Poster: Smart Speaker Privacy Control - Acoustic Tagging for Personal Voice Assistants

Peng Cheng, Ibrahim Ethem Bagci
Lancaster University
Lancaster, United Kingdom
{p.cheng2, i.bagci}@lancaster.ac.uk

Jeff Yan
Linköping University
Linköping, Sweden
jeff.yan@liu.se

Utz Roedig
University College Cork
Cork, Ireland
u.roedig@cs.ucc.ie

I. Introduction

Siri, Amazon Echo, Google Home and the like are now commonplace Personal Voice Assistants (PVAs). They are integrated in mobile phones (Siri, Cortana), consumer electronics such as TVs (SkyQ) and are also used as stand-alone devices (Amazon Echo, Google Home). PVAs are sometimes also referred to as Smart Speakers or Voice Controllable System (VCS). PVAs continuously monitor conversations and may transport conversation elements to a cloud back end where speech is stored, processed and maybe even passed on to other services.

A user has currently little control over how her conversations are treated. Not all PVAs are owned or managed by the user, and she is normally not in control of back-end systems and has no influence over how the services exchange conversation recordings. For example, when meeting people the user can switch off her own phone-based PVA but cannot control PVAs of others.

We argue that users desire more control on how their conversations are processed by PVAs. We propose to embed additional information (referred to as *tag*) into acoustic signals which can then be interpreted by the systems to implement security and privacy requirements of involved parties.

Many methods to generate acoustic tags exist, ranging from a simple signal overlay to a hidden acoustic watermark [1], which in turn are suitable for different application scenarios. For example, a simple acoustic tag can be employed by users to signal that they have given no consent to recording, processing and distribution of conversations recorded in their presence. A cooperating PVA back end looking out for such tags may then not process the recorded audio to honor the wishes of individuals. An acoustic watermark hidden within a recorded audio sample may be used by individuals to identify the origin of recorded speech at a later stage; it might give individuals an opportunity to keep track of recordings they have never agreed to. In such a scenario, cooperation of the PVA back end is not necessary.

Besides the design of a tag and its usage, there is also the question of how the acoustic tag is generated. A device is needed to generate the tagging signal; a likely candidate is a mobile phone with a suitable app. As it is not efficient to continuously transmit tag information, it must be determined when to emit a tag signal. This can be solved by having a tagging device listening for the same wake words as the PVA. Finally, as multiple users may want to tag, collisions must be avoided and a tagging protocol must be established.

This work explores the aforementioned design space of acoustic tagging for PVAs. Specifically we investigate:

- *Tagging Applications:* We give a description of application scenarios in which acoustic tagging can address user privacy and security concerns.
- *Tagging Signals and Protocols:* We provide a classification of tagging options and describe protocols for embedding tags of multiple users.
- *Tagging Evaluation:* We provide an evaluation of the signal path for simple overlay tagging using Google Home Mini. We show that tagging signals in the range between 4kHz and around 7.2kHz are usable.
- *Tagging Prototype:* We describe our prototype tagging device based on PocketSphinx [2] and an evaluation of the system. The prototype shows that tagging can be used to signal non-consent in public spaces.

II. Personal Voice Assistant (PVA)

The operation cycle of a PVA, shown in Figure 1, consists of two phases: *activation phase* and *recognition phase*.

In the activation phase the PVA waits for a user to activate voice recognition. In light of practicality, most systems utilize a wake word mechanism. The wake words may be speaker-dependent or speaker-independent [3].

On activation the PVA enters the recognition phase. In most scenarios, the PVAs streams the audio signals following the wake word to a back end for analysis. A response might be sent to the local device or another action may be triggered.

The captured audio streams are stored by PVA providers, and the storage duration and the specific usage of the data is not clearly articulated [4]–[6].

III. Application Scenarios

Acoustic tagging in the context of PVAs can be used for a number of security and privacy related application scenarios.

This poster is associated with the paper with the same title accepted by SafeThings 2019.

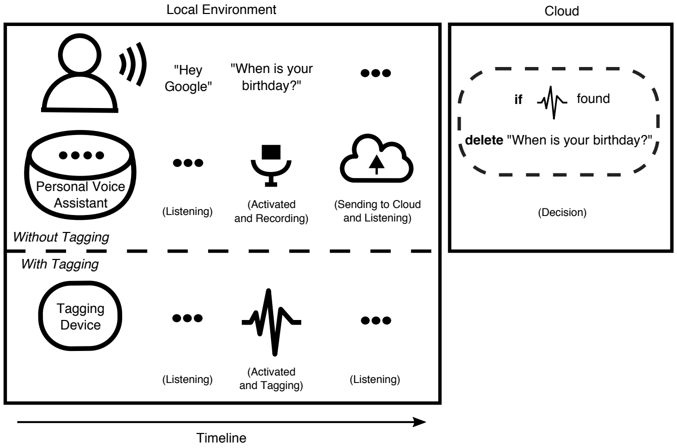

Fig. 1: The workflow of a personal voice assistant, without or with a tagging device.

Signalling Recording Consent: People generally object to conversations being recorded without their given consent. An acoustic tag will be emitted by users who give no recording consent. Any PVA system detecting a tag could then refrain from processing or even recording a conversation.

Recording Identification: It is reasonable to assume that conversations are recorded (by accident or on purpose) without consent by nearby PVAs. Such recordings may later be used and it might be desirable to identify the context (e.g., location, time, participants) of the conversation.

Data Trading: PVAs store conversation recordings on back-end systems. This data is an asset and the service providers employ it to improve their offerings. A provider may tag samples in order to control further distribution or to simply mark the sample source.

IV. Tagging Options

Audible Tag: A tag is embedded and its presence is clearly audible, e.g., in the form of audible noise.

Unnoticeable Tag: The tag is added to the audio signal such that its presence is not noticeable to a human.

Inaudible Tag: This approach is similar to the unnoticeable tag. The tag is added to the audio signal such that it cannot be perceived by a human.

Hidden Tag: The tag is added to the audio signal such that it cannot be perceived by the user. In addition, it cannot be determined by other tools (e.g., spectrum analyzer, frequency analysis) that a tag is embedded in the signal.

V. Tagging Analysis

We use a common PVA, the Google Home Mini, to evaluate tagging performance. The aim is to determine the usable tagging frequency range and to evaluate tag signal distortion.

We use the software Audacity to evaluate the MP3 recording and find it to be a stereo, 16kHz MP3 format. The audio signal passes through a low-pass filter which attenuates frequency elements higher than 8 kHz. Due to practical non-ideal low-pass filters, the attenuation will also affect frequencies just below 8 kHz.

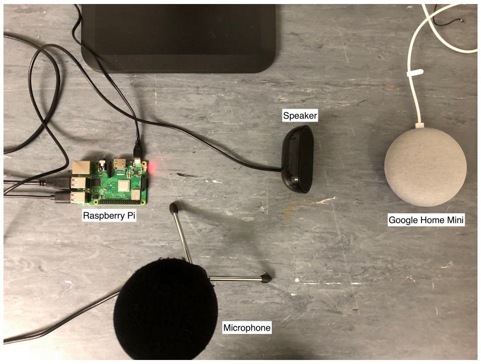

Fig. 2: Experiment Setup.

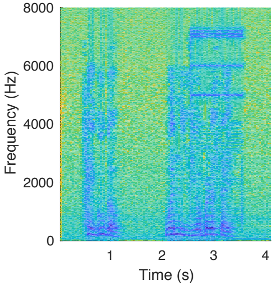

Fig. 3: The spectrogram of the downloaded signal resulting from the prototype tagging system

The sampling frequency of the audio encoding is 16kHz, which means ideally all of the audio contents below 8kHz should be retained. However, only the audio contents below 7.2kHz remains, and we assume this may result from the unavoidable imperfection of the filter design.

VI. A Prototype Tagging System

As the hardware platform we select a Raspberry Pi 3 Model B+ with a simple USB microphone and a commodity speaker.

To evaluate the tagging device we use the experiment setup shown in Fig 2. A Google Home Mini is used as the PVA and the tagging device with speaker and microphone are placed next to it.

The tag, an audible multi-tone signal, lasts one second. We then speak the sentence "*Hey Google, when is your birthday?*" to test the system. The wake word is recognized by the PVA and as well as the tagging device which emits the tag signal. Thereafter we use Google's Myactivity website to download the recording. Figure 3 is the spectrogram of the audio file representing the whole experiment.

References

[1] R. Jain, M. C. Trivedi, and S. Tiwari, "Digital audio watermarking: A survey," in *Advances in Computer and Computational Sciences*. Springer, 2018, pp. 433–443.

[2] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proc. ICASSP'06*, 2006.

[3] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "DolphinAttack: Inaudible Voice Commands," in *Proc. CCS'17*, 2017.

[4] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity," in *Proc. SenSys'18*, 2018.

[5] "Apple stores your voice data for two years," https://goo.gl/6hx1kh, 2013.

[6] "Google stores your voice input," https://goo.gl/7w5We1, 2017.