and it will likely expose new threats. Though it is likely difficult, it would be extremely useful if we could characterize the potential range of threats by, for example, increasing computational complexity.

e) Multi-user interaction: Privacy research is categorically focused on *personal* and *user-centric* privacy. However, speech is by definition communication between multiple agents, and when exposed, threatens *all participants simultaneously*. This is not an issue from a legal point of view, because privacy protections apply to all individual users equally. However, from an authorization and consent management perspective, this is an underappreciated issue. If user A records a discussion with user B, then both clearly have some level of ownership and privacy requirements on that recording. Another case is smart technology with multiple users, like smart TVs; even if one user has consented to data collection, that does not mean that others would agree. We do not yet have any widely accepted standard approaches for handling privacy, ownership, and consent in such multi-user scenarios.

f) Disentangelement: If we could disentangle all categories of speech information as in fig. 15, then it would be easy to anonymize each category to an appropriate degree. This approach thus seemingly solves all our problems. The issue is that we do not yet have sufficiently sophisticated methods to do that. The difficulty in developing disentanglement algorithms is that information categories in table I are vaguely and heuristically defined and there is significant overlap between them. We cannot even demonstrate that this would be a complete list of information categories. Without exact definitions of those categories, we have no hope of developing methods for them. An alternative approach is to use representation learning methods to create unsupervised clustering of information categories. The compromise is that we cannot guarantee that the learned representations correspond to heuristically meaningful categories. Still, since disentanglement is *the ideal* solution, that should continue to be a central focus of research.

g) Perception, Experience, and Design of Privacy: Most of the speech-specific research on privacy has focused on privacy-preserving processing and systems structures. This is useful because it is the mandatory prerequisite for privacy-preserving technology. However, as discussed in section V-B, users' experience of services is to some extent independent of the objective level of privacy. We need much more user studies on how, for example, voice characteristics and word choices influence trust, how the privacy level can be monitored during interactions and how changes are notified, how the environment and content of interaction influence user experiences, etc. By improving the user experience with respect to privacy, we are likely to improve user satisfaction and retention of the overall service, while also improving the service objectively.

In conclusion, threats and breaches of privacy have significant negative consequences on individual, societal, ethical, and economic levels. While further improvements in smart technology are expected to improve the utility of the technology, it likely also introduces new threats. The protection of privacy in speech technology has thus been important already for a long time and the importance is increasing. Fortunately, research in the area has picked up speed and this tutorial presents the most important concepts, approaches, and methodology. It is however likely that fundamental results and new technologies will be introduced in the near future. This is thus an exciting time for researchers in the area.

Acknowledgments

A majority of the content in this paper has risen directly from discussions with a long list of colleagues, who were far too many to include as co-authors. These colleagues include, but are not limited to, in alphabetical order, Chris Brzuska, Mads Græsbøll Christensen, Sneha Das, Nicholas Evans, Johannes Fischer, Florin Ghido, Meeri Haataja, Ivan Habernal, Aki Härmä, Dorothea Kolossa, Michael Laakasuo, Martha Larson, Rainer Martin, Joachim Meyer, Sebastian Möller, Andreas Nautsch, Birgit Popp, Nitin Sawhney, Ingo Siegert, Stephan Sigg, Rech Silas, Isabel Trancoso, Sanna Toropainen, Emmanuel Vincent, Jennifer Williams, and Pablo Pérez Zarazaga. I will remain forever grateful.

References

[1] Dorian Lynskey. "Alexa, are you invading my privacy? – the dark side of our voice assistants". In: *The Guardian* (2019). URL: https://www.theguardian.com/technology/2019/oct/09/alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants.

[2] A. Hern. "Apple contractors regularly hear confidential details' on Siri recordings". In: *The Guardian* (2019). URL: https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings.

[3] Thomas Brewster. "Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find". In: *Forbes* (2021). URL: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/.

[4] Karen Levy and Bruce Schneier. "Privacy threats in intimate relationships". In: *Journal of Cybersecurity* 6.1 (2020). URL: https://doi.org/10.1093/cybsec/tyaa006.

[5] Jacques Penders. "Privacy in (mobile) telecommunications services". In: *Ethics and Information Technology* 6 (2004), pp. 247–260. URL: https://doi.org/10.1007/s10676-005-5605-9.

[6] Rebecca Wong and Daniel Garrie. "Privacy in Electronic Communications: The regulation of VoIP in the EU and the United States". In: *Computer Telecommunications Law Review* (2009), pp. 139–146. URL: https://dx.doi.org/10.2139/ssrn.1466153.

[7] Myrto Arapinis et al. "Analysis of privacy in mobile telephony systems". In: *International Journal of Information Security* 16 (2017), pp. 491–523. URL: https://doi.org/10.1007/s10207-016-0338-9.

[8] Darshini Mahendran, Changqing Luo, and Bridget T. Mcinnes. "Review: Privacy-Preservation in the Context of Natural Language Processing". In: *IEEE Access* 9 (2021), pp. 147600–147612. DOI: 10.1109/ACCESS.2021.3124163.

[9] Jennifer Williams et al. "Revisiting Speech Content Privacy". In: *1st ISCA Symposium of the Security & Privacy in Speech Communication*. 2021. URL: https://arxiv.org/pdf/2110.06760.

[10] Tom Bäckström, Birgit Brüggemeier, and Johannes Fischer. "Privacy in speech interfaces". In: *VDE ITG News* (2020). URL: https://www.vde.com/resource/blob/1991012/07662bec66907573ab254c3d99394ec7/itg-news-juli-oktober-2020-data.pdf.

[11] Alexandru Nelus. "Privacy-preserving audio features for classification and clustering in acoustic sensor networks". PhD thesis. Ruhr-Universität Bochum, Universitätsbibliothek, 2022. DOI: 10.13154/294-9064.

[12] Sneha Das. "Robust and Efficient Methods for Distributed Speech Processing - Perspectives on Coding, Enhancement and Privacy". PhD thesis. Aalto University, 2021. URL: http://urn.fi/URN:ISBN:978-952-64-0576-6.

[13] Pablo Pérez Zarazaga. "Preserving Speech Privacy in Interactions with Ad Hoc Sensor Networks". PhD thesis. Aalto University, 2022. URL: http://urn.fi/URN:ISBN:978-952-64-0972-6.

[14] Manas A Pathak. *Privacy-preserving machine learning for speech processing*. Springer Science & Business Media, 2012.

[15] Rachel Cummings et al. "Challenges towards the Next Frontier in Privacy". In: *arXiv preprint arXiv:2304.06929* (2023). URL: https://doi.org/10.48550/arXiv.2304.06929.

[16] Sandra Petronio. *Boundaries of privacy: Dialectics of disclosure*. Suny Press, 2002.

[17] Valerian J. Derlega and Alan L. Chaikin. "Privacy and self-disclosure in social relationships". In: *Journal of Social Issues* 33 (1977), pp. 102–115. URL: https://doi.org/10.1111/j.1540-4560.1977.tb01885.x.

[18] Bart Van Der Sloot and Aviva De Groot. *The handbook of privacy studies*. Amsterdam University Press, 2018. URL: https://doi.org/10.1515/9789048540136.

[19] Natalia Tomashenko et al. "The VoicePrivacy 2020 Challenge: Results and findings". In: *Computer Speech & Language* 74 (2022), p. 101362. URL: https://doi.org/10.1016/j.csl.2022.101362.

[20] Mohamed Maouche et al. "A Comparative Study of Speech Anonymization Metrics". In: *Proc. Interspeech*. 2020, pp. 1708–1712. DOI: 10.21437/Interspeech.2020-2248.

[21] Yiming Wang et al. "Wake word detection and its applications". PhD thesis. Johns Hopkins University, 2021. URL: http://jhir.library.jhu.edu/handle/1774.2/64380.

[22] Andrew Liptak. "Amazon's Alexa started ordering people dollhouses after hearing its name on TV". In: *The Verge* (2017). URL: https://www.theverge.com/2017/1/7/14200210/amazon-alexa-tech-news-anchor-order-dollhouse.

[23] Angela Moscaritolo. "Amazon Alexa Sends Family's Private Conversation to Contact". In: *PCMag* (2018). URL: https://www.pcmag.com/news/amazon-alexa-sends-familys-private-conversation-to-contact.

[24] Francisco Teixeira, Alberto Abad, and Isabel Trancoso. "Patient Privacy in Paralinguistic Tasks". In: *Proc. Interspeech 2018*. 2018, pp. 3428–3432. URL: http://dx.doi.org/10.21437/Interspeech.2018-2186.

[25] Sneha Das and Tom Bäckström. "Enhancement by postfiltering for speech and audio coding in ad hoc sensor networks". In: *JASA Express Letters* 1.1 (2021), p. 015206. URL: https://doi.org/10.1121/10.0003208.

[26] Pablo Pérez Zarazaga, Tom Bäckström, and Stephan Sigg. "Acoustic fingerprints for access management in ad-hoc sensor networks". In: *IEEE Access* 8 (2020), pp. 166083–166094. DOI: 10.1109/ACCESS.2020.3022618.

[27] Stephan Sigg, Pablo Perez Zarazaga, and Tom Bäckström. "Provable consent for voice user interfaces". In: *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 2020, pp. 1–4. DOI: 10.1109/PerComWorkshops48775.2020.9156182.

[28] Timm Koppelmann et al. "Clustering-based Wake Word Detection in Privacy-aware Acoustic Sensor Networks". In: *Proc. Interspeech*. 2022, pp. 719–723. DOI: 10.21437/Interspeech.2022-842.

[29] Weisong Shi et al. "Edge computing: Vision and challenges". In: *IEEE internet of things journal* 3.5 (2016), pp. 637–646.

[30] Naftali Tishby, Fernando C Pereira, and William Bialek. "The information bottleneck method". In: *arXiv preprint physics/0004057* (2000). URL: https://arxiv.org/abs/physics/0004057.

[31] Naftali Tishby and Noga Zaslavsky. "Deep learning and the information bottleneck principle". In: *2015 IEEE Information Theory Workshop (ITW)*. 2015, pp. 1–5. DOI: 10.1109/ITW.2015.7133169.

[32] Ali Makhdoumi et al. "From the information bottleneck to the privacy funnel". In: *2014 IEEE Information Theory Workshop (ITW 2014)*. 2014, pp. 501–505. URL: https://doi.org/10.1109/ITW.2014.6970882.

[33] Juan M Perero-Codosero, Fernando M Espinoza-Cuadros, and Luis A Hernández-Gómez. "X-vector anonymization using autoencoders and adversarial training for preserving speech privacy". In: *Computer Speech & Language* (2022), p. 101351. URL: https://doi.org/10.1016/j.csl.2022.101351.

[34] Xin Wang et al. "A Vector Quantized Variational Autoencoder (VQ-VAE) Autoregressive Neural F_0 Model for Statistical Parametric Speech Synthesis". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019), pp. 157–170. URL: https://doi.org/10.1109/TASLP.2019.2950099.

[35] Mohammad Malekzadeh, Richard G Clegg, and Hamed Haddadi. "Replacement autoencoder: A privacy-preserving algorithm for sensory data analysis". In: *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*. 2018, pp. 165–176. URL: https://doi.org/10.1109/IoTDI.2018.00025.

[36] Xugang Lu et al. "Speech enhancement based on deep denoising autoencoder." In: *Interspeech*. Vol. 2013. ISCA. 2013, pp. 436–440. URL: https://doi.org/10.21437/Interspeech.2013-130.

[37] Michael Neumann and Ngoc Thang Vu. "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 7390–7394. URL: https://doi.org/10.1109/ICASSP.2019.8682541.

[38] Jan Chorowski et al. "Unsupervised speech representation learning using WaveNet autoencoders". In: *IEEE/ACM transactions on audio, speech, and language processing* 27.12 (2019), pp. 2041–2053. URL: https://doi.org/10.1109/TASLP.2019.2938863.

[39] Alexandru Nelus et al. "Privacy-Preserving Siamese Feature Extraction for Gender Recognition versus Speaker Identification". In: *Proc. Interspeech*. 2019, pp. 3705–3709. DOI: 10.21437/Interspeech.2019-1148.

[40] Alexandru Nelus et al. "Privacy-Preserving Variational Information Feature Extraction for Domestic Activity Monitoring versus Speaker Identification". In: *Proc. Interspeech*. 2019, pp. 3710–3714. DOI: 10.21437/Interspeech.2019-1703.

[41] Ranya Aloufi, Hamed Haddadi, and David Boyle. "Privacy-preserving voice analysis via disentangled representations". In: *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*. 2020, pp. 1–14. URL: https://doi.org/10.1145/3411495.3421355.

[42] Dimitrios Stoidis and Andrea Cavallaro. “Protecting Gender and Identity with Disentangled Speech Representations”. In: *Proc. Interspeech*. 2021, pp. 1699–1703. DOI: 10.21437/Interspeech.2021-2163.

[43] Kun Zhou, Berrak Sisman, and Haizhou Li. “VAW-GAN for disentanglement and recomposition of emotional elements in speech”. In: *IEEE Spoken Language Technology Workshop (SLT)*. 2021, pp. 415–422. URL: https://doi.org/10.1109/SLT48900.2021.9383526.

[44] Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. “Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge”. In: *Proc. Interspeech*. 2020, pp. 4836–4840. DOI: 10.21437/Interspeech.2020-1693. URL: http://dx.doi.org/10.21437/Interspeech.2020-1693.

[45] Reza Lotfidereshgi and Philippe Gournay. “Cognitive coding of speech”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 7772–7776. URL: https://doi.org/10.1109/ICASSP.2017.7953135.

[46] David A. Patterson et al. “Carbon Emissions and Large Neural Network Training”. In: *CoRR* abs/2104.10350 (2021). URL: https://arxiv.org/abs/2104.10350.

[47] International Speech Communication Association (ISCA). *Comments to ’Guidelines 02/2021 on Virtual Voice Assistants’ of the European Data Protection Board (EDPB)*. Ed. by Tom Bäckström and Andreas Nautsch. 2021. URL: https://edpb.europa.eu/sites/default/files/webform/public_consultation_reply/edpb_comments.pdf.

[48] Yi Xie and Siqing Peng. “How to repair customer trust after negative publicity: The roles of competence, integrity, benevolence, and forgiveness”. In: *Psychology & Marketing* 26.7 (2009), pp. 572–589. URL: https://doi.org/10.1002/mar.20289.

[49] S.C. Chen and G.S. Dhillon. “Interpreting Dimensions of Consumer Trust in E-Commerce”. In: *Information Technology and Management* 4 (2003), pp. 303–318. DOI: https://doi.org/10.1023/A:1022962631249.

[50] Cynthia Dwork, Aaron Roth, et al. “The algorithmic foundations of differential privacy”. In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407. URL: http://dx.doi.org/10.1561/0400000042.

[51] Stanley L Warner. “Randomized response: A survey technique for eliminating evasive answer bias”. In: *Journal of the American Statistical Association* 60.309 (1965), pp. 63–69. URL: https://doi.org/10.1080/01621459.1965.10480775.

[52] Frederik Armknecht et al. “A guide to fully homomorphic encryption.” In: *IACR Cryptology ePrint Archive* 2015 (2015), p. 1192.

[53] Abbas Acar et al. “A survey on homomorphic encryption schemes: Theory and implementation”. In: *ACM Computing Surveys (Csur)* 51.4 (2018), pp. 1–35. URL: https://doi.org/10.1145/3214303.

[54] Patricia Thaine and Gerald Penn. “Extracting Mel-Frequency and Bark-Frequency Cepstral Coefficients from Encrypted Signals”. In: *Proc. Interspeech*. 2019, pp. 3715–3719. DOI: 10.21437/Interspeech.2019-1136.

[55] Shi-Xiong Zhang, Yifan Gong, and Dong Yu. “Encrypted Speech Recognition Using Deep Polynomial Networks”. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 5691–5695. DOI: 10.1109/ICASSP.2019.8683721.

[56] Ronald Cramer, Ivan Damgård, and J.B. Nielsen. *Secure multiparty computation and secret sharing*. Cambridge, July 2015. URL: cambridge.org/9781107043053.

[57] Amos Treiber et al. “Privacy-preserving PLDA speaker verification using outsourced secure computation”. In: *Speech Communication* 114 (2019), pp. 60–71. URL: https://doi.org/10.1016/j.specom.2019.09.004.

[58] Francisco Teixeira et al. “Towards End-to-End Private Automatic Speaker Recognition”. In: *Proc. Interspeech*. 2022, pp. 2798–2802. DOI: 10.21437/Interspeech.2022-10672.

[59] Bo Liu et al. “When machine learning meets privacy: A survey and outlook”. In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–36. URL: https://doi.org/10.1145/3436755.

[60] Tiantian Feng, Raghuveer Peri, and Shrikanth Narayanan. “User-Level Differential Privacy against Attribute Inference Attack of Speech Emotion Recognition on Federated Learning”. In: *Proc. Interspeech*. 2022, pp. 5055–5059. DOI: 10.21437/Interspeech.2022-10060.

[61] Filip Granqvist et al. “Improving On-Device Speaker Verification Using Federated Learning with Privacy”. In: *Proc. Interspeech*. 2020, pp. 4328–4332. DOI: 10.21437/Interspeech.2020-2944.

[62] Abraham Zewoudie and Tom Bäckström. “Federated Learning for Privacy-Preserving Speaker Recognition”. In: *IEEE Access* 9 (2021), pp. 149477–149485. URL: https://doi.org/10.1109/ACCESS.2021.3124029.

[63] Alexandru Nelus, Rene Glitza, and Rainer Martin. “Estimation of Microphone Clusters in Acoustic Sensor Networks Using Unsupervised Federated Learning”. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 761–765. DOI: 10.1109/ICASSP39728.2021.9414186.

[64] Michael Shoemate et al. *Sotto Voce: Federated Speech Recognition with Differential Privacy Guarantees*. 2022. DOI: 10.48550/ARXIV.2207.07816. URL: https://arxiv.org/abs/2207.07816.

[65] Jae Ro et al. “Communication-Efficient Agnostic Federated Averaging”. In: *Proc. Interspeech*. 2021, pp. 871–875. DOI: 10.21437/Interspeech.2021-153.

[66] Tuan Nguyen et al. *Federated Learning for ASR based on Wav2vec 2.0*. 2023. arXiv: 2302.10790 [eess.AS]. URL: https://doi.org/10.48550/arXiv.2302.10790.

[67] Tiantian Feng et al. “Attribute inference attack of speech emotion recognition in federated learning settings”. In: *arXiv preprint arXiv:2112.13416* (2021). URL: https://arxiv.org/abs/2112.13416.

[68] Tribhuvanesh Orekondy et al. *Gradient-Leaks: Understanding and Controlling Deanonymization in Federated Learning*. 2018. DOI: 10.48550/ARXIV.1805.05838. URL: https://arxiv.org/abs/1805.05838.

[69] Jacob Donley, Christian Ritz, and W. Bastiaan Kleijn. “Improving speech privacy in personal sound zones”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pp. 311–315. DOI: 10.1109/ICASSP.2016.7471687.

[70] Jesper Kjær Nielsen et al. “Sound Zones As An Optimal Filtering Problem”. In: *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. 2018, pp. 1075–1079. DOI: 10.1109/ACSSC.2018.8645268.

[71] Daniel Wallace and Jordan Cheer. “Combining background noise and artificial masking to achieve privacy in sound zones”. In: *Computer Speech & Language* 72 (2022), p. 101285. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2021.101285. URL: https://www.sciencedirect.com/science/article/pii/S0885230821000875.

[72] Jacob Benesty, M Mohan Sondhi, Yiteng Huang, et al. *Springer handbook of speech processing*. Vol. 1. Springer, 2008. URL: https://doi.org/10.1007/978-3-540-49127-9.

[73] Tom Bäckström et al. *Introduction to Speech Processing*. 2nd ed. Aalto University, 2022. URL: http://speechprocessingbook.aalto.fi.

[74] *ISO/IEC 19795-1:2021 Information technology – Biometric performance testing and reporting – Part 1: Principles and framework*. Standard. Geneva, CH: International Organization for Standardization, May 2021. URL: https://www.iso.org/standard/73515.html.

[75] Niko Brümmer and Johan Du Preez. “Application-independent evaluation of speaker detection”. In: *Computer Speech & Language* 20.2-3 (2006), pp. 230–275. URL: https://doi.org/10.1016/j.csl.2005.08.001.

[76] Marta Gomez-Barrero et al. “General Framework to Evaluate Unlinkability in Biometric Template Protection Systems”. In: *IEEE Transactions on Information Forensics and Security* 13.6 (2018), pp. 1406–1420. DOI: 10.1109/TIFS.2017.2788000.

[77] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. URL: www.cambridge.org/9780521642989.

[78] Andreas Nautsch et al. “Preserving privacy in speaker and speech characterisation”. In: *Computer Speech & Language* 58 (2019), pp. 441–480. URL: https://doi.org/10.1016/j.csl.2019.06.001.

[79] Cynthia Dwork. “Differential privacy: A survey of results”. In: *International conference on theory and applications of models of computation*. Springer. 2008, pp. 1–19. URL: https://doi.org/10.1007/978-3-540-79228-4_1.

[80] Cristina Gârbacea et al. “Low bit-rate speech coding with VQ-VAE and a WaveNet decoder”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 735–739. URL: https://doi.org/10.1109/ICASSP.2019.8683277.

[81] Mohamed F Mokbel, Walid G Aref, and Ibrahim Kamel. “Performance of multi-dimensional space-filling curves”. In: *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*. 2002, pp. 149–154. URL: https://doi.org/10.1145/585147.585179.

[82] Ali Shahin Shamsabadi et al. "Differentially Private Speaker Anonymization". In: *Proceedings on Privacy Enhancing Technologies* 1 (2023), pp. 98–114. URL: https://doi.org/10.56553/popets-2023-0007.

[83] Patrick Cichy, Torsten-Oliver Salge, and Rajiv Kohli. "Extending the privacy calculus: the role of psychological ownership". In: *ICIS 2014 Proceedings*. Vol. 30. 2014. URL: https://aisel.aisnet.org/icis2014/proceedings/ISSecurity/30.

[84] Sarah Dawkins et al. "Psychological ownership: A review and research agenda". In: *Journal of Organizational Behavior* 38.2 (2017), pp. 163–183. URL: https://doi.org/10.1002/job.2057.

[85] Clifford Nass et al. "Anthropomorphism, agency, and ethopoeia: computers as social actors". In: *INTERACT'93 and CHI'93 conference companion on Human factors in computing systems*. 1993, pp. 111–112. URL: https://doi.org/10.1145/259964.260137.

[86] Samia Cornelius and Dorothy Leidner. "Acceptance of anthropomorphic technology: a literature review". In: (2021). URL: https://doi.org/10.24251/HICSS.2021.774.

[87] Steven John Thompson. *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*. IGI Global, 2021. DOI: 10.4018/978-1-7998-4894-3.

[88] Leonardo Pepino et al. *Detecting Distrust Towards the Skills of a Virtual Assistant Using Speech*. 2020. DOI: 10.48550/ARXIV.2007.15711. URL: https://arxiv.org/abs/2007.15711.

[89] Pablo Pérez Zarazaga et al. "Sound Privacy: A Conversational Speech Corpus for Quantifying the Experience of Privacy". In: *Proc. Interspeech*. 2019, pp. 3720–3724. DOI: 10.21437/Interspeech.2019-1172.

[90] Tom Bäckström et al. "Intuitive Privacy from Acoustic Reach: A Case for Networked Voice User-Interfaces". In: *Proc. 2021 ISCA Symposium on Security and Privacy in Speech*. 2021. URL: https://research.aalto.fi/files/67799021/SIG_Symposium2021_Privacy_2_.pdf.

[91] Anna Leschanowsky et al. "Perception of Privacy Measured in the Crowd — Paired Comparison on the Effect of Background Noises". In: *Proc. Interspeech*. 2020, pp. 4651–4655. DOI: 10.21437/Interspeech.2020-2299. URL: http://dx.doi.org/10.21437/Interspeech.2020-2299.

[92] Jikke Reinten et al. "Speech privacy in multiple-bed patient rooms". In: *Healthy Buildings Europe 2017, HB 2017*. International Society of Indoor Air Quality and Climate (ISIAQ). 2017, Paper–ID. URL: https://research.tue.nl/en/publications/speech-privacy-in-multiple-bed-patient-rooms.

[93] Birgit Brüggemeier and Philip Lalone. "Perceptions and reactions to conversational privacy initiated by a conversational user interface". In: *Computer Speech & Language* 71 (2022), p. 101269. URL: https://doi.org/10.1016/j.csl.2021.101269.

[94] Farida Yeasmin, Sneha Das, and Tom Bäckström. "Privacy Analysis of Voice User Interfaces". In: *Proc. 1st International Workshop on the Internet of Sound*. 2020. DOI: 10.5281/zenodo.4026514.

[95] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. "Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (Nov. 2018). DOI: 10.1145/3274371. URL: https://doi.org/10.1145/3274371.

[96] Christoph Lutz and Gemma Newlands. "Privacy and smart speakers: A multi-dimensional approach". In: *The Information Society* 37.3 (2021), pp. 147–162. URL: https://doi.org/10.1080/01972243.2021.1897914.

[97] Miriam Kurz, Birgit Brüggemeier, and Michael Breiter. "Success is not Final; Failure is not Fatal – Task Success and User Experience in Interactions with Alexa, Google Assistant and Siri". In: *Human-Computer Interaction. Design and User Experience Case Studies*. Ed. by Masaaki Kurosu. Cham: Springer International Publishing, 2021, pp. 351–369. ISBN: 978-3-030-78468-3.

[98] Preston So. *Voice content and usability*. A book a part, 2021. URL: https://abookapart.com/products/voice-content-and-usability.

[99] Michael Lewis. "Designing for human-agent interaction". In: *AI Magazine* 19.2 (1998), pp. 67–67. URL: https://doi.org/10.1609/aimag.v19i2.1369.

[100] Tobias Hildebrandt, Thomas Hermann, and Stefanie Rinderle-Ma. "Continuous sonification enhances adequacy of interactions in peripheral process monitoring". In: *International Journal of Human-Computer Studies* 95 (2016), pp. 54–65. URL: https://doi.org/10.1007/s00779-020-01394-3.

[101] Michael Iber et al. "Auditory augmented process monitoring for cyber physical production systems". In: *Personal and Ubiquitous Computing* 25 (2021), pp. 691–704. URL: https://doi.org/10.1016/j.ijhcs.2016.06.002.

[102] Lauren E Willis. "Deception by design". In: *Harvard Journal of Law & Technology* 34 (2020), p. 115. URL: https://ssrn.com/abstract=3694575.

[103] Andy Adler, Richard Youmaran, and Sergey Loyka. "Towards a measure of biometric information". In: *2006 Canadian conference on electrical and computer engineering*. IEEE. 2006, pp. 210–213. URL: https://doi.org/10.1109/CCECE.2006.277447.

[104] Andreas Nautsch et al. "The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment". In: *Proc. Interspeech*. 2020, pp. 1698–1702. DOI: 10.21437/Interspeech.2020-1815.

[105] Oubaïda Chouchane et al. "Privacy-Preserving Voice Anti-Spoofing Using Secure Multi-Party Computation". In: *Proc. Interspeech*. 2021, pp. 856–860. DOI: 10.21437/Interspeech.2021-983.

[106] Abraham Woubie Zewoudie, Tom Bäckström, and Pablo Pérez Zarazaga. "The Use of Audio Fingerprints for Authentication of Speakers on Speech Operated Interfaces". In: *Proc. 2021 ISCA Symposium on Security and Privacy in Speech*. 2021. URL: https://research.aalto.fi/files/75674953/Woubie_Use_of_Audio_Fingerprints_isca.pdf.

[107] Zhongxin Bai and Xiao-Lei Zhang. "Speaker recognition based on deep learning: An overview". In: *Neural Networks* 140 (2021), pp. 65–99. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2021.03.004.

[108] Yaowei Han et al. "Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release". In: *International Conference on Multimedia and Expo (ICME)*. IEEE. 2020, pp. 1–6. URL: https://doi.org/10.1109/ICME46284.2020.9102875.

[109] Tomi Kinnunen and Haizhou Li. "An overview of text-independent speaker recognition: From features to supervectors". In: *Speech Communication* 52.1 (2010), pp. 12–40. ISSN: 0167-6393. URL: https://doi.org/10.1016/j.specom.2009.08.009.

[110] John HL Hansen and Taufiq Hasan. "Speaker recognition by machines and humans: A tutorial review". In: *IEEE Signal processing magazine* 32.6 (2015), pp. 74–99. URL: https://doi.org/10.1109/MSP.2015.2462851.

[111] Candy Olivia Mawalim et al. "X-Vector Singular Value Modification and Statistical-Based Decomposition with Ensemble Regression Modeling for Speaker Anonymization System". In: *Proc. Interspeech*. 2020, pp. 1703–1707. DOI: 10.21437/Interspeech.2020-1887.

[112] Andreas Nautsch et al. "Privacy-Preserving Speaker Recognition with Cohort Score Normalisation". In: *Proc. Interspeech*. 2019, pp. 2868–2872. DOI: 10.21437/Interspeech.2019-2638.

[113] Paul-Gauthier Noé et al. "Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation". In: *Proc. Interspeech*. 2021, pp. 1902–1906. DOI: 10.21437/Interspeech.2021-1712.

[114] Paul-Gauthier Noé et al. "Speech Pseudonymisation Assessment Using Voice Similarity Matrices". In: *Proc. Interspeech*. 2020, pp. 1718–1722. DOI: 10.21437/Interspeech.2020-2720.

[115] Champion Pierre, Anthony Larcher, and Denis Jouvet. "Are disentangled representations all you need to build speaker anonymization systems?" In: *Proc. Interspeech*. 2022, pp. 2793–2797. DOI: 10.21437/Interspeech.2022-10586.

[116] Gauri P. Prajapati et al. "Voice Privacy Through x-Vector and CycleGAN-Based Anonymization". In: *Proc. Interspeech*. 2021, pp. 1684–1688. DOI: 10.21437/Interspeech.2021-1573.

[117] Brij Mohan Lal Srivastava et al. "Design Choices for X-Vector Based Speaker Anonymization". In: *Proc. Interspeech*. 2020, pp. 1713–1717. DOI: 10.21437/Interspeech.2020-2692.

[118] Brij Mohan Lal Srivastava et al. "Privacy and Utility of X-Vector Based Speaker Anonymization". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 2383–2395. DOI: 10.1109/TASLP.2022.3190741.

[119] Fuming Fang et al. "Speaker Anonymization Using X-vector and Neural Waveform Models". In: *Proc. 10th ISCA Speech Synthesis Workshop*. 2019, pp. 155–160. DOI: 10.21437/SSW.2019-28. URL: http://dx.doi.org/10.21437/SSW.2019-28.

[120] Phuoc Nguyen et al. "Automatic classification of speaker characteristics". In: *International Conference on Communications and Electronics 2010*. IEEE. 2010, pp. 147–152. URL: https://doi.org/10.1109/ICCE.2010.5670700.

[121] Ranya Aloufi, Hamed Haddadi, and David Boyle. "Configurable Privacy-Preserving Automatic Speech Recognition". In: *Proc. Interspeech*. 2021, pp. 861–865. DOI: 10.21437/Interspeech.2021-1783.

[122] Abhinav Garg et al. "Streaming On-Device End-to-End ASR System for Privacy-Sensitive Voice-Typing". In: *Proc. Interspeech*. 2020, pp. 3371–3375. DOI: 10.21437/Interspeech.2020-3172.

[123] Muhammad A. Shah et al. "Evaluating the Vulnerability of End-to-End Automatic Speech Recognition Models to Membership Inference Attacks". In: *Proc. Interspeech*. 2021, pp. 891–895. DOI: 10.21437/Interspeech.2021-1188.

[124] Brij Mohan Lal Srivastava et al. "Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?" In: *Proc. Interspeech*. 2019, pp. 3700–3704. DOI: 10.21437/Interspeech.2019-2415.

[125] Natalia Tomashenko et al. "Privacy attacks for automatic speech recognition acoustic models in a federated learning framework". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6972–6976. URL: https://doi.org/10.1109/ICASSP43922.2022.9746541.

[126] Yuchen Liu, Apu Kapadia, and Donald Williamson. "Preventing sensitive-word recognition using self-supervised learning to preserve user-privacy for automatic speech recognition". In: *Proc. Interspeech*. 2022, pp. 4207–4211. DOI: 10.21437/Interspeech.2022-85.

[127] Mohamed Maouche et al. "Enhancing speech privacy with slicing". In: *Proc. Interspeech*. 2022. URL: https://hal.inria.fr/hal-03369137/.

[128] David Ifeoluwa Adelani et al. "Privacy Guarantees for De-Identifying Text Transformations". In: *Proc. Interspeech*. 2020, pp. 4666–4670. DOI: 10.21437/Interspeech.2020-2208.

[129] Scott Novotney, Yile Gu, and Ivan Bulyko. "Adjunct-Emeritus Distillation for Semi-Supervised Language Model Adaptation". In: *Proc. Interspeech*. 2021, pp. 866–870. DOI: 10.21437/Interspeech.2021-27.

[130] Ivan Habernal. "When differential privacy meets NLP: The devil is in the detail". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 1522–1528. URL: http://dx.doi.org/10.18653/v1/2021.emnlp-main.114.

[131] Timm Koppelmann et al. "Privacy-Preserving Feature Extraction for Cloud-Based Wake Word Verification". In: *Proc. Interspeech*. 2021, pp. 876–880. DOI: 10.21437/Interspeech.2021-262.

[132] Dimitrios Stoidis and Andrea Cavallaro. "Generating gender-ambiguous voices for privacy-preserving speech recognition". In: *Proc. Interspeech*. 2022, pp. 4237–4241. DOI: 10.21437/Interspeech.2022-11322.

[133] Chao-Han Huck Yang, Sabato Marco Siniscalchi, and Chin-Hui Lee. "PATE-AAE: Incorporating Adversarial Autoencoder into Private Aggregation of Teacher Ensembles for Spoken Command Classification". In: *Proc. Interspeech*. 2021, pp. 881–885. DOI: 10.21437/Interspeech.2021-640.

[134] Fasih Haider and Saturnino Luz. "A System for Real-Time Privacy Preserving Data Collection for Ambient Assisted Living". In: *Proc. Interspeech*. 2019, pp. 2374–2375. URL: https://isca-speech.org/archive/pdfs/interspeech_2019/haider19_interspeech.pdf.

[135] Matt O'Connor and W. Bastiaan Kleijn. "Distributed Summation Privacy for Speech Enhancement". In: *Proc. Interspeech*. 2020, pp. 4646–4650. DOI: 10.21437/Interspeech.2020-1977.

[136] Pablo Pérez Zarazaga et al. "Cancellation of Local Competing Speaker with Near-Field Localization for Distributed ad-hoc Sensor Network". In: *Proc. Interspeech*. 2021, pp. 676–680. DOI: 10.21437/Interspeech.2021-1329.

[137] Silas Rech. "Multi-Device Speech Enhancement for Privacy and Quality". MA thesis. Aalto University, 2022.

[138] European Data Protection Board. *Guidelines 02/2021 on virtual voice assistants*. 2021.

[139] Deepika Natarajan et al. "PRIORIS: Enabling Secure Detection of Suicidal Ideation from Speech Using Homomorphic Encryption". In: *Protecting Privacy through Homomorphic Encryption*. Springer, 2021, pp. 133–146. URL: https://doi.org/10.1007/978-3-030-77287-1_10.

[140] Alexandru Nelus and Rainer Martin. "Gender Discrimination Versus Speaker Identification Through Privacy-Aware Adversarial Feature Extraction". In: *Speech Communication; 13th ITG-Symposium*. 2018, pp. 1–5. URL: https://ieeexplore.ieee.org/abstract/document/8578003.

[141] Christophe Olivier Schneble, Bernice Simone Elger, and David Shaw. "The Cambridge Analytica affair and Internet-mediated research". In: *EMBO reports* 19.8 (2018), e46579. URL: https://doi.org/10.15252/embr.201846579.

[142] Melissa Heikkilä. *A Cambridge Analytica-style scandal for AI is coming*. 2023. URL: https://www.technologyreview.com/2023/04/25/1072177/a-cambridge-analytica-style-scandal-for-ai-is-coming/.

[143] Chris Culnane, Benjamin IP Rubinstein, and Vanessa Teague. "Health data in an open world – A report on re-identifying patients in the MBS/PBS dataset and the implications for future releases of a Australian gogernment data". In: *arXiv:1712.05627* (2017). URL: https://doi.org/10.48550/arXiv.1712.05627.

[144] M. Day, G. Turner, and N. Drozdiak. "Amazon Workers Are Listening to What You Tell Alexa". In: *Bloomberg.com* (2019). URL: https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio.

[145] Donna Lu. "How Abusers Are Exploiting Smart Home Devices". In: *Vice* (2019). URL: https://www.vice.com/en/article/d3akpk/smart-home-technology-stalking-harassment.

[146] Stan Hulsen. "Stalkers intimideren ex-partners via slimme camera's, lampen en alarmen". In: *RTL Nieuws* (2022). URL: https://www.rtlnieuws.nl/nieuws/nederland/artikel/5345441/stalkers-intimideren-ex-partners-slimme-cameras-lampen-en-alarmen.

[147] Avast Software s.r.o. *Use of Stalkerware and Spyware Apps Increase by 93% since Lockdown Began in the UK*. 2021. URL: https://press.avast.com/use-of-stalkerware-and-spyware-apps-increase-by-93-since-lockdown-began-in-the-uk.

[148] European Parliament. *Directive 95/46/EC General Data Protection Regulation*. 2016. URL: http://data.europa.eu/eli/reg/2016/679.

[149] Andreas Nautsch et al. "The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding". In: *Proc. Interspeech*. 2019, pp. 3695–3699. DOI: 10.21437/Interspeech.2019-2647.

[150] State of California. *California Consumer Privacy Act of 2018 (CCPA)*. 2018. URL: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.

[151] *Open Voice Network*. 2023. URL: https://openvoicenetwork.org.

[152] *MyData Global*. 2023. URL: https://www.mydata.org.

[153] A Poikola, K Kuikkaniemi, and H Honko. *MyData–White Paper*. 2016.

[154] Antti Poikola and Kai Kuikkaniemi end Harri Honko. *MyData – A Nordic Model for human-centered personal data management and processing*. 2015. URL: http://urn.fi/URN:ISBN:978-952-243-455-5.

[155] *ISCA SIG "Security and Privacy in Speech Communication"*. 2023. URL: https://www.spsc-sig.org/.

[156] Batya Friedman, Edward Felten, and Lynette I. Millett. *Informed consent online: A conceptual model and design principles.* Tech. rep. University of Washington Computer Science & Engineering Technical Report 00–12–2 8, 2000. URL: https://dada.cs.washington.edu/research/tr/2000/12/UW-CSE-00-12-02.pdf.

Tom Bäckström D.Sc. (tech.) is an associate professor at Aalto University, Finland (2019-). He obtained his Master's and Doctoral degrees at Helsinki University of Technology (the predecessor of Aalto) in 2001 and 2004, respectively. During his time at International Audio Laboratories Erlangen, Germany (2008-2016), he made contributions to several international speech and audio coding standards such as MPEG USAC and 3GPP EVS, and became a professor (W2) at Friedrich-Alexander University Erlangen-Nürnberg (FAU) (2012-2016). Before his current position, he was a professor practice at Aalto University (2016-2019). He was the initiator and chair of ISCA SIG "Security and Privacy in Speech Communication" (2019-2022). His current research interests include privacy, coding, enhancement and transmission of speech as well as machine learning.