Pattern-based Alignment of Audio Data for Ad-hoc Secure Device Pairing

Ngu Nguyen[*] and Stephan Sigg, An Huynh, Yusheng Ji[**]

[*] *University of Science, Ho Chi Minh City, Vietnam*
[**] *National Institute of Informatics, Tokyo, Japan*
Email: *nlnngu@fit.hcmus.edu.vn*, *sigg@nii.ac.jp*, *zanton.zzz@gmail.com*, *kei@nii.ac.jp*

Abstract

When studying the use of ambient audio to generate a secure cryptographic shared key among mobile phones, we encounter a misalignment problem for recorded audio data. The diversity in software and hardware causes mobile phones to produce badly-aligned audio chunks. It decreases the identical fraction in audio samples recorded in nearby mobile phones and consequently the common information available to create a secure key. Unless the mobile devices are real-time capable, this problem can not be solved with standard distributed time synchronisation approaches. We propose a pattern-based approximative matching process to achieve synchronisation independently on each device. Our experimental results show that this method can help to improve the similarity of the audio fingerprints, which are the source to create the communication key.

1. Introduction

With recent advances in smart-phone dissemination and their computational capabilities, smart-phones can be seen as a kind of wearable device for the masses. These general-purpose devices are capable of solving several wearable computing tasks. Due to their high penetration, security in communication among devices becomes a relevant issue. Common security schemes for mobile devices require explicit user input to provide a shared piece of information. A wearable device, however, should not distract its holder from other tasks. How can we provide security among possibly unacquainted devices without any user interaction?

We consider an interaction-free common key generation scheme for proximate devices conditioned on ambient audio. Each device computes a binary characteristic sequence for a synchronised recording: An audio-fingerprint [4, 2]. This binary sequence is designed to fall onto a code-space of an error correcting code [11]. Devices then exploit the error correction capabilities of the error correcting code to map fingerprints to codewords as described in [6, 12]. For fingerprints with a Hamming-distance within the error correction threshold of the error correcting code the resulting codewords are identical and then utilized as secure keys. The Hamming distance in fingerprints rises with increasing distance between devices so that distant devices are unlikely to guess the correct key. Our fingerprint extraction scheme is adapted from [4] to extract fingerprints from synchronized ambient audio recordings in a noisy environment without exchanging any information about the resources among devices. However, when audio sequences utilized are not well aligned, similarity in fingerprints decreases. This is due to the fingerprint generation which exploits the relative fluctuation of energy over time. Simple time synchronisation approaches, for instance the network time protocol (NTP), are not suitable to sufficiently synchronise audio recordings due to the additional delays in the recording hardware.

This paper addresses the accurate alignment of recorded audio sequences from remote devices. The challenging point here is to achieve an alignment between audio samples taken from distinct devices interaction free and without any inter-device communication other than an initial plain pairing request.

We will in section 2 discuss related work on secure ad-hoc pairing of mobile devices. Problems that prevent accurate audio sequence alignment and our pattern-based approximative matching method to reduce the mismatching are detailed in section 3. Section 4 describes a case study with smart-phones to investigate the accuracy of our alignment scheme. Section 5 draws our conclusion.

1550-4816/12 $26.00 © 2012 IEEE
DOI 10.1109/ISWC.2012.14

IEEE computer society

2. Related Work

Contextual or sensor information of mobile devices can be incorporated for authentication [5]. When the seed to the key is implicit with the context, no information that could be used to reconstruct it need to be transmitted over a wireless channel during key generation. For instance, McCune et al. [10] introduced *Seeing-Is-Believing*, utilizing the camera of a mobile device to capture a 2D barcode which is displayed on the screen of another device. *Loud and Clear* of Goodrich et al. [3] implements a similar scheme but exploits spoken audio. A user reads a text message displayed on one device and a second device recognizes the speech for authentication. Another mechanism by Mayrhofer et al. [9] uses accelerometer readings when devices are shaken simultaneously by a single person. Mayrhofer derived in [8] that the sharing of secret keys is possible with a similar protocol by repeatedly exchanging hashes of key-subsequences until a common secret is found. Bichler et al. generalise this approach to noisy acceleration readings [1]. They utilize a hash function that maps similar acceleration patterns to identical key sequences. These approaches require explicit user interaction.

By utilising a context source that provides a sufficient amount of unique, context-related information, such as audio or radio frequency (RF), it is possible to get the user out of the loop. Mathur et al. introduced ProxiMate which enables wireless devices in proximity to pair automatically and securely using their shared ambient RF-signals [7]. They generate fingerprints from RF-channel fluctuations and map these onto a codespace of an error-correcting code. By correcting potential errors in the fingerprints, they are mapped onto the closest regular codeword in the codespace. When the similarity between fingerprints is high, codewords are identical. Sigg et. al proposed to use audio instead of RF in a similar implementation [13]. They study the entropy of audio fingerprints and identify time synchronisation as a main hindrance to practically apply the method for mobile devices. Their instrumentation requires idealized conditions regarding the synchronisation of devices and to account for this a high number of fingerprints must be created (201 in their experiments) in order to find one matching fingerprint. For extensive computational load, this is feasible only in an offline approach. The high number of fingerprints created, however, was necessary since the utilized NTP synchronisation is not sufficiently accurate.

In this paper, we present an alignment mechanism that enables a synchronisation accuracy of recorded audio in the order of less than 10 milliseconds among mobile devices in the same context. The synchronisation is achieved by processing a weakly NTP-synchronised recording without transmitting information about the audio sequences over the wireless channel.

3. Pattern-based alignment of audio data

When developing the scheme of unobtrusive secure device pairing with audio fingerprints for Android-based mobile devices, we encountered practical issues not evident when considering the problem theoretically. One issue in practical implementation is differing audio hardware. For instance, the Samsung Google Nexus S[1] and HTC Google Nexus One[2] devices we utilized apply different audio-preprocessing routines that render the unprocessed audio outputs on these devices unusable for the generation of identical fingerprints. Furthermore, time synchronisation is a serious problem for the approach. In particular, not only the clocks on remote devices have to be synchronised as usual for distributed devices, but also the generally unknown and possibly non-constant hardware specific delays on both devices need to be taken into account.

3.1. Misalignment of audio data

Due to the differences in software and hardware of the devices, we observe that the recorded audio sequences are not exactly the same. An example of this phenomenon is shown in figure 1. All audio files are recorded with the same settings and at the same time (clocks synchronised by a NTP service[3]). From the similar visual appearances in the waveform format of the signals in figure 1, the recording start time of the Nexus One was heavily delayed when compared to the Nexus S.

We observe that the offset when both recordings start is fluctuating and in all cases much higher than the accuracy expected from NTP synchronisation. In our recorded audio files, time difference ranges from 0.3 seconds to more than 1 second. Additionally, as depicted in figure 1, the higher frequency bands of the signal available at the Nexus One device completely differ from the Nexus S readings because the Nexus One employs a hardware noise cancellation. There is no way to bridge the noise cancellation on that device to obtain the unmodified signal. Both these effects are unfortunate for our fingerprinting method.

3.2. Aligning recorded audio data

One important condition for our secure device pairing scheme is that no information regarding the recorded audio shall be exchanged between the devices. Otherwise, the security of the key might be impaired by information leaking from these transmissions. To reduce the mismatch between audio data from neighbouring devices, we

[1] `http://www.google.com/nexus/tech-specs.html`

[2] `http://www.google.com/phone/detail/nexus-one`

[3] Navy Clock II application: `https://market.android.com/details?id=com.cognition.navyclock`

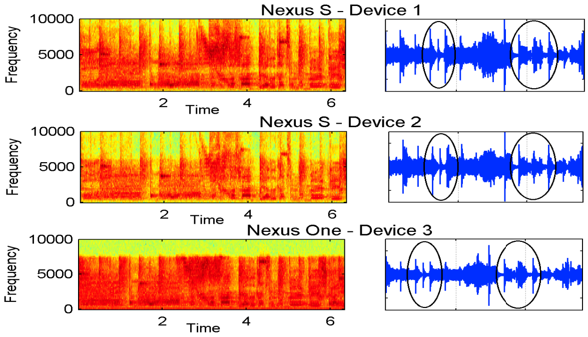

Figure 1. Visualization of audio data.

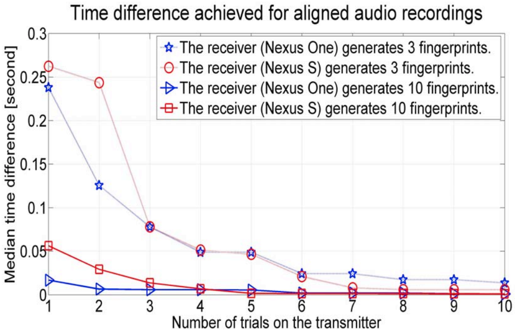

Figure 2. Time difference of aligned audio.

propose a synchronization scheme which is based on the Smith-Waterman pattern matching algorithm [14]. Since the matching is approximative, it will always find a best matching position even though the absolute similarity of this very position might not be high. We exploit this property to be able to resign from any information transmitted among devices on the actual recorded audio. In particular, we utilize a predefined, characteristic pattern on both devices. since the pattern is known in advance, no information need to be transmitted. The downside to this implementation is, of course, that the pattern utilised might be very different from the actual audio recorded. However, since the Smith-Waterman algorithm always computes a best matching, we can speculate that this best matching is found at similar positions in the audio recordings, provided that the data has significant similarity. The specific pattern used for matching is extracted randomly from consecutive samples of an arbitrary audio sequence. In our experiments, its length is 100 samples (longer patterns increase running time of the algorithm but have insignificant impact on the accuracy). The pattern p is matched in the first 100000-sample part of each audio file. The matching score of p is the sum-difference between amplitude values. In to our experiments, the gap penalty that can yield an acceptable matching is 150. After finding the matching positions, we eliminate all samples preceding the matching positions and generate the audio fingerprints from the remaining sequence.

We experienced that only seldom a perfect synchronizing result among two best matching points on both devices is found. We therefore calculated a set of k best matching points. A device then chooses the best matching points, encodes a data sequence possibly multiple times with several keys and transmits this to the second device. The receiver then attempts to decode the data with the keys generated from its best matching points. This process might require the transmitter to send a data sequence several times, encrypted with different keys each time. Although this increases the communication load, this process could be implemented in an iterative fashion and is required only for the first encounter of the devices to derive the secret key. Since all data blocks are encrypted, only marginal additional information is provided to a potential adversary.

4. Experimental results

We utilize two Android-based mobile phones in our experiments. The HTC Google Nexus One smart-phone has a secondary microphone dedicated for dynamic noise suppression, while the Samsung Google Nexus S device uses software-based noise cancellation. The ambient audio data are recorded for 6375 milliseconds at a sampling rate of $44100\ Hz$. We attempt to generate a secure key among a Nexus S and a Nexus One device after applying a bandpass filtering and aligning audio sequences with the pattern matching approach. The bandpass filtering only retains signals whose frequency is between $4000Hz$ and $4500Hz$. The smart-phones are put at the same distance d from an audio source and we increase the distance between two devices from $10cm$ to $100cm$. With each inter-device distance, we record 10 audio files on each phone.

Figure 2 depicts the median time difference after alignment when increasing the number of trials in the transmitter. The receiver attempts to decode the data with its sequences from the top 3 and 10 matching positions. We can achieve a synchronisation in the order of 10 milliseconds with the 3 best matching points considered at the transmitter (cf. figure 2). Moreover, with the 10 best matching points, we observe that, already with the best matching trial at the transmitter, the matching is greatly improved. With 3 trials of the transmitter, the synchronisation time is sufficiently accurate for our secret key generation approach. Since the alignment approach calculates all possible alignments at once, the cost for additional trials at the receiver is low. Additional trials at the transmitter, however, directly impact the communication load.

We conclude that the alignment approach greatly improves the synchronisation of audio recordings on remote devices without additional inter-device communica-

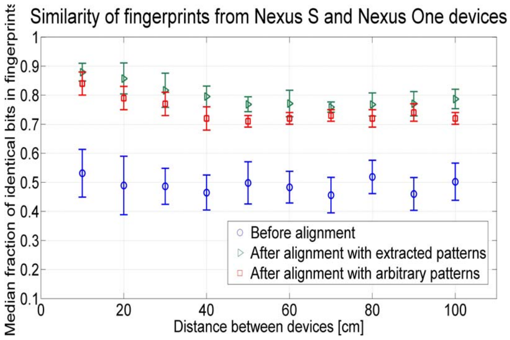

Figure 3. Similarity of audio fingerprints.

tion. Figure 3 depicts the fraction of identical bits among fingerprints before and after the pattern-based alignment is applied. While the fingerprint similarity initially only marginally deviates from the similarity to a random sequence, it is greatly improved after the alignment. The fraction of identical bits in the audio fingerprints decreases when inter-device distance increases. In particular, also the variance in the data is reduced. These characteristics allow for a sharper threshold of the error correcting code. For comparison, we also extracted the pattern from one of the recorded audio files to align it with the audio sequence on the other device (Exploiting the undesirable case of data transmission between devices). The quality of the audio fingerprints is comparable to the results achieved with arbitrary patterns, even with the best matching results (cf. figure 3). Consequently, disclosing data on the audio files over the wireless channel does not pay off in improving fingerprints.

5. Conclusion

When implementing an audio-based ad-hoc secure device pairing mechanism for previously unacquainted mobile devices, the diversity of hardware and software can affect the offset in audio recordings of even clock-synchronised mobile devices. We propose an approximative pattern matching method to align the corresponding audio without communication between the devices. The devices synchronise their audio sequences without any knowledge about the recorded audio on the remote device other than their own recorded contextual information. Hence, no information about the audio utilized as a seed for the secure key generation, can leak. To improve the alignment quality, we can choose more than one matching position on each device at the cost of increasing the communication load. We can obtain a synchronization among devices of less than 2 milliseconds when both devices utilize up to 10 matching positions. With 3 trials, a synchronisation in the order of 10 milliseconds is reasonable.

Acknowledgment

This work was supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD).

References

[1] D. Bichler, G. Stromberg, M. Huemer, and M. Loew. Key generation based on acceleration data of shaking processes. In J. Krumm, editor, *Proceedings of the 9th International Conference on Ubiquitous Computing*, 2007.

[2] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A review of audio fingerprinting. *Journal of VLSI Signal Processing Systems*, 41 Issue 3, 2005.

[3] M. T. Goodrich, M. Sirivianos, J. Solis, G. Tsudik, and E. Uzun. Loud and clear: Human-verifiable authentication based on audio. In *Proceedings of the 26th IEEE International Conference on Distributed Computing Systems*, 2006.

[4] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In *3rd International Conference on Music Information Retrieval*, pages 107–115, Paris, France, 2002.

[5] L. E. Holmquist, F. Mattern, B. Schiele, P. Schiele, P. Alahuhta, M. Beigl, and H. W. Gellersen. Smart-its friends: A technique for users to easily establish connections between smart artefacts. In *Proceedings of the 3rd International Conference on Ubiquitous Computing*, 2001.

[6] A. Juels and M. Wattenberg. A Fuzzy Commitment Scheme. *Sixth ACM Conference on Computer and Communications Security*, pages 28–36, 1999.

[7] S. Mathur, R. D. Miller, A. Varshavsky, W. Trappe, and N. B. Mandayam. Proximate: proximity-based secure pairing using ambient wireless signals. In *MobiSys*. ACM, 2011.

[8] R. Mayrhofer. The Candidate Key Protocol for Generating Secret Shared Keys from Similar Sensor Data Streams. *Security and Privacy in Ad-hoc and Sensor Networks*, 2007.

[9] R. Mayrhofer and H. Gellersen. Shake well before use: Authentication based on accelerometer data. *Pervasive Computing*, pages 144–161, 2007.

[10] J. M. McCune, A. Perrig, and M. K. Reiter. Seeing-is-believing: Using camera phones for human-verifiable authentication. In *Proceedings of the 2005 IEEE Symposium on Security and Privacy*, 2005.

[11] I. Reed and G. Solomon. Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics*, pages 300–304, 1960.

[12] B. Schneider. *Applied Cryptography: Protocols, Algorithms, and Source Code in C*. John Wiley and Sons, Inc., 2 edition, 1996.

[13] S. Sigg, D. Schuermann, and Y. Ji. Pintext: A framework for secure communication based on context. In *Proceedings of MobiQuitous 2011*, 2011.

[14] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, Mar. 1981.