Provable Consent for Voice User Interfaces

Stephan Sigg
Communications & Networking
Aalto University
Espoo, Finland
stephan.sigg@aalto.fi

Le Ngu Nguyen
Communications & Networking
Aalto University
Espoo, Finland
le.ngu.nguyen@aalto.fi

Pablo Perez Zarazaga
Signal Processing and Audio
Aalto University
Espoo, Finland
pablo.perezzarazaga@aalto.fi

Tom Backstrom
Signal Processing and Audio
Aalto University
Espoo, Finland
tom.backstrom@aalto.fi

***Abstract*—The proliferation of acoustic human-computer interaction raises privacy concerns since it allows Voice User Interfaces (VUI) to overhear human speech and to analyze and share content of overheard conversation in cloud datacenters and with third parties. This process is non-transparent regarding when and which audio is recorded, the reach of the speech recording, the information extracted from a recording and the purpose for which it is used. To return control over the use of audio content to the individual who generated it, we promote intuitive privacy for VUIs, featuring a lightweight consent mechanism as well as means of secure verification (proof of consent) for any recorded piece of audio. In particular, through audio fingerprinting and fuzzy cryptography, we establish a trust zone, whose area is implicitly controlled by voice loudness with respect to environmental noise (Signal-to-Noise Ratio (SNR)). Secure keys are exchanged to verify consent on the use of an audio sequence via digital signatures. We performed experiments with different levels of human voice, corresponding to various trust situations (e.g. whispering and group discussion). A second scenario was investigated in which a VUI outside of the trust zone could not obtain the shared secret key.**

***Index Terms*—Privacy, audio processing, device pairing, fuzzy cryptography**

I. Introduction

We have witnessed the spread of voice user interfaces (VUIs) embedded in smartphones and digital assistants such as Siri, Alexa, and Cortana. These systems are able to participate in conversational social activities [1]. VUIs can provide an acoustical front-end for voice-based services [2]. It is common, that significant part of the processing and analysis of the audio as well as interpretation and reasoning on its content is conducted in a remote cloud operated by the VUI manufacturer. When a specific VUI is recording, information on which audio it shares, as well as the further use of the information content is intransparent and not under the control of the individual generating the audio. It is further not possible for an individual to, even temporarily, opt out of the audio recording and sharing. In addition, the recording is usually conducted without the consent of the individuals generating the audio. Still worse, audio recording and sharing is done without even notifying the individuals being overheard.

In this paper, we address these issues by proposing a technical solution that would empower individuals to establish

- an adaptive trust-zone based on audio SNR
- verifiable trust relationships and consent on audio recording and use

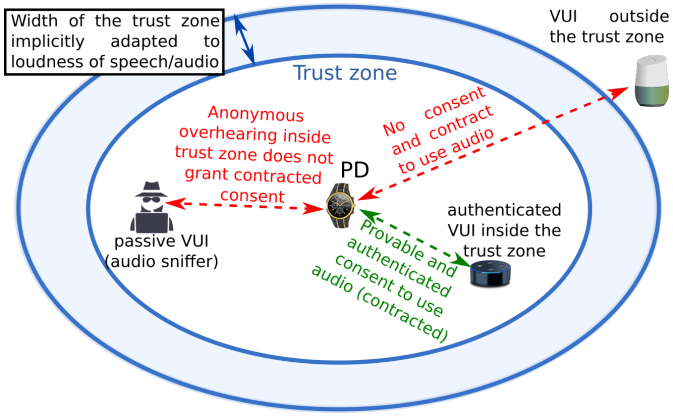

Fig. 1. Concept of our proposed scheme. Only VUIs in the variable-width trust zone are eligible for authenticated consent on audio use. Non-trusted or invisible/passive VUIs are not granted consent by the Personal Device (PD)

In a nutshell (see figure 1), we generate audio fingerprints at VUIs in proximity. These fingerprints are mapped to trust-space-representative patterns via fuzzy cryptography. Only VUIs in the same trust space are able to generate identical patterns from the overheard audio. In addition, due to properties of the fuzzy commitment protocol, a VUI has to disclose its presence in order to generate the representative pattern.

Only VUIs inside the trust-zone, that disclose their presence and receive consent are able to compute the representative pattern. These patterns are then used to verify presence in the trust space while exchanging keys to sign the recorded audio via Diffie-Hellman key exchange. We performed experiments to prove that our method can prevent a system outside the trust zone from obtaining the secret key.

II. Related Work

Speech recognition by mobile VUIs has been studied intensively over the last decades [1]. Through advanced signal processing [3], noise filters [4], processing of recorded audio [5], and language-specific models [6], impressive speech-recognition accuracy has been achieved [7], both in speaker-dependent [8] and speaker-independent [9] approaches. This success has spurred a proliferation of speech-audio interfaces in all kinds of commercial devices in recent years [10].

Often, audio recording and processing is conducted in a remote data center, non-transparent for the subjects monitored

by these devices. For an individual it is not possible to control the audio recording and sharing of privacy-relevant data.

In this work, we propose a mechanism with which individuals can re-gain control over their data in that each individual could independently control when and with which devices audio information shall be shared by providing explicit, verifiable consent that is secured by digital signature.

In a nutshell, we utilize the concept of audio-based secure VUI pairing [11] to distribute private keys among devices within a trust zone, which are in turn utilized for pseudo-authenticated[1] Diffie-Hellman key exchange [12]. Via this key exchange, VUIs agree on a shared secret key of a public key cryptography system [13] which is utilized to digitally sign (and thereby prove consent and presence in the trust zone) recorded audio. With this protocol in place, any party can thus later verify that an audio sequence was recorded with consent of the subject being observed.

III. Methodology and Protocol

The individual steps in our methodology are briefly sketched in figure 2. We intend to empower an individual to

1) implicitly control the width of a trust zone by lowering or raising her voice such that all VUIs outside of the trust zone are incapable of establishing a secret key that is required to generate a digital signature to verify consent on the use of speech signals
2) force disclosure of presence and potential authentication of VUIs as a precondition to sharing this secure key
3) allow verification of consented use of an audio sequence

To achieve these goals, we assume that the individual wears a Personal Device (PD) which serves as the central component in the protocol to grant consent and to generate and distribute secure keys. This item can, for instance, be an audio-capable smart watch or a smartphone worn by the individual. A trust zone is created, surrounding the PD via audio-based distributed key generation as it is described in [14]. In particular, similarity in audio-fingerprints (e.g. [15], [16]) is exploited to define the boundaries of the trust zone: VUIs that are able to generate a fingerprint with bit-difference within the error correction threshold t of an error correcting code (e.g. [17], [18], are considered inside the trust zone and share a master secret k_{MS}. Utilizing Diffie-Hellman key exchange, authenticated with the hashed secret $h(k_{MS})$, VUIs inside the trust zone agree on public and private keys k_{ts}^{+}, k_{ts}^{-} to sign recorded audio. The respective steps are detailed in the following.

A. Trust Zone

To create audio-fingerprints, following the process in [15], we split an audio sequence S with length $|S| = l$ and sample rate r up into n frames $F_1, \ldots, F_n$ of identical length $d = |F_i| = r \cdot \frac{l}{n}$. On each frame a discrete Fourier transformation (DFT) weighted by a Hanning window (HW) is applied:

$$\forall i \in \{0, \ldots, n-1\},$$
$$S_i = DFT\left(HW(F_i)\right)$$

[1] authentication/verification of presence presence within a trust zone

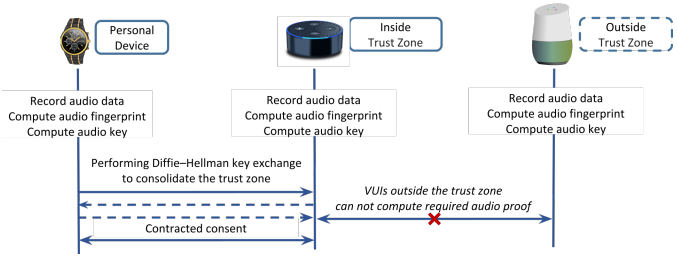

Fig. 2. Protocol for fingerprint generation and distinction between VUIs inside and outside of the trust zone. Affiliation with the trust zone is conditionedthe individual step on VUIs ability to compute the correct audio key.

The frames are divided into m non-overlapping frequency bands of width

$$b = \frac{\text{maxfreq}(S_i) - \text{minfreq}(S_i)}{m}. \tag{1}$$

On each band the sum of the energy values is calculated and stored to an energy matrix E with energy per frame per frequency band.

$$\forall j \in \{0, \ldots, m-1\},$$
$$S_{ij} = \text{bandfilter}_{b \cdot j, b \cdot (j+1)}(S_i) \tag{2}$$
$$E_{ij} = \sum_k S_{ij}[k] \tag{3}$$

Using the matrix E, a fingerprint f is generated, where $\forall i \in \{1, \ldots, n-1\}, \forall j \in \{0, \ldots, m-2\}$ each bit describes the difference between the energy on frequency bands between two consecutive frames:

$$f(i,j) = \begin{cases} 1, & (E(i,j) - E(i,j+1)) - \\ & (E(i-1,j) - E(i-1,j+1)) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

To generate a master key k_{MS} for a pair of fingerprints f_p and f_{tz} generated for the PD and any VUI inside the trust zone, we utilise Reed-Solomon $RS(q, m, n)$ codes, with $q = 2^k$, $k \in \mathbb{N}$ and $n < 2^k$ (with message and codespace as $\mathcal{A} = \mathbb{F}_q^m, \mathcal{C} = \mathbb{F}_q^n; q = p^k, p \text{ prime}, k \in \mathbb{N}$). First, the PD chooses a random secret $a \in \mathcal{A}$ which is then encoded following the Reed-Solomon scheme to a specific codeword c_p. It further computes $\delta = f_p \ominus c_p$ and shares δ publicly. All VUIs in the trust zone can then generate $c_{tz} = f_{tz} \oplus \delta$ and derive a_{tz} from c_{tz} following the Reed-Solomon scheme. IFF $sim(f_p, f_{tz}) = sim(c_p, c_{tz}) \leq t$ it follows that $a_p = a_{tz}$ due to the error correction.

Consequently, all VUIs within the trust zone share the same master key $k_{MS} = a_p$. This process is detailed in figure 2.

B. Authentication of VUIs in the trust zone

To authenticate VUIs in the trust zone, Diffie-Hellman authenticated key exchange is executed between the PD and any VUI within the trust zone to derive keys k_{ts}^{+}, k_{ts}^{-} of a public key cryptosystem. Presence in the trust zone is verified via the hash of the master key $h(k_{MS})$, exchanged over an encrypted channel using k_{ts}^{+}, k_{ts}^{-}. The PD will discard k_{ts}^{+}, k_{ts}^{-},

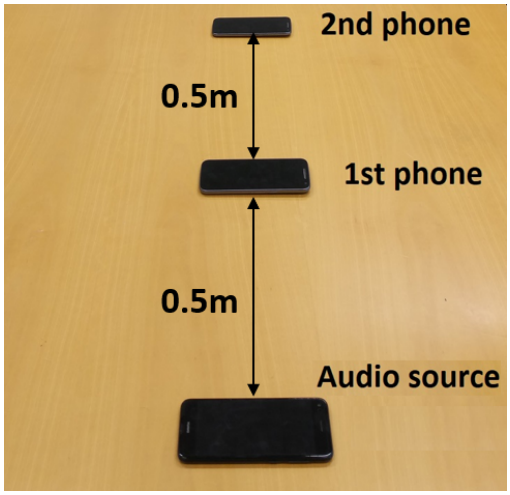

Fig. 3. Layout of VUIs in our experiment. The 1st phone represents the personal device (PD), which is close to the audio source, while the 2nd phone represents a VUI. The between the VUI and the PD is varied throughout the experiment to $\{1.0, 1.5, 2.0\}$ metres respectively

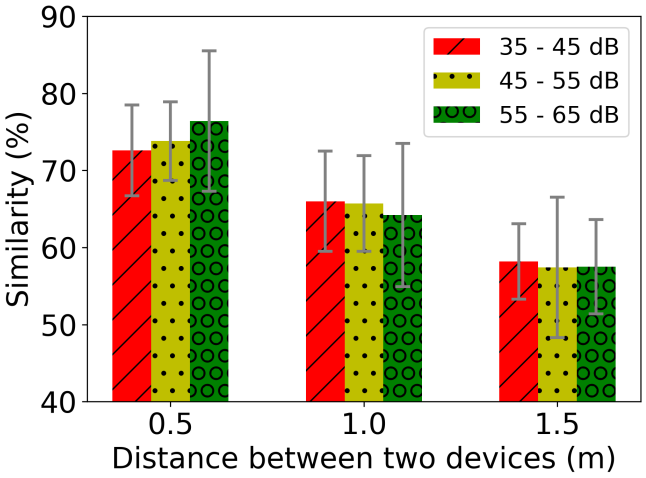

Fig. 4. Similarity of audio fingerprints with respect to distance between the PD and the VUI. The similarity decreases with increasing distance and increasing audio loudness. With increasing distance, the impact of loudness diminishes.

should the authentication fail. In order to link k_{ts}^{+}, k_{ts}^{-} to the respective audio sequence, the PD will keep k_{MS}, δ and k_{ts}^{+} for verification. Note that, additionally, a trusted authority and certificates could be employed to establish authentication of VUIs.

C. Signing audio

Using the private key k_{ts}^{-}, the VUI in the trust zone will digitally sign all audio sequences published or shared with other devices (e.g. a cloud server). The signature proves that the audio was recorded by a VUI inside the trust zone and with ocnsent ofthe PD, since otherwise, k_{ts}^{-} would not be known to the VUI.

D. Proof of consent

Any party is able to verify that a specific signed audio sequence has been recorded and shared with the consent of the PD via the signature and the fingerprint of the audio sequence that should lead to k_{MS} when combined with δ from above.

IV. Experiments

We conducted two experiments to demonstrate the concept of the trust zone. In the first experiment, we consider VUIs at different distance to an audio source and in the second experiment, a VUI is placed outside the trust zone to demonstrate that this remote device is not capable of obtaining a sufficiently similar audio fingerprint in order to compute c_{ts}.

A. Impact of SNR on audio pairing

In this experiment, we simulate a scenario in a small room with multiple VUIs. The layout of the experiment is displayed in Figure 3. In order to generate a setting that can be repeated and verified, we utilized a phone broadcasting continuously recorded speech audio. Two phones (the VUIs) were placed in $d_1 = 0.5$ metres and $d_2 = \{1.0, 1.5, 2.0\}$ metres from the audio source (see Figure 3). This shall model a scenario with multiple VUIs scattered across a room. We controlled the sound level of the office (i.e. the audio source) in $35-45$, $45-55$, and $55-65$ dB, which correspond with verbal conversation loudness (i.e. an individual raising or lowering her voice to intuitively adapt the range of the trust zone). For all combinations of these settings, fingerprints have then be recorded according to the steps described in section III-A. The similarity of audio fingerprints, expressed by relative Hamming distance, is shown in Figure 4.

The similarity of audio fingerprints decreases when the distance increases. Hence, it is possible to configure a fuzzy cryptography scheme to allow only VUIs in a certain proximity (within the trust zone to share an audio-based key (see section III-A). Furthermore, the proposed approach can be intuitively controlled by verbal conversation loudness.

B. VUI outside the trust zone

In the second experiment, we investigate VUIs outside the trust zone. An example is a meeting room that defines the trust zone since audio significantly degrades outside the room. We assume a VUI inside the trust zone and another VUI outside of the room and trust zone. The setting is depicted in figure 5.

A VUI inside the trust zone is located 40cm from the PD and audio source. Another VUI is located 1.40 metres away from the PD in the same direction and outside the room in front of the opened door. Speech is clearly audible outside the room but the SNR is lower at that distance. After performing similar computation as above, we observe that within a distance of one metre, audio fingerprints achieve a similarity of 68.2% while the VUI outside of the secure zone achieved lower similarity (52.3%) and therefore did not obtain the correct master secret k_{MS}.

V. Conclusions

We have proposed a mechanism with which individuals can re-gain control over their audio data, in particular individual control when and with which devices audio information shall be shared by providing explicit, verifiable consent that is secured by digital signature.

We utilized the concept of audio-based secure VUI pairing to distribute private keys among devices within a trust zone, which are in turn utilized for authenticated Diffie-Hellman key exchange. Finally, VUIs agree on a shared secret which is utilized to digitally sign recorded audio, and thereby prove consent and presence in the trust zone. Any party can thus later verify that an audio sequence was recorded with consent of the subject being observed.

We envision that this protocol can be implemented and required by legislation so that the obligation to prove consented use of a piece of audio is passed to VUI device manufacturers. In particular, the approach is lightweight since it can be implemented irrespective of the VUI manufacturers cooperation and protocol used.

We performed experiments with different levels of human voice, corresponding to various trust situations (e.g. whispering and group discussion). A second scenario was investigated in which a VUI outside of the trust zone could not obtain the shared secret key.

VI. Future Work and Discussion

This paper demonstrated the general principle and feasibility to establish a provable, trust and consent based contract relation between individuals and VUIs in proximity. The proposed protocol places the individual in control of her produced audio content and, in particular, does not require technical compliance by VUI manufacturers but, in contrast, proper legislation. Indeed, with a VUI is not prevented to record and use audio without consent, but the solution presents a technical solution that does not allow a VUI to forge a proper signature to that would prove consent on the use of a particular audio sequence, since the private key k_{tz}^{-} used for the signature can only be obtained by computing a sufficiently similar fingerprint f_{tz} and after authentication towards the PD.

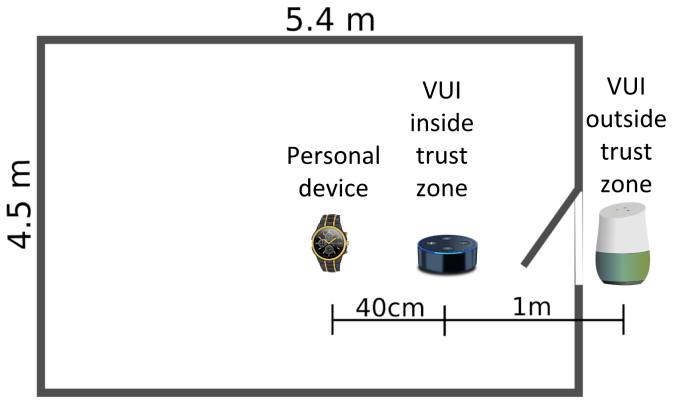

Fig. 5. Distinction between VUIs inside and outside of the trust zone. The protocol separates devices inside and outside the trust zone as VUIs outside the trust zone are incapable to compute a sufficiently similar audio fingerprint.

It should be mentioned though that the scheme also requires a trusted party or some decentralized trusted data structure (e.g. distributed ledger technology or other distributed commitment schemes) to maintain public information (k_{MS}, δ and k_{ts}^{+}) which is needed for later verification.

Future work is required to investigate the performance and scalability of such solutions that would support application for ubiquitously distributed VUIs.

References

[1] M. Porcheron, J. E. Fischer, S. Reeves, and S. Sharples, "Voice interfaces in everyday life," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: ACM, 2018, pp. 640:1–640:12.

[2] T. Bäckström, F. Ghido, and J. Fischer, "Blind recovery of perceptual models in distributed speech and audio coding," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 2483–2487.

[3] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6645–6649.

[4] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 5, pp. 49:1–49:28, Apr. 2018.

[5] S. Pouyanfar, Y. Yang, S.-C. Chen, M.-L. Shyu, and S. S. Iyengar, "Multimedia big data analytics: A survey," *ACM Comput. Surv.*, vol. 51, no. 1, pp. 10:1–10:34, Jan. 2018.

[6] K. Li, H. Xu, Y. Wang, D. Povey, and S. Khudanpur, "Recurrent neural network language model adaptation for conversational speech recognition," in *Proc. Interspeech 2018*, 2018, pp. 3373–3377.

[7] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19 143–19 165, 2019.

[8] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5934–5938.

[9] A. A. Allen, H. C. Shane, and R. W. Schlosser, "The echo™ as a speaker-independent speech recognition device to support children with autism: an exploratory study," *Advances in Neurodevelopmental Disorders*, vol. 2, no. 1, pp. 69–74, Mar 2018.

[10] M. B. Hoy, "Alexa, siri, cortana, and more: An introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018, pMID: 29327988.

[11] D. Schürmann and S. Sigg, "Secure communication based on ambient audio," *IEEE Transactions on Mobile Computing*, vol. 12, no. 2, pp. 358–370, 2013.

[12] W. Diffie and M. Hellman, "New directions in cryptography," *IEEE Trans. Inf. Theor.*, vol. 22, no. 6, pp. 644–654, Sep. 1976.

[13] W. Stallings, *Cryptography and Network Security: Principles and Practice*, 6th ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2013.

[14] D. Schürmann and S. Sigg, "Secure communication based on ambient audio," *IEEE Transactions on Mobile Computing*, vol. 12, no. 2, pp. 358–370, Feb 2013.

[15] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *International Society for Music Information Retrieval Conference*, 2002, pp. 107–115.

[16] A. Wang *et al.*, "An industrial strength audio search algorithm." in *ISMIR*, vol. 2003. Washington, DC, 2003, pp. 7–13.

[17] I. S. Reed and G. Solomon, "Polynomial Codes Over Certain Finite Fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960.

[18] S. Li, C. Li, C. Ding, and H. Liu, "Two families of lcd bch codes," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5699–5717, 2017.