Security and privacy in speech technology

Contents

DISCLAIMER: *This document is meant to be an introduction to questions in security and privacy in speech technology for engineering students, such that they would understand the main problematic. In particular, this is not a legal document. In real-life application of technology and data collection, you must consult legal experts to determine whether you follow the law. You are responsible.*

The right to privacy is a widely accepted concept though its definition varies. It is however clear that people tend to think that some things are “theirs”, that they have ownership of *things*, including information about themselves. A possible definition of privacy would then be “the absence of attention from others” and correspondingly security could be defined as the protection of that which one owns, including material and immaterial things. It however must be emphasised that there are no widely shared and accepted definitions and in particular, the legal community has a wide range of definitions depending on the context and field of application.

From the perspective of speech technology, security and privacy has two principal aspects;

- Security and privacy of data related to speech signals and
- Protection against attacks which use speech signals as a tool.

The latter aspect is mainly related to speaker identity; fraudsters can for example synthesise speech which mimics (spoofs) a target person to gain access to restricted systems, such as access to the bank account of the target person. Such use cases fall mainly under the discussions under speaker recognition and verification, and not discussed further here.

Observe that in the isolated category of telephony (classical telephone connections) privacy and security already have well-established ethical standards as well as legislation. In typical jurisdictions, telephone calls are private in the sense that only the “intended” participants can listen to them and sometimes even recording them is restricted. Covert listening is usually allowed only for the police and even for them only in specially regulated situations, such as with a permission granted by a court or judge.

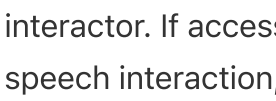
[Lareo, 2019, Nautsch *et al.*, 2019]

13.1. System models

13.1.1. All-human interaction

Speech is a tool for communication such that it is generally sensible to always discuss interactions between two agents, say, Alice and Bob. The interaction between them is the desired function such that the information exchanged there is explicitly permitted. By choosing to talk with each other, they both reveal information to the extent speech contains such information.

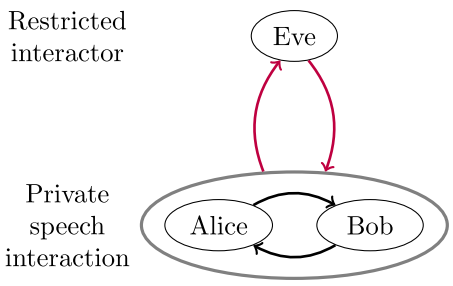
13.1.1.1. Primary interaction



Though Alice and Bob knowingly and intentionally interact, they might reveal private things. This is the classic "*slip of the tongue*". [Petronio, 2002]

13.1.1.2. Secondary interactions

A second-order question are third parties, who are not part of the main speech interaction. The pertinent question is the degree to which the third party is allowed to partake in an interaction. As a practical example, suppose Alice and Bob have a romantic dinner at a restaurant. To which extent is the waitress Eve allowed to interact with the discussion of Alice and Bob? Clearly Eve has some necessary tasks such that interaction is unavoidable. Will Alice and Bob, for example, pause their discussion when Eve approaches?



Observe that we have here labelled Eve as a "*restricted*" and not as an "*unauthorized*" interactor. If access is unauthorized, then it is clear that Eve should not have any access to the speech interaction, which is generally straightforward to handle. The word restricted, on the other hand, implies that unimpeded access should not be granted, but that some access can be allowed. It is thus not question of "*if*" access should be granted but "*how much?*".

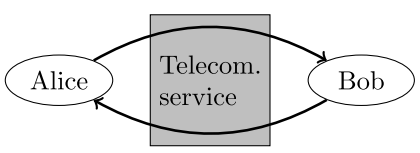
13.1.1.3. Ownership and personal privacy

Privacy is closely connected to ownership of immaterial property, that is, information. Such ownership can also be translated to the question of *who has the control* over some information? In terms of *personal privacy*, it is clear that it relates to information to which a single person can claim ownership.

Speech is more complicated. Speech is a form of communication and thus relates to an interaction between two parties. Dialogues can also commonly lead to co-creation of meaning, where new information is generated through the dialogue in a form which none of the involved parties could have alone produced [Gasiorek, 2018]. None of the users can thus claim sole ownership of the information, but the ownership is shared. Currently we do not have the tools for handling such shared ownership.

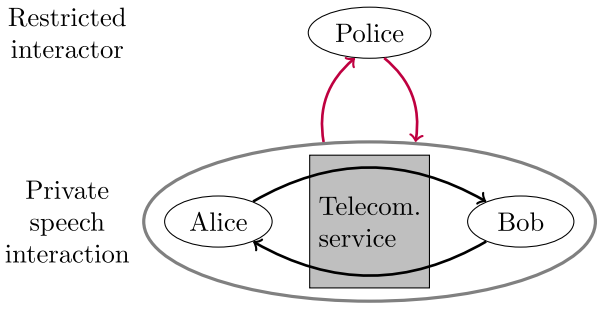
13.1.2. Interactions which involve devices or services

13.1.2.1. Telecommunication



Talking over the phone and video conference involve transmission of speech through a telecommunication service. Here we consider scenarios where the telecommunication device does not include any advanced functionalities or artificial intelligence. Most countries have

clearly definied rules that specify the situations when such communication can be eavesdropped. In most jurisdictions, only the police is allowed to intercept such traffic and only in specific situations.

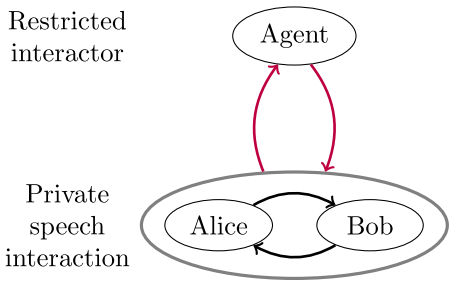


Such interception of private communication is interesting primarily from ethical and legal perspectives, but does not contain technological challenges related to the communication itself. The main technological challenges are related to forensics;

- What information can the police extract (e.g. speaker identity, emotions and health)?
- How can speakers (and service providers) protect themselves from unauthorized interception by, for example, stripping away information such as speaker identity, from the transmitted signal?

13.1.2.2. Discussion in the presence of speech interfaces

A commonly occuring scenario is one where two or more users engage in a discussion such that there is one or more speech operated devices nearby. For example, a user could have their mobile phone or there could be a smart speaker nearby.

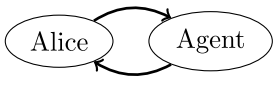


Often, one of the users will be the primary user of said devices (e.g. it is *their* phone), so the question is how the devices should relate to the other users. For example, suppose Alice has a smart speaker at home and Bob comes to visit. What would be the appropriate approach then for both Alice and the agent? Should Alice or the agent notify Bob of the presence of an agent? Or should the agent automatically detect the presence of Bob and change its behaviour (e.g. go to sleep)?

We seem to lack both the cultural habits which dictate how to handle such situations, the legal tools which regulates such situations as well as the technical tools to manage multi-user access.

13.1.2.3. Interaction with a speech interface

An interaction with a speech interface or agent is surprisingly free of problems as long as the agent is not connected to any outside entity. We can think of the agent as a local device. If nobody else has access to that device, then all the information remains in the user's direct control. Even if the agent exist in a remote cloud-service, if information remains strictly within the desired service, there are very little problems to consider.

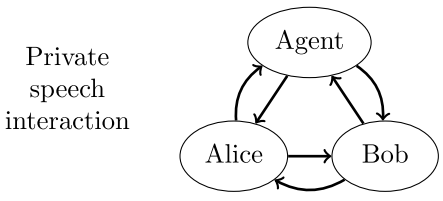


An exception is analysis, which the agent can perform, which is not related to the desired service, which can take abusive forms. For example, suppose the agent analyses the user's voice for health problems and identifies that the user has Alzheimer's disease. What should the agent then do with that information? Not doing anthing seems unethical - getting early access to medical services could greatly improve life quality. Informing the user, on the other

hand, involves risks. How will the user react to the information? Is the user sufficiently psychologically stable to handle it? What if the analysis is incorrect and the agent thus causes suffering? It is also easy to think of further problematic scenarios. [König *et al.*, 2015]

13.1.2.4. Multi-user interaction with a speech interface

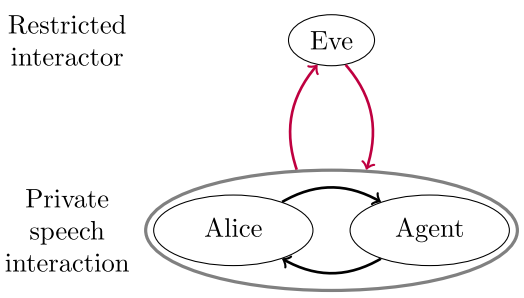
An agent can be involved in an interaction with multiple users at the same time.



This scenario differs from the single-user case in particular in the way the device can store information. To which extent should different combinations of users be permitted to have access to information from prior interactions? Quite obviously Alice should not have default, unrestricted access to Bob's prior interactions without Bob's permission. Where Alice and Bob have engaged in a joint discussion, the question of access becomes more complicated. It would seem natural that both can have access to information about their prior discussions. However, if Bob is in a discussion with Eve, then access to prior discussion between Bob and Alice should again be restricted. The rules governing access will thus be complicated, often non-obvious and they will have many exceptions.

13.1.2.5. Interaction with a speech interface in the presence of others

In the early days of mobile phones, a common *faux pas* was to speak loudly on the phone in public places such as on a bus or subway. It seems that it causes an uncomfortable feeling to people when they overhear private discussions. It can also be hard to ignore speech when you hear it. Obviously, the reverse is also true, participants of a private discussion often feel uncomfortable if they fear that outsiders can hear their discussion.

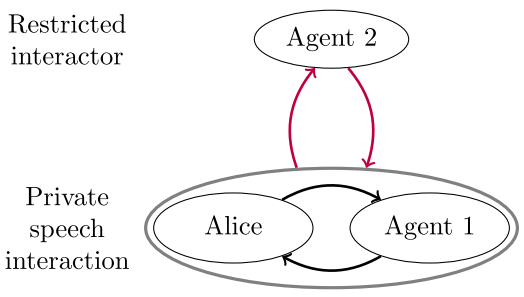


The same applies to speech operated devices. Such interactions can be private, but even when they are not, the fact that they can be overheard is often uncomfortable to all parties.

This is a problem when designing user interfaces to services. Speech interaction is often a natural way to user a service or device, but it is not practical in locations where other people can overhear private information and where it the sound is annoying to other people present.

13.1.2.6. Interaction with a speech interface connected to other services

When interacting with a speech interface, users typically do it with a specific objective in mind. For example, suppose Alice wants to turn off the lights in the bedroom and says "*Computer, lights off*". To which extent is it permissible that that information is relayed to a cloud-service? If the local device is unable to or uncapable of deciphering the command, it can transmit it to the cloud. The cloud-service then obtains information from a very private part of Alice's life.

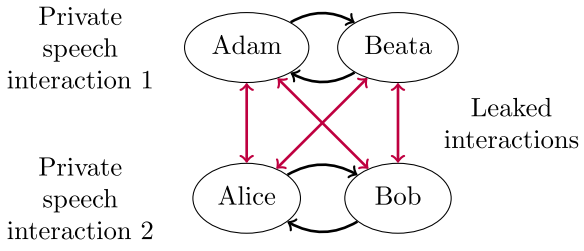


Information obtained this way can be very useful for example, to advertisers. By analyzing the habits of users, they can serve more meaningful advertisements. Arguably, by better targeting of users, advertisement can more effective, which could *potentially* reduce the need for advertisement. It is however questionable whether advertisers ever would have incentives to *reduce* the amount of advertisement. Still, some people are creeped out by "*overly fitting*" advertisement.

There are however plenty of other scenarios which are more potent sources of danger. What if insurance agencies analyze users life patterns and increase payments for at-risk users such as substance abusers? Some smart devices already today can call the emergency services if they recognize cries of help or other obvious signs of distress. What are the moral dilemmas of that?

13.1.3. Multi-user and multi-device scenarios

Things get even more complicated when multipled users and/or users co-exist in the same space. Consider, for example, an open office with two users simulatenously engaged in independent video conferences.



13.2. Information contained in speech signals

Speech signals contain a wide variety of information which are often potentially private or even sensitive and it is difficult or impossible to list all categories of potential information. However, information which speech at least contains includes for example;

- The linguistic (text) content of a speech signal can contain almost anything
  - If you reveal private information about yourself in a conversation, then clearly that information is contained in a transcription of the conversation.
  - Word-choices and manners of speaking can reveal things about the speaker, without the speaker realizing it himself or herself.
- Para-linguistic (i.e. non-linguistic viz. non-text) content has a wide variety of information:
  - Speaker identity
  - Physical traits, including gender, body size and age
  - Emotional state (happy, sad, angry, excited etc.)
  - Speaking style (public, intimate, theatrical etc.)
  - State of health, such as flu, mental health and diseases like Alzheimer's, but also including tiredness and intoxication.
  - Association and affiliation with reference groups, including groups of gender-identity, ethnicity, culture background, geographic background, nationality, political and religious affiliations etc.
- In interaction with other speakers:
  - Proof that you have met with the other speaker
  - Level of familiarity, intimacy and trust.
  - Power structures (leader/follower/partner)
  - Family, romantic and other relationships

In other words, speech contains or can contain just about all types of private and sensitive information you could imagine. As speech is a tool for communication, this is not surprising; anything we can communicate about, can be spoken. Conversely, if we find that (and we do find that) privacy is important, then speech is among the most important signals to protect.

13.3. Types of privacy

In a more general scope than just speech, privacy can be categorized into seven types: [Finn *et al.*, 2013]

- **Privacy of the person**, which refers to the privacy with respect to our physical body as well as any information about our body such as fingerprints, voice characteristics and medical history.
- **Privacy of behaviour and action**, refers to privacy with respect to what we do; for example, nobody needs to know what I do within the confine of my home.
- **Privacy of communication**, is particularly important in speech communication and refers to privacy of the content and meta-content of communication. That is, it is not only about the literal content of a communication, but also about the style of communication and also about the fact that communication has happened.
- **Privacy of data, including images and sound**, which extends the privacy of possessions to immaterial things. In particular, this addresses privacy with regard to sharing information about a third person.
- **Privacy of thoughts and feelings**, is paraphrasing, the privacy of thinking. We have the right to share our thoughts and emotions with whom we like, or to choose not to share our feelings with anyone. This does not mean that people would have to listen to you, but that your are allowed to offer them your thoughts.
- **Privacy of location and space**, has become increasingly important, since so many of our mobile devices has the ability to track your location. However, this type of privacy covers both your location and location history (as in location tracking).
- **Privacy of association**, refers to the privacy of whether you belong or to which extent you otherwise associate yourself to a particular group (religious, political, ethnic, gender identity, professional, interest groups etc.).

Note that this list does not make any claims with respect to *rights* to these types of privacy, but that privacy-issues can be often be split into these sub-topics. Whether someone has a right to privacy is a society-level decision and political choice, where psychological and cultural aspects play a big role.

13.4. Threat and attack scenarios

Threats to privacy in speech communication can almost always be defined as covert extraction of information as well as storage, processing and usage of that information in ways of which the speaker is not aware, and/or with which the speaker does not agree. Variability in scenarios is then almost entirely due to the type of information involved as well as the stakeholders. In particular,

- Companies can extract information from private users for unethical advantage
  - Insurance policies and mortgages can be denied based on covertly extracted information
  - The price of services can be increased for vulnerable groups
  - Access to information (search results) can be restricted and information can be targeted (advertisement) to covertly influence users for unethical advantage. Such behaviour is often advertised under the pretence of customizing services to the users preferences, but without giving the user any tools for choosing how services are customized.
- State operators can use surveillance and access restrictions on their own citizens
  - Authoritative regimes can track and eavesdrop the political opposition, dissent and other groups such as religious, ethnic and sexual orientations.
  - Also legitimate uses by police for surveillance in criminal investigations
- State operators can use surveillance on foreign citizens
  - Spies can use speech technology for (remote) eavesdropping and information extraction

- Criminals can steal information and use it for their advantage
  - Identity theft
  - Paparazzi's can steal private information of famous people
  - Private information can potentially be used for extortion
  - Explicit content could be sold as entertainment
- Private persons can covertly use speech technology for eavesdropping and extracting information of other persons
  - For example, the command history of smart speakers can give access to past commands of all users, also when speech commands have been made in private.

13.5. Privacy and security scandals

Most of the threats and attack scenarios are not familiar to the common public and some of them might be too abstract to be relevant to average users. Typically, we can hypothesize that scenarios which do not touch directly on the life of an individual, probably do not get much attention in the media. Topics in privacy and security which however have received attention in the public media include:

- Amazon workers are listening to what you tell Alexa (Bloomberg, 2019). Later it was revealed that Google, Apple and others are doing the same.
- Amazon Sends 1,700 Alexa Voice Recordings to a Random Person (Threatpost, 2018).
- Amazon's Alexa recorded private conversation and sent it to random contact (The Guardian, 2018).
- Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case (The Wall Street Journal, 2019).

Note that the fact that many of the above examples are related to Amazon/Alexa is probably more coincidence than an indication that Alexa would treat privacy differently than its competitors.

13.6. Approaches for safeguarding privacy in and improving usability of speech technology

13.6.1. Basic design concepts

At least from the European perspective, the following design concepts are seen as basis of good design for privacy. In fact, they are mandated by the General Data Protection Regulations of the European Union.

- *Privacy by default*; systems should always default to the least invasive configuration. Additional services, which are more invasive, can be chosen (opt-in) if the user so chooses.
- *Privacy by design*; privacy should not be an after-thought but systems should be designed for privacy. The overall systems structure should be chosen such that it supports privacy.
- *Data minimization*; data can be extracted from users only the extent explicitly required by the system. For example, if you order a pizza through a web-portal, you need to tell which pizza you want and where it should be delivered. Otherwise you cannot receive the service. In other words, all services require some level of information transfer, but the services cannot ask about information which is irrelevant to the service. For example, the pizza service cannot ask you about your gender-identity. Data minimization cab also be interpreted as *data ecology*, which would underline the fact that "more data" usually means also a higher consumption of energy and other resources. In fact, data could (or should) be treated as a natural resource itself. This line of argumentation

connects privacy with the UN sustainability goals.
An important aspect of data minimization is that information should stored only as long as it is necessary for the service.

- *Informed and meaningful consent*; when the user chooses a service, he should receive information about its implications to privacy in an understandable and accessible way, and the service provider should receive the users consent without any form of coercion.

Typically it is reasonable that service providers have access to aggregated data such as ensemble averages, but not to information about individuals. For example, in a hypothetical case, a smart speaker operator could receive the information that 55% of users are male, but would not get the gender of any individual user.

13.6.2. Definitions

A central problematic in privacy with speech signals is the concept of “uniquely identifiable”. Legal frameworks such as the GDPR state that private information is such data where individual users are “uniquely identifiable”, but there is no accurate definition of what it really means. If your partner recognizes your “Hello” on the phone, it means that for her, your “Hello” is uniquely identifiable. However, if you give 10.000 speech samples to your partner, one of which is your “Hello”, then there’s a significant likelihood that your partner would not find your “Hello” from the pile. An unanswered question is thus, “What is the size of the group where a user should be uniquely identifiable?”.

A more detailed aspect is that of significance. The speaker recognition approach is to find the most likely speaker, out of the reference group of size N, whereas speaker verification tries to determine whether we, within some confidence intervals, can be sure that you are who you claim you are. In engineering terms, this means that we want to find the speaker with the highest likelihood, but with a sufficient margin to all other speakers. In the opposite direction, we can also use a lower threshold; we could say that statistically significant correlation already exposes the users privacy. For example, if we find that the speaker is either you or your father/mother, then we have a significant statistical correlation, but your are not uniquely identified.

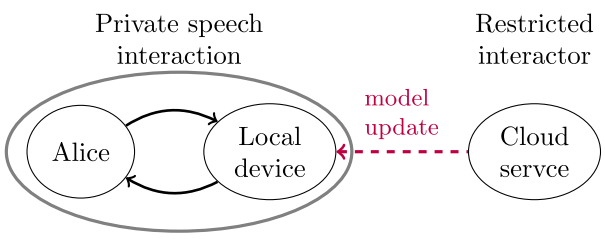
A further consideration is that of adjoining data; Suppose there is a recording of a speaker A, and that you happen to know a speaker A very well. Then it will be easy for you to recognize the voice of A in that recording. That is, you have a lot of experience (stored data) about how A sounds, therefore it is easy for you to identify A. Does that mean that A is uniquely identifiable in that recording? After all, A would not be identifiable if you did not know A (= if you would not have prior, stored data about A).
A slight variation of the above case is a recording of a speaker B, where B is relatively famous public person, such that there are readily available sound samples of his voice on-line. Does that make the recording of B uniquely identifiable? Or if there is a recording of a currently non-famous person C, who later becomes famous. Does that change the status of the recording of C to uniquely identifiable?

Today, this question remains unanswered and we have no commonly agreed interpretation of what “uniquely identifiable” really means. What level of statistical confidence is assumed? What level of adjoining data is assumed (in terms of GDPR probably: any and all data which exists)? Can it change over time if new information becomes public (probably: yes)? Can it change over time if new technologies are developed (probably: yes)?

13.6.3. Local/edge processing

Privacy is an issue only if some other party has access to data about you. Data which resides on a device which is in your control is therefore relatively safe, assuming that no outsider has access to that device. If data is sent to a cloud server then there are more entities which could potentially have access to your data. Therefore all storage and processing which can be done

on your local device is usually by design more private than any cloud server. Typically, a cloud server would only provide software updates (downlink), but no data would be sent in the other direction (uplink). [Shi *et al.*, 2016]



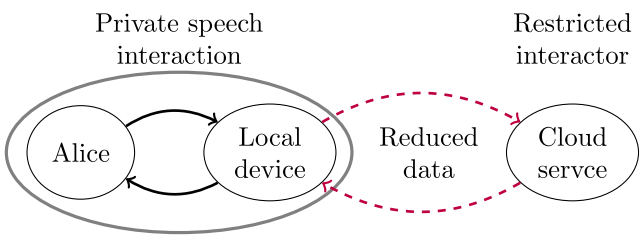
Observe that this does not protect you from other local users. For example, if multiple persons are using one smart speaker at home, then the other users could have access to information about you through that device and any connected other devices.

Central limitations of edge processing are

- Many services require outside access; say if you ask your phone "What's the weather tomorrow?", the phone cannot know that by itself, but has to retrieve the information from a cloud server. The essential content of your speech is therefore relayed to the cloud and you don't have much benefit from edge processing.
- Improving voice operated services requires a lot of data. Moreover, data which reflects features of actual users is much better than any simulations. Service providers thus argue that they need to collect data from users to provide high-quality services, and local processing could prevent the services providers from getting that data.
- Edge devices would use their full capacity only when they pick up speech in their microphone, which means that most of the time, edge devices would lie dormant, waiting for speech commands. This is a wasteful use of resources; a cloud server can better balance the load because with a large number of users, the resource requirements would likely be more stable.

13.6.4. Anonymization, pseudonymization and disentanglement

A central issue with voice communication is that, in addition to the intended message, it also contains so much other information. For example, if you want to order a pizza delivery, the service provider needs to know only the content of the order, destination where the order should be delivered and how it is paid. The provider does not need to know, for example, your state of health or your cultural affiliation. *Anonymization* refers to methods which try to strip away such private and extra information such that only the intended message remains. With *pseudonymization* we refer to similar methods, where private information is replaced by some other information. For example, we could replace personal identifying information of a user *Alice* with an avatar-identity *Adam*. The process and methodology of separating the different streams of information is known as *disentanglement*.

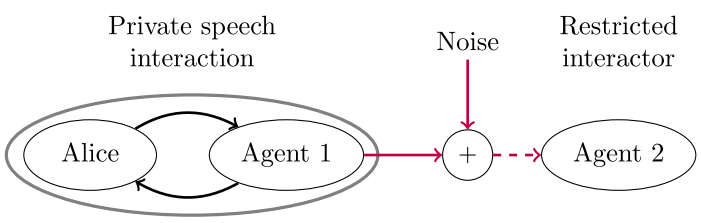


An issue with current methodology is that there are no theoretical guarantees of anonymity. That is, we can try to deduce private information from the anonymized data stream, and if we fail, we can say that the anonymization has succeeded with respect our attempts to break it. However, we have no guarantee that some other more advanced method for breaking the anonynization would not succeed.

13.6.5. Differential privacy

Even when operating with aggregate data, like the mean user age, it is still possible to extract private information in some scenarios. For example, if we know the mean user age and the number of users at a time t, and we also know that the age of user X was added to the mean

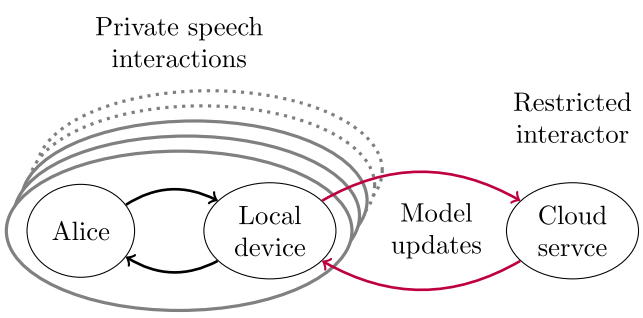
at time $t+1$, as well as the mean user age at $t+1$, then we can deduce the age of user X with basic algebra. As a safeguard against such differential attacks, to provide differential privacy, it is possible to add noise to any data transfers. Individual data points are then obfuscated and cannot be exactly recovered. However, the ensemble average can still be deduced if the distribution of the added noise is known.



The required compromise here is that the level of privacy corresponds to amount of noise, which is inversely proportional to the accuracy of the ensemble mean. That is, if the amount of noise is large, then we need a huge number of users to determine an accurate ensemble average. On the other hand, if the amount of noise is small, then we can get a fair guess of an individual data point, but also the ensemble average is accurate.

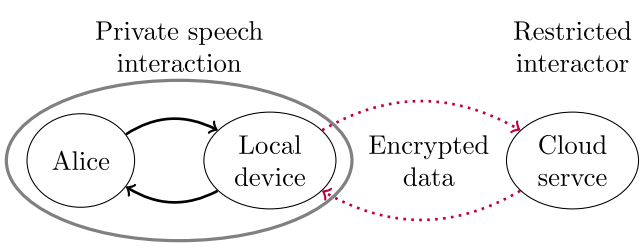
13.6.6. Federated learning

To enable machine learning in the cloud without the need to provide access to private data, we can use federated learning, where private data remains on the local device, but only model updates are sent to the cloud. Clearly this is approach has better privacy than one where all private data is sent to the cloud. At the same time, by combining information from a large number of local devices in the cloud, the machine learning models can be optimized to be highly accurate. However, currently we do not yet have clear understanding of the extent of privacy with this type of methods; some data is sent to the cloud, but can some private data still be traced back to the user?



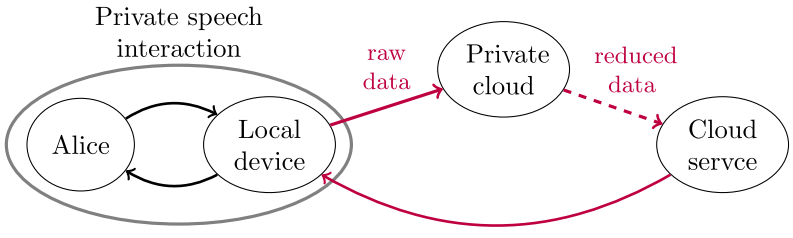
13.6.7. Homomorphic encryption

Suppose a service provider has a proprietary model, say an analysis method for Alzheimer's disease from the voice, and your doctor would like to analyse your voice with that method. Naturally your voice is also private, so you do not want to send your voice to the third-party service provider, but also the service provider does not want to send the model to you. Homomorphic encryption provides a method for applying the secret model on *encrypted* data, such that you have to only send your data in an encrypted form to the service provider. Your doctor would then receive only the final diagnosis, but not the model nor your speech data. The concept is in principle beautiful, it solves the problem of mutual distrust very nicely. However, the compromise is that currently available homomorphic encryption methods require that all processing functions can be written as polynomial functions. In theory, we can transform any function to a corresponding polynomial, but the increase in complexity is often dramatic. Consequently, privacy-preserving methods based on homomorphic encryption typically have a prohibitively high computational complexity. [Armknecht *et al.*, 2015]



13.6.8. myData

In addition to privacy-preserving algorithms, we can also design privacy-preserving architectures. The myData paradigm is based on a three-tier design, where the user can choose where all his/her data is stored and where the user can give access for service providers to his/her data when required. The idea is to separate service providers from data storage, such that users have better control over his/her data. To transform existing services to adhere with the myData concept requires that new storage services for private data are created and that APIs between storage and processing services are specified. [Poikola and end Harri Honko, 2015]



Note that, if a user chooses to store private data on a cloud-server, then it is still susceptible for abuse by the storage-service-provider, unless appropriate encryption methods are used. However, the user could in principle choose to store private data on a edge device, such that the storage-service-provider is cut out of the loop.

A further risk is that in the myData concept, we usually assume that data is stored at a single central location, which becomes a central point of weakness. Should someone gain illegitimate access to the storage, then all your data would be compromised. Distributing data to several different storage locations might therefore be reasonable.

13.7. Design goals, human computer interfaces and user experience

A common prejudice is that privacy and security requirements cause problems for developers and make systems more difficult for users to use. Such prejudice are unfortunate and patently misguided. The problem is that many privacy problems are not visible to the casual observer and their effects become apparent only when it already is too late. Another argument is "*privacy is not my concern because I haven't seen any privacy problems*", which is like saying that "*rape is not my concern because I haven't seen any rapes*". This is thus an absurd argument. Privacy safeguards are meant to protect users and developers from very bad consequences. These problems are real. *You* cannot ignore them.

However, designing for privacy is also *not only* about protection of users. It is also very much about designing technology which is *easy to use* and where the user experience feels intuitive and natural. For example, speaker recognition can be used to grant access to voice technology such that the user does not have to be bothered with passwords, PIN-codes or other cumbersome authentication methods. Overall speech technology promises to give access to services without the need scroll through menus on your washing machine to find that one mode which is optimized for white curtains made out of cotton.

The overall design goal could be that people should be able to *trust* the system. In particular, a trustworthy system will be [Chen and Dhillon, 2003, Xie and Peng, 2009]

- *consistent*; It does what it says, and it says what it does.
- *competent*; It is able to do what it should do and what it says it does.
- *benevolent*; It cares for the user and it shows that in both *what* is says and *how* it says it.

These goals are best illustrated by examples;

- We should avoid cognitive dissonance; if the computer has the intellectual capacity corresponding to a 3-year-old child, but if it then speaks with the authoritative voice of a middle-aged person, it is giving mixed signals which can confuse users. Similarly, if a

computer leaks out all your information to the world, while speaking with an intimate voice, it is also giving the wrong impression to the user. Conversely, the intuitive impression which a device gives should be consistent with its true nature.
- Consistency is extremely important and one mistake can be very costly; If your friend Greg once tells about your intimate health-incident to your friends, it casts a shadow of doubt over all future and all of the past 10 years. Did he already earlier tell my secrets to them? Who else has he told my secrets to? It takes a very long time to patch such breaches of trust.

13.8. Ethical dilemmas

The following is a list of hypothetical questions which can (and do) arise in the design of speech operated systems:

- If a device hears cries of help, should it call the police automatically, even when it breaches privacy and the trust of the user?
- If a device hears indications of domestic abuse, but has no direct evidence, should it call the police? What level of evidence is required?
- If a device or service recognizes that you have Alzheimer's or some other serious illness, should it tell it to you? Even if your doctor would not yet have noticed it? Even if the diagnosis might be incorrect? Even if you might choose not to want to hear it? Even if you'd have a history of depression and such bad news might trigger suicidal thoughts?
- Suppose you have been dreaming of buying a new bicycle, but haven't told anyone. Your smart TV, though, knows about it because you have searched for information about that bicycle. Suppose that your spouse is simultaneously trying to come up with a nice surprise birthday present. Should your smart TV suggest to your spouse that she buys the bicycle?
- Suppose your local cafeteria has automatic speech operated ordering of drinks. On this day, last year, you bought a birthday-surprise drink from this cafeteria. Should the computer remember that and congratulate you today about your birthday?
- The better your services know you, the better they could serve you. That's undoubtedly a fact. (The fact that current services are not optimal is not a contradiction.) However, do you want to have privacy from devices in the same way you have privacy from your friends. For example, your friends do not follow you to the toilet or the bedroom; is it ok if your device does that?

13.9. Security and privacy in speech research

Scientific research is based on arguments supported by evidence, where evidence, in the speech sciences, is recordings of speech. Access to speech data is therefore a mandatory part of research in the speech sciences. To obtain trustworthy results, independent researchers have to be able to verify each others results, which means that they have to have access to the same or practically identical data sources. Shared data is the gold standard for reproducible research. However, the sharing of speech data can be problematic with respect to speaker privacy.

The concept of "uniquely identifiable" is here the key. If an individual is *not* uniquely identifiable in a data set, then you are allowed to share that data. Conversely, if you remove all identifying data, then you can share data relatively freely. However, in perspective of the discussions above, it should be clear that it is not clear what constitutes identifying data nor is it clear what makes that data "uniquely" identifying.

A second important consideration is *consent*. The persons whose voices are recorded must be allowed to choose freely whether they want to participate and that choice has to be explicit; you need to ask them clearly whether they want to participate in a recording. The

research needs to be able to prove that consent has been given, and therefore that consent must be documented carefully. Consent must also be given freely such that there are no explicit or hidden penalties of rejecting consent. Furthermore, if any uniquely identifying data of a participant is stored, then the participant must be allowed to withdraw consent afterwards. There are however some important exemptions to this rule; the right to withdraw consent can be rejected, for example, if that would corrupt the integrity of the data set, such as

- If withdrawal of a participant could bias the results, then that *could* be grounds for denying withdrawal. For example, if a dataset is constructed in a way that it represents a balanced subset of the population (the amount of say, males and females, different age, cultural and education backgrounds are chosen to match the general population), then we cannot remove any participants without corrupting the distribution. Moreover, people more educated in questions of privacy could hypothetically be more prone to withdrawing their consent, such that the population becomes biased.
- If withdrawal of a participant could jeopardize the reproducibility of results, then that *could* be grounds for denying withdrawal. For example, if a machine learning algorithm is used on a dataset, then we can recreate that algorithm only if we have *exactly* the same data available. This is especially problematic if the dataset is relatively small, where small changes in the dataset can have big consequences on the output.

To allow plausible grounds for denying the right to withdraw consent, datasets can then be designed to be either balanced or relatively small. Collecting balanced datasets is good practice in any case, such that this is not a limitation but can actually improve quality. Conversely, good data is balanced and that should be our goal; A consequence is that we *might be forced* to deny the right to withdraw consent. Avoiding the collection of excessively large data sets is also good from the perspective of *data minimization* and data ecology.

In a request for consent, the data collector should state the purpose of the dataset (i.e. *purpose binding*). For instance, a dataset could be collected for development of wake-word detection methods and consent is received for that purpose. Then *it is not permissible* to use the same data set for speaker detection experiments or medical analysis of the voice. Period. It is therefore good practice to ask for consent in a sufficiently wide way, such that researchers have some flexibility in using the data. Blanket consent to all research purposes is however not good practice. In particular, it is recommend that processing of sensitive information such as health, ethnic, political information is excluded if it is not the express purpose of the dataset (cf. data minimization).

If a dataset by nature does include uniquely identifiable data, then the researchers need to apply stronger layers of safeguards. In particular, typically researchers have to keep track of who has access to the data, to ensure purpose binding and to allow withdrawal of consent. This could also require that any researcher who downloads the data signs a contract with the data provider, where the terms of usage are defined. Such a contract can be required in any case, not only with uniquely identifiable data.

Data such as medical information, data about children or other exposed groups, political, religious and gender-identity affiliations etc. are particularly *sensitive*. If your dataset contains *any* such information, then you have to apply stronger safeguards. To begin with, access to such data has to be, in practice, always limited to only persons who are included in a legally binding contract specifying access rights and allowable uses, processing and storage.

As an overall principle, note that the principal investigator (research group leader) is legally responsible for the use of the data that is collected, stored and processed. In particular, if a third party downloads the data and misuses it, for example by analysing health information even if no consent has been acquired for that purpose, then it is the principal investigator who is responsible. However, the principal investigator is only required to apply *reasonable safeguards* to ensure that data is not misused. What level of safeguards are sufficient has however not yet been agreed. It is likely that there will never be rules which specify exactly a sufficient level of safeguards.

In the above discussion it has become clear that the nature of *unique identifiability* can change over time, when new information is published and new technologies emerge. This means that datasets which previously were adequately protected, over time become exposed to privacy problems. It is therefore important that researchers monitor their published datasets over time such that if new threats emerge, they can take appropriate action. For example, they could withdraw an dataset entirely. Reasonable ways for implementing this could be:

- *Expiry date*; All datasets should have a clearly stated shelf-life and use of the dataset after the expiry date should be prohibited. The manager of the dataset could update the expiry date if no new threats are discovered. Academic publications based on expired datasets should not be accepted.
- *Controlled access to datasets*; To enforce purpose binding and to enable withdrawal of datasets, the data manager can require that all users are registered and sign a formal contract which specifies accepted uses.

As a last resort, when data is so sensitive and private that it cannot be publicly released, it is possible to require on-site processing of data. For example, you can design a computing architecture, where data resides on a secure server, to which researcher have access through a secure API. Data never leaves the server such that privacy is always preserved. For an even higher level of security, data can be stored on an air-gapped computer system, which means that access to the data requires that researchers physically come to the computer (no network access). This level of security is usually the domain of military-grade systems.

13.10. References

- ISCA Special Interest Group “Security and Privacy in Speech Communication”, https://www.spsc-sig.org/

[ABC+15] Frederik Armknecht, Colin Boyd, Christopher Carr, Kristian Gjøsteen, Angela Jäschke, Christian A Reuter, and Martin Strand. A guide to fully homomorphic encryption. *IACR Cryptology ePrint Archive*, 2015:1192, 2015. URL: https://ia.cr/2015/1192.

[CD03] S.C. Chen and G.S. Dhillon. Interpreting dimensions of consumer trust in e-commerce. *Information Technology and Management*, 4:303–318, 2003. doi:https://doi.org/10.1023/A:102296263124.

[FWF13] Rachel L Finn, David Wright, and Michael Friedewald. Seven types of privacy. In *European data protection: coming of age*, pages 3–32. Springer, 2013. URL: https://doi.org/10.1007/978-94-007-5170-5_1.

[Gas18] Jessica Gasiorek. *Message processing: The science of creating understanding*. UH Mānoa Outreach College, 2018. URL: http://pressbooks-dev.oer.hawaii.edu/messageprocessing/.

[KonigSS+15] Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H Robert, and others. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124, 2015. URL: https://doi.org/10.1016/j.dadm.2014.11.012.

[Lar19] Xabier Lareo. Smart speakers and virtual assistants. *TechDispatch #1:*, 2019. URL: https://data.europa.eu/doi/10.2804/004275.

[NJK+19] Andreas Nautsch, Catherine Jasserand, Els Kindt, Massimiliano Todisco, Isabel Trancoso, and Nicholas Evans. The gdpr & speech data: reflections of legal and technology communities, first steps towards a common understanding. *arXiv preprint arXiv:1907.03458*, 2019. URL: https://doi.org/10.21437/Interspeech.2019-2647.

[Pet02] Sandra Petronio. *Boundaries of privacy: Dialectics of disclosure*. Suny Press, 2002.

[PeHH15] 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.

[SCZ+16] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: vision and challenges. *IEEE internet of things journal*, 3(5):637–646, 2016. URL: https://doi.org/10.1109/JIOT.2016.2579198.

[XP09] Yi Xie and Siqing Peng. How to repair customer trust after negative publicity: the roles of competence, integrity, benevolence, and forgiveness. *Psychology & Marketing*, 26(7):572–589, 2009. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mar.20289, doi:10.1002/mar.20289.

By Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Mariem Bouafif Mansali, Daniel Ramos