How can private information recorded by voice-enabled systems be identified?

Álvaro Moretón, Ariadna Jaramillo

Report

This part of the EDPL hosts reports in which our correspondents keep readers abreast of various national data protection developments in Europe, as well as on the most recent questions in different privacy policy areas. The Reports are organised in cooperation with the Institute of European Media Law (EMR) in Saarbrücken (www.emr-sb.de) of which the Reports Editor Mark D. Cole is Director for Academic Affairs. If you are interested in contributing or would like to comment, please contact him at mark.cole@uni.lu.

Practitioners Corner

How can Private Information Recorded by Voice-enabled Systems be Identified?

Alvaro Moretón and Ariadna Jaramillo*

I. Introduction

People are keener to use voice-enabled technologies than ever before. Whether it is to improve employees' productivity in corporate environments or to perform random queries from smartphones, voice technologies are revolutionising user experience around the world. Voice assistants can facilitate everyday tasks, but they raise serious privacy concerns, particularly by the way these technologies are trained to complete users' requests. The different elements of the user's speech (message content, voice signal), together with background sounds can be used to extract personal information from the speaker at the moment of his/her interaction with the voice-enabled system, which conflicts with privacy aspects.

This report provides an overview of how personal data recorded by voice-enabled systems can be identified through the categorisation of personal information and the analysis of the context, and how these methods can be used to design private-by-design voice-based solutions that intend to neutralise personal data and information/words that may reveal private information.

II. Voice Technologies

Currently, voice technologies rely strongly on deep learning, which has led to major improvements in speech-to-text, spoken language understanding, and dialogue management. These technologies operate as cloud-based services. The user's speech is sent to the cloud, where it is automatically transcribed and processed, and the system's reply is sent back to the user's device. This reply can be rendered orally employing spoken language generation followed by text-to-speech and/or through different modalities (eg visual display, tactile feedback, etc.), depending on the use case.

In order to expand the range of languages and application cases offered by these technologies, it is necessary to collect massive amounts of training data for each language and each application domain. While certain voice technology companies hire paid speakers to do so, this process is expensive and results in out-of-domain data hence suboptimal machine learning performance. Therefore, many companies collect in-domain data from their users instead. Although the users have given their consent – although the question of validity of the consent in light of the re-

* Alvaro Moretón is the corresponding author and Project manager at Rooter Analysis SL in Madrid; for correspondence alvaro.moreton@rooter.es; Ariadna Jaramillo is project intern with Rooter.
The work described in this report was partly supported by the European Union's Horizon 2020 Research and Innovation Program, under Grant Agreement No. 825081 COMPRISE (<https://www.compriseh2020.eu/>). Emmanuel Vincent (Senior Research Scientist and Project Coordinator) and Marc Tomassi (Professor in Computer Science at Lille University), both part of the COMPRISE project, contributed to this article by providing comments and feedback in their areas of expertise.

quirement of informed consent is not further discussed in this report – when accepting the terms and conditions of the voice-enabled application, this raises serious privacy concerns, as the company or a third-party attacker that would get access to the stored speech data can access personal information about the user revealed by this.[1]

III. The Concept of Personal Data

Before moving into categorisation and contextualisation *per se*, we recall the concept of personal data. Article 4 (1) of the GDPR defines personal data as 'any information relating to an identified or identifiable natural person'. Moreover, it states that 'an identifiable natural person is one who can be identified, directly or indirectly, by reference of an identifier, an online identifier or one or more factor specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person'. Moreover, Article 9 (1) provides an exhaustive list on special categories of personal data (racial or ethnic origin, political opinion, religious or philosophical beliefs, or trade union membership, genetic and biometric data, data concerning health, and data concerning a natural person's sex life or sexual orientations).

The Article 29 Working Party (WP29, now the European Data Protection Board established under the GDPR) referred to the concept as it was contained in the preceding Data Protection Directive as well in its Opinion 4/2007. According to this Opinion, by including the phrase 'any information' in the definition, the legislator called for a broad definition of personal data.[2]

Identification of the data subject is commonly achieved through particular pieces of information called identifiers. The extent to which certain identifiers are enough to achieve identification depends on the context of each situation. Most of the time, however, identifying personal data is not an easy task. In this sense, categorisation of personal data could be the key to facilitate the identification of personal data contained in user speech.

To begin with, the GDPR does not provide a proper categorisation of personal data. Instead, Article 4(1) exemplifies possible categories 'such as a name, an identification number, location data, an online identifier or [...] one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person'.

However, Article 9(1) GDPR provides the abovementioned closed list of special categories of personal data which are considered particularly sensitive due to the risk they pose to the fundamental rights and freedoms of data subjects. It is the only article in the GDPR that provides an exhaustive categorisation of personal data.

Despite the lack of an appropriate categorisation in the GDPR, it is possible to extract some categories from it, taking into consideration Opinion 4/2007 of the WP29, too. Such categories of data relate to private and family life, working relations, types of activity that are undertaken by the individual, economic behaviour, social behaviour, living traits, appearance of the person, among others.

IV. Personal Data Categorisation

The following subsections analyse the sources from which voice-enabled devices collect information, as well as the types of personal information that can be collected from each of them. Furthermore, this section intends to analyse how the combination of information from various sources can reveal details about the context in which the speaker is interacting with the device.

1. Speech Content

Speech recognition-based technologies (speech-to-text) recognise speech and convert it into readable form or text. In this regard, spoken messages often carry private information that may be collected and processed, such as:

– Words or utterances that explicitly mention the user's identity, general traits related to the speaker's background (age, nationality, etc.), informa-

1 Centre for Data Ethics and Innovation (CDEI), 'Smart Speakers and Voice Assistants' (Centre for Data Ethics and Innovation, September 2019), <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831180/Snapshot_Paper_-_Smart_Speakers_and_Voice_Assistants.pdf> accessed 10 December 2019.

2 Opinion 4/2007 on the concept of personal data, WP 136, available at <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf> 4.

tion relating to the user's health status, or otherwise critical information (eg. credit card number, home address, etc.)
- Information not revealed by the dialogue outcome (eg. user preferences revealed by asking the system about similar products or a general category of products before settling on one)
- Private information about other persons (eg age of a user's family member)

Identifying personal data within the speech content can be a difficult task. Categorising personal data could facilitate the identification of words and phrases that could potentially contain personal information (eg. it is likely that the phrase 'my credit card number is [...]' triggers the inclusion of a credit card number, which is considered to be personal data). These types of phrases could be included in specific category of data (eg. a 'financial information' or 'account information' category). Some other examples of possible categories of personal information (including words and phrases that may trigger the inclusion of personal information) are listed below:
- Racial and ethnic (eg Indo-European, Caucasic, English, Spanish)
- Behavioural (eg 'I usually visit the _____website', 'I buy through_____', 'I practice_____, 'I usually_____')
- Demographic (eg 'My age is ____', single, married, divorced, middle class, upper class, working class)
- Health (eg 'I've been prescribed _____', 'I suffer of_____', deaf, blind, physical disability, mental disability, pills, tablets, antibiotic, treatment)

Words or phrases that fall into any of the previous categories would be considered personal data only if they can be related to an identified or identifiable person. In the context of self-contained data processing in a speech-enabling applicationn, one single word cannot necessarily be linked to an identified or identifiable person. However, different pieces of information collected together can lead to the identification of a particular person and, therefore, be considered as personal data.

2. Voice Signal

Personal information could also be extracted from the speaker's voice signal, including, among others, general traits, and physical, emotional and health state.

The list below contains examples of information that could be revealed through the speaker's voice signal[3]:
- General traits of the speaker (eg gender, age, ethnic origin, etc.)
- Mental state (eg stress, relaxation, depression, etc.)
- Physical health state (eg bronchitis, smoking habits, intoxication)

3. Background Sounds

Depending on its recording capacity and the usage mode (close-talk vs hands-free), voice-based devices might be capable of collecting background sounds, which may provide private information about the user and be employed for different purposes (eg profiling for marketing purposes).

Some examples of background sounds that could be revealing personal information about the speaker interacting with the device[4] could be:
- Background conversations (eg between the speaker's family members)
- Background sounds (eg a TV program, traffic, the radio)

In many cases, background sounds alone may not be sufficient to provide private information about the user interacting with the voice-enabled device. However, combined with additional elements such as the user's voice and/or the content of the speech message itself, the probability of private information being disclosed increases.

V. Contextualisation

Contextualisation focuses on the circumstances surrounding the user at any given moment. The concept must not be confused with personalisation, which in-

3 Ranya Aloufi, Hamed Haddadi, David Boyle, 'Emotionless: Privacy-Preserving Speech Analysis for Voice Assistants' (2019) < https://arxiv.org/pdf/1908.03632.pdf> accessed 12 December 2019

4 Bob Siegel, 'How voice recognition will affect privacy in the Internet of Things' (CSO, 14 Nov 2016) <https://www.csoonline.com/article/3140633/how-voice-recognition-will-affect-privacy-and-the-internet-of-things.html> accessed 12 December 2019

volves learning individual users' traits, preferences or other personal information over time based on his/her past experiences.[5] For example, following a personalisation approach, if the user is continuously searching for animated films being played at cinemas, the voice-enabled system is very likely to recommend him/her upcoming films based on this preference. By contrast, contextualisation concerns the information collected by the voice-enabled system about the user (eg is he /she going alone to the movies?) at the moment he/she is interacting with the system exclusively (which is limited regarding space and time).

Contextualisation could be used to improve services and functionalities provided by voice-enabled systems, which, in turn, can lead to an increase in the amount of personal information revealed as it allows for a more precise interpretation of the information given during the interaction.

Understanding the specific context in which a voice-enabled device is used (eg. by whom, app type/nature, purpose) facilitates the task of identifying more accurately which private or sensitive information is more likely to be collected and processed in each specific context.

This report analyses three different types of contexts: usage context, cultural and linguistic context, and context provided by the combination of different elements. Furthermore, it focuses on the contextualisation that occurs when the user interacts with the voice-enabled system only.

1. Usage Context

Most voice devices and voice-based apps are used for concrete or limited purposes and in specific sectors or contexts. Being able to determine this information would facilitate the detection of private or sensitive information more likely to be revealed in each specific context, and to improve privacy functionalities. Some examples are:

- In a legal environment, personal information that could be revealed to the voice-based solution is more likely to be related to client's identities, criminal records, financial situation, etc.
- In a medical environment, personal information that could be revealed to the voice-based solution is more likely to be related to the identification and health status of the user.

2. Cultural and Linguistic Context

The analysis of the user's cultural and linguistic background can be useful for identifying words or pieces of information that may be private or sensitive as well. Depending on the region in which the voice-enabled system is being used or the origin of the speaker, one single word can contain personal or very sensitive information or, in contrast, provide completely neutral (or less sensitive) information. For example:

- IRA: In British English, it might stand for Irish Republican Army, in American English for Individual Retirement Account.
- A carry-on: for the British, this could mean having a love affair, for the Americans, this is a luggage that can be carried aboard an aircraft, bus, or train.

Another critical aspect to be considered related to the linguistic context is the terminology associated with a field or area of activity, as well as the jargon or slang vocabulary. For example, the term 'FX' may have no meaning outside the medical activity but, within it, could be revealing a health condition (bone fracture).

3. Information Revealed by the Combination of Speech Elements and Background Sounds

The combination of all different elements of the user's speech (message content, voice signal), together with background sounds, may increase the amount of information provided by one or a few words alone considerably, and provide additional information on the context in which the speaker is interacting with the voice-enabled system.

a. Context Provided by the Message as a Whole

The first source of information used to identify the meaning behind one or a few words and to assess whether private or sensitive information is being revealed in the message in which these words are integrated, is the context of the whole message. By

5 Nuance Communications, 'Personalization & Contextualization: Empowering content domains with Artificial Intelligence (AI)' (Nuance, 9 May 2017) <https://www.nuance.com/content/dam/nuance/en_us/collateral/mobile/automotive/white-paper/wp-personalization-en-us.pdf> accessed 13 December 2019

analysing the content of the message as a whole, words can be contextualised individually. This knowledge facilitates the task of identifying whether the meaning of a word has changed given the content in which it is being employed.

For example, the phrase 'find an Indian restaurant' could provide information about the preferences of the user, which is not considered particularly sensitive. However, the phrase 'find a gluten-free restaurant' would be providing much more sensitive information as it may concern the speaker's health.

However, it is challenging to identify loose or siloed words as personal information (due to the different meanings they may have, the dependency on the context, etc.). One possible approach to overcome this issue could be to identify short phrases that may lead to personal information.

b. Context Provided by the Voice

Together with the content of the message, the voice itself can provide important information about the speaker state (personal information) or a given situation (context).

The following example shows how voice tone may help contextualise the message provided by the speaker and extract additional information about its meaning:

– The phrase 'find me a place to buy my pills' together with a voice tone suggesting a depressive state or nervousness could be providing sensitive information about the speaker's mental health.

Voice tone could also be used to determine how the speaker feels regarding a specific situation revealed in the message, and his/her thoughts on concrete action, a place, a product, etc. For example:

– If the speaker is dictating a voice note containing the sentence 'Auditors will be really surprised with such a good job made by the accounting manager', the recipient may be immediately aware of the irony behind the message because he or she already knows the context behind it (eg. the accounting manager is a disaster). On the contrary, people unaware of the situation in the accounting department may discover this just by the ironic tone of the speaker.

Lastly, voice tone could also reveal information about the situation the speaker is living when interacting with the voice-enabled device. The following situation exemplifies this possibility:

– The phrase 'I need a doctor, where is the closest hospital?' together with a concerned or nervous tone, could reveal that the speaker is living a serious or life-threatening situation (eg. a domestic accident).

c. Context Provided by Background Sounds

Background sounds may provide additional information about the context the speaker is in when interacting with the voice-enabled system (eg his/her location). The following situation exemplifies this possibility:

– The phrase 'recommend me a playlist for travelling' together with the sound of an airport megaphone may reveal that the speaker was in the airport ready to travel at the moment of the interaction. Sometimes, background sound could even reveal which airport (eg. an audible welcome message from loudspeakers).

Background noise could also reveal information about a given situation experienced by the speaker:

– If the speaker asks the device to call the police or to send an officer to his/her address, and people yelling or a glass break can be heard in the background, this could provide information to assess a possible case of domestic violence.

VI. Towards Privacy Preservation

To offer a privacy-driven alternative to the classical workflow, research projects such as COMPRISE (Cost-effective, Multilingual, Privacy-Driven voice-enabled Services), which has received funding from the European Union H2020 program, are aiming to deliver fully privacy-by-design methodology and tools to protect the users of voice-enabled systems. In this respect, the COMPRISE training branch aims to collect large-scale in-domain speech and language data for multiple languages and application domains and learn domain-specific personalised models from this data for speech-to-text, spoken language understanding, and dialogue management in a privacy-preserving way. For this purpose, it relies on four research advances:

– The users' speech and the transcribed text will be transformed into 'neutral' data, in which the pri-

vate information has been deleted or replaced in a provable way before being sent to the cloud. To properly develop this system, it must be capable of identifying words that could be considered as personal data and categorisation is therefore key.
- The neutral speech and text data will be weakly labeled by means of multiple automatic labelers in order to perform weakly supervised learning. As neutral data are continuously being collected from the users, only a small portion of that data will be human-labeled.
- User-independent speech and dialogue models will be trained in the cloud to leverage the large amounts of data required to achieve state-of-the-art performance, and personalised to each user on the user's device.

Furthermore, other privacy-preserving speech and language technologies and methodologies have been developed and implemented, such as[6]
- Partially/fully homomorphic encryption (HE): an encryption technique that allows the performing of calculations on encrypted data without having to decrypt it
- Secure two-party computation (STPC): allows calculations without revealing intermediate information. To do this, two parties jointly operate a computable function on two inputs, without mutually revealing the value of their respective inputs
- Searchable encryption: enables a server to search on a database of encrypted documents without learning the client's secret keys
- Functional encryption (FE): lets an authorized entity evaluates the value of an encrypted function on encrypted data and obtain the result in clear
- Hardware-assisted security: trusted execution environment, (eg. Intel SGX)

VII. Conclusions

After analysing how all different types of content, either alone or in combination, could reveal personal and sensitive information about the speaker, it can be concluded that contextualisation can be used for disambiguation between private and non-private information. Based on the fact that context itself supports personal data, the text could argue that one privacy driven way to perform such a disambiguation task is to follow a workflow similar to the one in the COMPRISE project described before.

The above leads to consider the possibility of:
- Neutralising the speaker's voice, so no traits (eg. general, physical, mental, etc.) could be identified through it.
- Suppressing background sounds and noises so no information on the speaker situation at the moment of the interaction could be revealed (eg location).
- Suppressing words and phrases that could lead to revealing personal information.

The analysis of context may help identify which word or phrases containing personal information should be neutralised. However, voice-enabled systems' users are diverse, and privacy levels preferred may vary from one person to another and from one context to another.

For example, in medical environments, the use of contextualization could facilitate the creation of a privacy driven-voice enabled app that focuses on words or expressions more likely to be used in this particular context, and that may reveal personal information (eg health-related words/expressions). Again, the level of privacy desired will depend on the needs of the system's users. In corporate environments, contextualization could allow the use of voice-enabled systems to increase employee's productivity (eg. replacing typing by dictation) without compromising sensitive information on the company's activities, as devices would be capable not only of suppressing word and phrases containing this information but also background sounds and conversation that could reveal information not intended by the employees.

Another example would be that a private user may not care for his or her information to be collected in a domestic environment, as long as his or her kids are not present. On the contrary, as just mentioned, a voice-enabled system in a corporate environment might be preferred to have strict privacy levels to avoid sensitive information of the company to be recorded. Contextualization could facilitate the detection of private or personal information, allowing users to select the levels of privacy they desire. Research will have to continue in that direction.

6 Andreas Nautsch et al. 'Preserving Privacy in Speaker and Speech Characterisation' (2019) <https://www.sciencedirect.com/science/article/pii/S0885230818303875> accessed 13 December 2019