Robust and Responsive Acoustic Pairing of Devices Using Decorrelating Time-Frequency Modelling

Pablo Pérez Zarazaga, Tom Bäckström
Department of Signal Processing and Acoustics
Aalto University, Espoo, Finland
(pablo.perezzarazaga; tom.backstrom)@aalto.fi

Stephan Sigg
Department of Communications and Networking
Aalto University, Espoo, Finland
stephan.sigg@aalto.fi

***Abstract*—Voice user interfaces have increased in popularity, as they enable natural interaction with different applications using one's voice. To improve their usability and audio quality, several devices could interact to provide a unified voice user interface. However, with devices cooperating and sharing voice-related information, user privacy may be at risk. Therefore, access management rules that preserve user privacy are important. State-of-the-art methods for acoustic pairing of devices provide fingerprinting based on the time-frequency representation of the acoustic signal and error-correction. We propose to use such acoustic fingerprinting to authorise devices which are acoustically close. We aim to obtain fingerprints of ambient audio adapted to the requirements of voice user interfaces. Our experiments show that the responsiveness and robustness is improved by combining overlapping windows and decorrelating transforms.**

***Index Terms*—Voice User Interface, Acoustic Pairing, Audio Fingerprint, DCT**

I. Introduction

Voice User Interfaces (VUI) enable intuitive interaction with devices, similar to interaction between people. As VUI technology has rapidly improved, VUIs have gained in popularity. Applications include, for example, voice assistants [1]–[3], where people talk and interact to an artificial intelligence (AI), and telecommunications applications such as Skype, which allow voice communication. However, these applications bind the user to the device that provides this VUI. In the developed world, the majority of the population owns multiple devices containing microphones, e.g., smart phones, tablets or laptops. If all these collaborated in a distributed network sharing the recorded information, free mobility would be possible without carrying a recording device. Voice could be recorded by the device that received the cleanest signal or, if multiple devices heard the same speech, multichannel coding could achieve a more efficient extraction of the speech signal [4], [5].

Multi-device VUIs, however, give rise to privacy concerns. Specifically, if all devices are continuously recording and transmitting the voice of the user, one has to consider the effect that it may have on the privacy of the user. The voice does not only contain features that reveal personal information, like emotion or health condition [6], but people are also less careful in information disclosure via voice [7]. Therefore, the devices of a distributed VUI need robust and intuitive privacy management.

This research was supported by the Academy of Finland project No 312490.

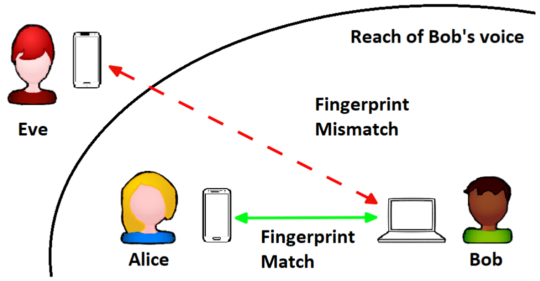

Fig. 1. Scene representing the acoustic space and how the pairing algorithm should ideally work.

Figure 1 illustrates the intended functionality of the proposed device pairing. An acoustic space can be defined as the area where a person's voice is audible and comprehensible. This acoustic space depends on the distance between the devices, but it is also strongly affected by the acoustic conditions of the room. For example, closing the door of a room implies the creation of an acoustic barrier that will set a boundary for the acoustic space. The same applies to a noisy cafeteria, where a high noise level masks the speech signal and reduces the reach of the acoustic space.

This paper proposes an intuitive access management system for acoustic information. Access is awarded to devices residing in the same acoustic space. Therefore, we present pairing methods based on acoustic fingerprints. Successfully paired devices are then granted permission to collaborate. To improve on the state-of-the-art in Section II, we propose a method based on overlapping windows and decorrelation of the frequency bands in every time frame, presented in Section III. The performance of the proposed method is evaluated and compared to state-of-the-art methods in Section IV, by measuring robustness against noise and signal delay, as well as through statistical analysis of the generated fingerprints to evaluate their cryptographic strength.

II. Acoustic Pairing

Device pairing from acoustic features can be classified into two categories: Active and passive methods. In active methods, one device will generate a known signal, which is analysed by other devices to determine if pairing is feasible [8]. However, active methods require that the devices play an acoustic signal

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

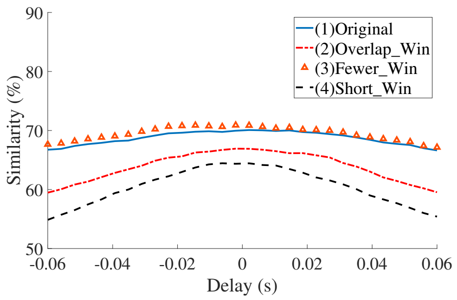

Fig. 2. Similarity of fingerprints between distinct microphones with varying delay using the original method with (1) 300 ms non-overlapping windows, (2) 200 ms overlapping windows, (3) reduced number of windows and (4) 100 ms non-overlapping windows.

every time a new device needs authentication, which is not viable in an intuitive voice user interface. Alternatively, passive methods take advantage of external sounds to evaluate if two devices are located close to each other. Passive methods usually generate a fingerprint which contains different properties of the recorded audio that can be compared to evaluate their similarity [9]–[11]. Passive methods depend on the environment sounds and not on a controlled signal as active methods, therefore, their robustness is lower. However, they adapt better to the requirements of intuitive authentication in a distributed VUI.

A passive method can obtain a fingerprint that can be used to calculate a cryptographic key [12]. This key will be used to authenticate the devices in the acoustic sensor network. To calculate the fingerprint, 6.375 s of audio are recorded and divided into 17 non-overlapping windows with a duration of 375 ms. The DFT of every time frame is calculated and the frequency bands are grouped into 33 energy bands. The fingerprint is then calculated as the sign of the difference between consecutive bands in time and frequency [10].

Real-life environments however feature background and sensor noises, whereby we have to use error-correction methods to avoid bit-errors [12]. Therefore, if two fingerprints are sufficiently similar, a shared secret is generated and the devices will be able to share information without disclosing their fingerprint.

The main issue with the above method is that the length of the recording is required to be at least 6 s. A whole conversation can take place in 6 s, therefore, if the fingerprint calculation takes longer than that, the whole conversation would be lost. A faster response is necessary to establish the initial pairing and to stop sharing information if a device in the distributed VUI leaves the acoustic space.

Trivial approaches to reducing the length of the recordings would be to use a smaller number of windows or reduce their length. Figure 2 shows the effect of these modifications on the original fingerprint [12] to reduce the length of the recordings to 2.2 s. The performance of the fingerprints is represented as a similarity value between two fingerprints from different recording devices. The similarity is calculated as the ratio of matching bits between the compared fingerprints.

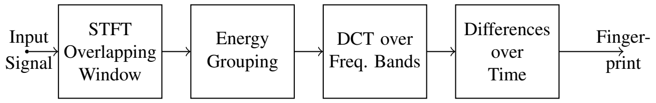

Fig. 3. Block diagram of the proposed fingerprint calculation.

Delay between the recorded signals is the main source of degradation in the fingerprint matching. Therefore, shorter windows (100 ms) reduce the quality of the fingerprint. Alternatively, by reducing the number of windows from 17 to 6, fingerprint quality is mostly retained but its length degrades from 512 to 160 bits, which impairs cryptographic strength.

III. Proposed Method

The objective of the proposed method is robust authentication of devices with low delay, such that it is suitable for VUI applications. As a consequence, we use passive acoustic pairing of devices that avoids long recordings. The novelty is a fingerprint from reduced-length recordings which improves the robustness to noise compared to previous algorithms [12].

Consider two microphone signals, x_1 and x_2, which are synchronised to within 10 ms, with different sensor noises, and distorted by slightly different room impulse responses. We aim to obtain a fingerprint $f(x)$ such that $|f(x_1) - f(x_2)|$ is minimised when the signals are considered to match and maximised otherwise. Error-correcting codes can then rectify the remaining differences and similar fingerprints will match despite bit errors [12].

To design such a fingerprinting function $f(x)$, we use a modification of the previously proposed time-frequency transform as presented in Figure 3. We reduce the recording length to 2.2s using 17 overlapping windows with a duration of 200 ms and an overlap of one third of a window. The windows are transformed to the frequency domain using the discrete Fourier transform. To keep the number of coefficients equal to [12], we calculate the energy of 32 uniformly distributed energy bands in the spectrum of each frame. Decorrelation of the frequency components is then proposed to compensate the degradation produced by the shorter overlapping windows.

A further benefit of this approach is that it makes the systems structure more similar to speech coding methods also present in the device [13], [14]. The energy band grouping resembles an estimation of the spectral envelope, which is a common operation in speech codecs. If the length of the windows could be reduced to a similar length as the device's codec, we could use the codec's spectral envelope to to obtain the fingerprint. The benefit is that by sharing modules across different tasks, we could reduce the overall computational complexity and resource consumption.

A discrete cosine transform (DCT) is then applied over the frequency axis in every frame as a decorrelation function. The decorrelating properties of the DCT are widely used in speech coding [14], [15] and in this case, the decorrelation of the energy bands increases the robustness of the fingerprint

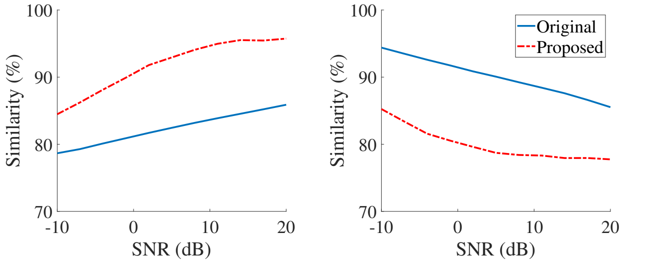

Fig. 4. Similarity values for special non-matching scenarios. (Left) Matching speech with different noises. (Right) Matching noise and different speech signals

and improves its accuracy. The decorrelation function also benefits the fingerprint from a cryptographic viewpoint. The fingerprint should resemble a random sequence, for which its serial correlation has to be minimised.

Finally, we define the bits of the fingerprint as the sign of the difference between a DCT component and the corresponding one in the next time slot. The outcome is a sequence of 512 bits, in accordance with the fingerprint in [12], which allows an easy comparison of methods.

IV. Performance Evaluation

To quantify the performance of the proposed method in comparison to state-of-the-art, we measured the similarity between two fingerprints from different devices in matching and non-matching scenarios. First, we evaluate the robustness of the fingerprint with respect to noise. Second, we investigate how delay between signals reduces the performance. Finally, we analyse the statistical properties of the generated fingerprints to evaluate their cryptographic strength.

A. Noisy Speech Database and SNR Analysis

Our first experiment evaluates the fingerprints at different SNR values using a noisy speech database. We used the TIMIT speech corpus [16] with additive noise from the QUT noise database [17]. Specifically, first we randomly selected 100 speech files and 20 types of noise. Second, noises were added at different SNR levels from -10 to 20 dB with 3 dB increments. Every speech-noise pair is used to generate 3 files with random delays around ± 15 ms between audio files to simulate the different positions of the recording devices. The resulting database contained 60000 noisy speech files with a duration of 10 s each.

A speech signal affected by a single additive noise source is a simplification of sound propagation as signals are altered by the impulse response of the room and the noise comes from multiple directions depending on the position of its sources. This is useful to control the SNR level and analyse the influence of speech and non-speech signals in the calculation of the fingerprints. However, the expected performance in this section will be higher than a real situation and it can not be considered a realistic measurement.

The original fingerprint was implemented with 300 ms non-overlapping windows [12]. For the proposed method, 200 ms windows with an overlap of one third of a window are used. To obtain a fingerprint with the same length using both methods, the analysed frame is divided into 17 windows. The recordings are 6.375 and 2.2 s respectively. To determine the impact of the windowing function, multiple windows used in speech coding were tested, e.g., Hamming, Hann or half-sine. Informal experiments did not show any noticeable difference over the multiple windows, therefore, the window can be chosen according to the requirements of the final application.

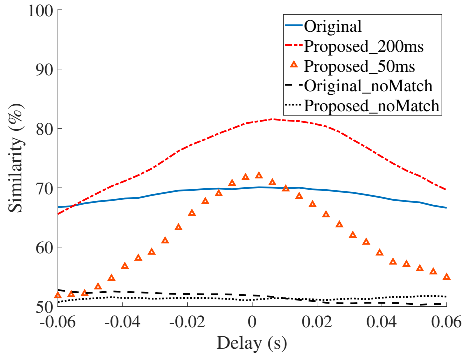

Fig. 5. Similarity of fingerprints between distinct microphones with varying delay in an office environment. Original, proposed with 200 ms and 50 ms overlapping windows, as well as comparison to microphones that do not match for original and proposed.

TABLE I
Average similarity values over 100 matching fingerprint pairs using 200 ms windows in different scenarios.

Scenario	Office	Conference Room	Street
Original	62.3 %	66.0 %	57.7 %
Proposed	71.5 %	77.3 %	68.2 %

Every file was compared to all the other files with the same SNR and classified depending on whether the speech and noise match. When both speech and noise signals match, the number of identical bits remains almost constant around 95 % over all the SNR values in both methods. The result is also independent to the SNR when both speech and noise do not match, showing a reduction in the similarity of the fingerprints.

Figure 4 shows the results when the speech signals match but the background noise is different and vice-versa. When both signals contain the same speech, we observe that the proposed method provides higher accuracy at high SNR levels. When the speech signal differs but the background noise is the same, fingerprint similarity reached with the proposed method is lower than the original method. This represents the effect of the acoustic space in places with a high level of noise. Consequently, in the proposed method, signals resembling speech have a higher impact on the calculated fingerprint than in the original method.

B. Real-scenario Recordings and Synchronisation Effect

To evaluate how the proposed method would behave in a real VUI, we collected informal recordings of conversations in multiple realistic scenarios. The chosen locations were a

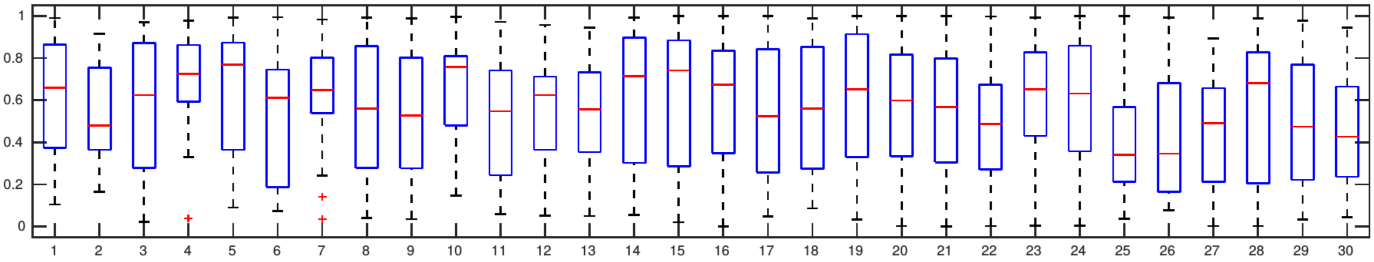

Fig. 6. Distribution of p-values for keys after 20 runs of the DieHarder set of tests. (1) birthdays (2) operm5 (3) rank32x32 (4) rank6x8 (5) bitstream (6) opso (7) oqso (8) dna (9) count-1s-str (10) count-1s-byt (11) parking (12) 2D circle (13) 3D sphere (14) squeeze (15) runs (16) craps (17) marsaglia (18) sts monobit (19) sts runs (20) sts serial [1-16] (21) rgb bitdistr. [1-12] (22) rgb min dist. [2-5] (23) rgb perm. [2-5] (24) rgb lagged sum [0-32] (25) rgb kstest (26) dab bytedistr. (27) dab dct (28) dab filltree (29) dab filltree 2 (30) dab monobit 2.

TABLE II
ENT PSEUDORANDOM NUMBER SEQUENCE TEST PROGRAM RESULTS FOR FINGERPRINTS GENERATED FROM AUDIO SEQUENCES (430080 BITS).

Parameter	Results	Optimum	Worst case
Entropy	0.9956	1	0
Optimum Compression	0 %	0 %	100 %
Arithmetic mean	0.5389	0.5	0 or 1
Serial correlation	0.0815	0	1

small private office, a conference room with a long table in the middle and the street. Three mobile phones (OnePlus 5T, HTC10, HUAWEI Honor 5X) were used to record different conversations between three people. In the first two scenarios, two of the phones were placed on the table at a distance of 1 to 2 meters from each other, and the third one was in one of the user's pockets. The third scenario was staged on the street and the users held the recoding devices in their hands.

Table I shows the similarity between matching synchronised fingerprint pairs in the different scenarios described above. These were calculated averaging the similarity of 100 matching fingerprint pairs in every scenario. The proposed method provides a similarity 10 % higher in all the evaluated scenarios, which would increase the robustness of the authentication.

Informal experiments show that the main source of distortion in the fingerprint matching is the delay between the signals, specifically delays caused by acoustic distance and transmission delays. The fingerprints should match even with synchronisation errors, such that $f(x_1(k)) \approx f(x_2(k-\Delta_{sync}))$. The effect of delays is attenuated in [12] by using long non-overlapping windows. The degradation due to delay in overlapping windows is noticeable in Figure 2. The decorrelation should also compensate the degradation for short delays.

To evaluate the effect of discrepancies in synchronization, we measured how such delays degrade fingerprint similarity. Figure 5 depicts the performance of the two methods in an office scenario under delays in the range of -60 to 60 ms. Shorter windows have a negative effect on the robustness to delays, however, when delays are under 60 ms, the performance of the proposed method is still higher than the original. Note that in non-matching scenarios, both methods present a similarity between fingerprints around 50 %, which is the expected value of comparing two random binary sequences.

C. Entropy and Statistical Analysis

We also need to evaluate the performance of the fingerprints from a cryptographic perspective, which can be measured using their statistical properties. Our objective is to generate fingerprints that resemble a random sequence of bits. To analyse the statistical properties of the generated fingerprints and their entropy, multiple fingerprints were calculated using the recordings from three mobile devices in multiple scenarios as it is described in Section IV-B. In total, 3780 fingerprints were used.

The DieHarder statistical tests [18] were used to evaluate whether the proposed scheme is robust against bias in the generated random sequences. Figure 6 depicts the p-values computed from 20 runs of the DieHarder tests. While these tests can not replace cryptanalysis they are designed to uncover bias and dependency in the pseudo random sequence.

Every test has an expected distribution of outcomes; test runs produce a value that is compared to the theoretical outcome. A p-value between 0 and 1 is then computed, describing the probability that a real random number generator (RNG) would produce this outcome. A good RNG will have a range of p-values that follows a uniform distribution. A p-value below a fixed significance level $\alpha = 0.001$ indicates a failure of the RNG with probability $1 - \alpha$. 100 p-values are computed in a single run of DieHarder and their distribution is compared to a uniform one.

As depicted in Figure 6, the p-values are well distributed and centred around the mean at roughly 0.5. While few tests deviate slightly, and few tests were weaker than the majority, not a single test failed in the 20×100 repetitions of all tests. Therefore, no statistical bias was found for the random sequences generated by the proposed method.

Additionally, the results of an *Ent Pseudorandom Number Sequence* Test [19] are summarised in Table II. In this test, the information density of bit sequences is computed together with reduction capabilities through optimal compression, arithmetic mean of data bytes as well as the serial correlation coefficient. We observe that the results obtained from the calculated fingerprints barely deviate from the optimum values.

V. Conclusion

This paper presents a method for authentication of devices in voice user interfaces based on an acoustic fingerprint. In comparison to previous methods [8], [9], [12], the proposed method provides a higher accuracy in determining fingerprints and a lower latency in the authentication process, thus increasing its responsiveness.

We propose shorter, overlapping windows, such that a shorter segment of audio is used to calculate the acoustic fingerprint. This, simultaneously, makes the calculation more similar to typical speech processing [13], [14]. This may prove useful in future applications where parts of the processing could be shared between different speech processing methods [4], [5]. The decorrelation of the energy bands provides a higher accuracy in the fingerprint matching, compensating the degradation produced by the shorter windows. The shorter speech segment also contributes to sharing a lower amount of information, consequently protecting the users' privacy.

In conclusion, the presented method shows that a combination of overlapping windows and decorrelating functions can be used to generate an acoustic fingerprint using shorter audio segments while maintaining the performance of state-of-the-art. This method provides a robust and responsive authentication for devices in a distributed voice user interface that adapts to the properties of the acoustic space in a variety of scenarios.

References

[1] 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.

[2] M. J. Callaghan, V. B. Putinelu, J. Ball, J. C. Salillas, T. Vannier, A. G. Eguíluz, and N. McShane, "Practical use of virtual assistants and voice user interfaces in engineering laboratories," in *Online Engineering & Internet of Things*, Michael E. Auer and Danilo G. Zutin, Eds. 2018, pp. 660–671, Springer International Publishing.

[3] P. Bartie, W. Mackaness, O. Lemon, T. Dalmas, S. Janarthanam, R. L Hill, A. Dickinson, and X. Liu, "A dialogue based mobile virtual assistant for tourists: The spacebook project," *Computers, Environment and Urban Systems*, vol. 67, pp. 110–123, 2018.

[4] T. Bäckström, F. Ghido, and J. Fischer, "Blind recovery of perceptual models in distributed speech and audio coding.," in *Interspeech*, 2016, pp. 2483–2487.

[5] T. Bäckström and J. Fischer, "Fast randomization for distributed low-bitrate coding of speech and audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 19–30, Jan 2018.

[6] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572 – 587, 2011.

[7] J. Schroeder and M. Schroeder, "Trusting in machines: How mode of interaction affects willingness to share personal information with machines," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[8] H. T. T. Truong, J. Toivonen, T. D. Nguyen, S. Tarkoma, and N. Asokan, "Proximity verification based on acoustic room impulse response," *arXiv preprint arXiv:1803.07211*, 2018.

[9] W. Tan, M. Baker, B. Lee, and R. Samadani, "The sound of silence," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2013, p. 19.

[10] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system.," in *ISMIR*, 2002, vol. 2002, pp. 107–115.

[11] V. Chandrasekhar, M. Sharifi, and D. A. Ross, "Survey and evaluation of audio fingerprinting schemes for mobile query-by-example applications.," in *ISMIR*, 2011, vol. 20, pp. 801–806.

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

[13] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*, Springer, 2007.

[14] T. Bäckström, Ed., *Speech Coding with Code-Excited Linear Prediction*, Springer, 2017.

[15] G. Fuchs, C. R. Helmrich, G. Marković, M. Neusinger, E. Ravelli, and T. Moriya, "Low delay LPC and MDCT-based audio coding in the EVS codec," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5723–5727.

[16] J. S. Garofolo et al., "TIMIT acoustic-phonetic continuous speech corpus ldc93s1. web download," Philadelphia: Linguistic Data Consortium, 1993.

[17] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-noise-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Interspeech 2010*, Makuhari Messe International Convention Complex, Makuhari, Japan, September 2010.

[18] R. G. Brown, Eddelbuettel D., and Bauer D., "DieHarder: a random number test suite," https://webhome.phy.duke.edu/~rgb/General/dieharder.php, 2019, Online, accessed 28-February-2019.

[19] J. Walker, "ENT: a pseudorandom number sequence test program," *Software and documentation available at/www. fourmilab. ch/random/S*, 2008.