Speaker anonymization solution for public voice-assistant interactions –
Presentation of a Work in Progress Development

Ingo Siegert

1 Mobile Dialog Systems, Institute for Information Technology and Communications,
Otto von Guericke University Magdeburg, Germany

https://www.dhmd.de/ausstellungen/ab-611-kuenstliche-intelligenz/

Abstract

The use of voice assistants has rapidly grown in recent years. They can be found in millions of households. And a lot of effort has been made by researchers to improve the usage of these systems. One issue that remains open is the usage of voice assistants and recording of interactions for research purposes in public environments due to privacy concerns. Additionally, data collections, offering unconstrained, unscripted public interactions are quite rare and mainly only focus on transcribed content or have focused on private usage, short pre-defined tasks, or specific domains. The current paper presents an approach on how voice data recordings of user interactions with voice assistants in a public space can be recorded and processed in conformity with the GDPR.

Index Terms: speech assistant, public recordings, GDPR

1. Introduction

In recent years, the market for commercial voice assistants has been continuously rising, resulting in a nearly doubled user-base in the last three years among the U.S. adults [1, 2] and an increasing number of owners using voice assistants in their daily routine [3]. The reason is mainly the given naturalness of speaking as a form of communication in contrast to the use of additional external periphery, and on top of that, the independence from additional training based on the single application. In practical effect, the usage of a mobile phone does not differ substantially from the control of a smart home application with intelligent voice control.

But for many (public) applications it has to be stated that it is still unclear how users used this technology, and its given freedom, and overcome certain pitfalls. Especially, as evidence on the use of voice assistants is scarce, often anecdotal, and usually focuses on private uses and contexts. For this purpose, larger interaction studies of public, unconstrained interactions are needed. But, recordings in public environments, in combination with commercial voice assistants, are a privacy risk, due to the danger of traceable voiceprints. In this paper, a possible solution for a GDPR compliant voice assistant interaction in a public environment is presented.

2. Constraints due to the Setup

As part of a science exhibition on artificial intelligence at the Deutsches Hygiene-Museum Dresden (German Hygiene Museum in Dresden)¹, an exhibit on the limits of the interaction with modern voice assistants was prepared by the Otto von Guericke University Magdeburg. This exhibit is designed as a quiz in which questions of different difficulty levels are displayed on a touch screen. To answer these questions, visitors are assisted by a commercial voice assistant (Amazon Alexa). The setting is similar to the use of “telephone jokers” from well-known TV shows. The voice assistant has no knowledge of the quiz and the current question. The visitor must activate it independently. To demonstrate the limits of actual voice assistants, the questions are formulated in such a way that the voice assistant cannot answer them directly and the visitor must therefore either rephrase the questions or ask partial questions. In addition to the benefit for the visitor, however, the added value should also be created for science by storing the voice interactions with the voice assistant in order to be able to study how users interact with voice assistants and, above all, how certain pitfalls in the interaction are solved. However, this setup is problematic from a data protection perspective in several places: 1) unintended transmission of voice data (accidental activation), 2) unintended recording of background voices, 3) transmission of voice recordings, and 4) storage of visitors voice recordings, and 5) preventing users’ identification via their speech content.

3. Discussion of possible solutions

3.1. Prevention of Unintended Activation

It has already been shown that voice assistants can be activated unintentionally [4, 5]: Either by saying words phonetically similar to the activation word or because, in a public place, visitors are not aware of being next to a voice assistant and using the activation word with a different intention. As a consequence, these unintentional activations lead to voice utterances being recorded and transmitted for processing. Therefore, an unintentional activation must be prevented.

The presented approach uses an “activate-by-button” solution. The visitor wanting to interact with the voice assistant has to press a button to activate the voice assistant. To ease the interaction, pressing the button will automatically play a pre-recorded awake-word, as visitors in pre-tests interpret the button press as activation and did not utter the wake-word.

3.2. Prevention of Unintended Background Recording

Modern voice assistants use a microphone array in which the surroundings are recorded in addition to the speaker. By applying beam-forming and speech enhancement algorithms it is then ensured that the area in which the speaker is located (when activated) is acoustically amplified [6, 7]. Unfortunately, these algorithms do not completely suppress background conversations. Thus these background conversations may be an audible

¹Due to the judgement “Schrems II” the adequacy of the EU-US data protection shield (Privacy Shield) is invalid.
part of the voice recording transmitted to the external servers during the processing of the request. This circumstance is due to the intended use of the voice assistants in private households, so that it must be assumed that optimization has been made for the acoustic environment of a household rather than for a much broader and acoustically different museum area or similar [8]. As many background voices are to be expected in a public exhibition, it has to be secured that in the intended setting the recording of background noise is avoided. For this purpose, the actual microphone array of the voice assistant with omni-directional directivity is replaced by a single microphone with the narrowest possible lobe-like directivity. Pre-tests with data recordings under similar room-acoustics have showed that state-of-the-art ASR systems are able to work with heavily noised speech data [9, 10].

3.3. Transmission of Speaker’s Voice Recordings

The voice of a user represents personal information, which can also be seen as biometric information (cf. Art 4(14) GDPR), as it allows the speaker to be identified in large data collections [11, 12]. Although it is possible to withdraw consent to the use of one’s own data by Amazon for the purpose of quality improvement, the processing on American servers cannot be completely ruled out and different legal opinions are to be applied to the transmission of personal voice data, which is to be treated particularly sensitively in a public environment. Hence, the voice of the speaker needs to be altered so that no conclusions can be drawn in regard to a particular speaker. Therefore, the actual solution relies on speaker anonymization. However, the application scenario poses difficulties. From the experience with a similar previous study ([13]), it can be concluded that women and men, but also children and elderly people interact with the voice assistant. Therefore, the anonymization must be able to produce on the one hand an intelligible anonymized version for many different voice types and their fundamental frequencies but on the other hand without exhaustive training, as the speakers are not known before and should use the system without any further voice-tuning. Thus, simple pitch shifters and vocoders are rejected. Furthermore, it was also shown in ([13]) that the speech content varies a lot with also quite complex clauses and that state-of-the-art speech recognizers already have difficulties here (cf. [9]). Thus, voice conversion algorithms relying on a correct ASR output are also ruled out [14]. Therefore, it was decided to use the McAdams algorithm [15]. The basic idea of McAdams is to apply a slight contraction/expansion to the poles derived from linear predictive coding (LPC) coefficients of speech content on a frame-by-frame basis, which leads to a transformation of the related formants and thus to a change of the voice impression. A further advantage is that this algorithm can be applied to the transmission of personal voice data, which is to be anonymized before the transmission of the request. This circumstance is due to the intended use of the voice assistants in private households, so that it must be assumed that optimization has been made for the acoustic environment of a household rather than for a much broader and acoustically different museum area or similar [8].

The presented solution presents a first approach to use voice assistants in a public environment that cannot be fully pre-defined in terms of users’ voice characteristics and speech content complexity. The interaction with a voice assistant has to be activated consciously and the voice data is anonymized before the transmission. To which extent the anonymized voice recordings are still recognized by the voice assistant and how intuitive the usage is will be evaluated continuously during the exhibition.

4. Preliminary Evaluation

A preliminary evaluation is currently conducted, first results will be shortly presented in the final paper. Hereby the intelligibility, the ASR performance, and the processing run-time within the final setup will be analyzed.

5. Conclusions and Outlook

The presented solution presents a first approach to use voice assistants in a public environment that cannot be fully pre-defined in terms of users’ voice characteristics and speech content complexity. The interaction with a voice assistant has to be activated consciously and the voice data is anonymized before the transmission. To which extent the anonymized voice recordings are still recognized by the voice assistant and how intuitive the usage is will be evaluated continuously during the exhibition.

6. Acknowledgements

The author of the paper would like to thank Andreas Nautsch for his hint to the McAdams anonymization technique and the Deutsches Hygiene Museum in Dresden for hosting the scientific exhibit and supporting the research on public voice-assistant interactions. This research has been funded by the Federal Ministry of Education and Research of Germany in the project Eronymous (project number S21060A).
7. References


