# Literature Review

Aidan Fray

May 31, 2019

## 1 Review - Fingerprint representations

### 1.1 Fingerprint representation comparison

**TODO:** - Create written introduction setting the scene.
**TODO:** - Add participant size to M. Shirvanian (25)
**TODO:** - Create a background section for Compare-and-X

Current research into human-based validation of fingerprints has almost exclusively focused on the usability of such schemes. The following section will individually discuss available research findings alongside an overall comparison of the research recommendations.

Some of the first work in this area was performed by **Hsu-Chun Hsiao, *et al.***[1]. The main aim of the study was to provide the first insight into the best encoding schemes used to represent a hash fingerprint. The schemes used were Base32, English words, Random Art[2], Flag[3], T-Flag[4], Flag Ext[1] and finally Chinese, Korean and Japanese symbolic encodings. A total of 436 participants were assessed in the study.

In addition to the central aims of the paper, the authors intended to provide empirical links between participant attributes (such as age or gender) and performance with the encoding scheme. This is something rarely seen in the related research. Hsu-Chun Hsiao *et al.* also provided two categories of similarities; "hard" and "easy". "Hard" pair only have a very small difference, therefore, difficult to distinguish. These were designed to be the worst and best case respectively. The reasoning behind inclusion of Chinese, Korean and Japanese characters was to investigate the links between language ability and distinguishing subtle differences between encodings in the respective language. The entropy of encodings ranged from 22-bits to 28-bits. This is the lowest level of entropy used in the assessed literature. To quantify this a hash rate of 10 MH/s SHA-1

[1] The authors' own improvement on Flag

compressions would be able to find a match for 28-bits in 26.8 seconds. This is addressed by the author at a stage in the paper where they state *"However, increasing entropy is not a solution because it sacrifices usability and accuracy. With more entropy, representations will contain longer sequences of characters or more minute details which will lead to increased time and errors during comparisons."*. This claim is entirely unsubstantiated with any empirical evidence alongside no discussion into why the level of entropy was chosen. This, however, was followed by in interesting recommendation into the use of commitments[5]. This provides an attacker with a single chance to find a collision, thus, providing a valid use case for low levels of entropy.
Another limitation is the authors' decision to exclude hexadecimal and numerical encodings (some of the most widespread encoding schemes). This was based on their own claims on similarities to Base32 and "well-known deficiencies" in the excluded schemes. This point was provided with no further justification or quantified in any way. It is also not consistent with available research, for example, in *"Empirical Study of Textual Key-Fingerprint Representations"*[6] it was shown numerical representations performed significantly better than that of Base32.

Findings from the paper showed that age and gender do not affect the accuracy of the scheme, however, younger participants were significantly faster.
The paper recommends Base32, Random Art, T-Flag or their own improvement; Flag Ext. These encoding recommendations were also supplemented with a review of the requirements required by these schemes. The recommendation of graphical encoding schemes is inconsistent with alternative literature, where most other papers have found graphical encodings easy to use but insecure. In terms of language comprehension; being able to speak the language assists in the ability to discern the differences between hard pairs. Subsequently, knowledge of the language did not assist in differentiating easy pairs as these had high accuracy regardless of the spoken language.

In conclusion, the paper touched upon a wide variety of topics. Moreover, the consideration into the attributes of participants and consideration into how this would affect their performance with the scheme was a substantial addition. This provides a higher level of validity to the results and is something severely lacking from the literature assessed. However, the age of the study limits the applicability to modern day scenarios, and oversights in the development of the study such as the exclusion of hexadecimal encoding and the lack of information gained on the perceived usability of the scheme from users results in an incomplete study. In addition, the paper failed to consider the quantification of attacker strength and, therefore, failed to discuss the feasibility of these attacks.

**Kainda *et al.***[7] in 2009 investigated fingerprint representations and comparison methods in the context of using humans as the Out-Of-Band (OOB) channel for secure device pairing. This study was comprehensive and assessed a wide variety of concepts. This is one of the few papers in the literature that assessed the way to compare encodings alongside the representations themselves. For

comparison methods the paper looked into "Compare-and-Confirm", "Compare-and-Select", "Compare-and-Enter" and "Barcode" scanning. This set includes the most common ways of comparison according to available research. Alongside this, the paper reviews Numerical, Alphanumerical, Words, Sentences, Images, Melodies and Sound (Numerical/Alphanumerical). Again, this in comparison to similar literature is highly comprehensive.

In terms of empirical data gathered by the paper, they reviewed the time completed to asses the representation and the number of security and non-security related errors. Furthermore, the authors also collected user opinion on ease-of-use, satisfaction and confidence of the reviewed schemes. The overall schemes were then quantified using a metric known as the Single Usability Metric (SUM)[8]. This provides an elementary way to rank and compare the schemes, thus improving the potential validity of the results. However, the one apparent downside to the experimental side of the study is the 30 participants enrolled. This is one of the smallest sizes of enrolment in the literature and thus, puts a possible limit on the reliability and applicability of the obtained results. This limitation was, however, supplemented by a consideration into the demographics of the participants enrolled. The participants were evenly split between male and female alongside an age range of 18 to 75.

Overall, if usability is the only consideration the paper ranked the comparison methods: Compare-and-Confirm, Compare-and-Enter, Compare-and-Select and Barcode. If, however, the consideration of security alongside usability is required the best is: Compare-and-Enter, Barcode, Compare-and-Confirm and Compare-and-Select. Research has been limited in this area of modes of comparison, however, the results from this paper are consistent with the limited alternative literature. In terms of encoding schemes, the paper ranked Numerical, Alphanumerical and Words as the top three encoding schemes (These were all with Compare-and-Confirm). This recommendation took into consideration the usability and security of the scheme and is relatively consistent with other research where numerical and English language based encoding schemes seem to be the most effective overall.

In conclusion, this paper provided a highly comprehensive review of all the aspects of human fingerprint verification, where it reviewed all the elements involved in the process. Limitations of the paper involved the inconsistent and low range of entropy (20-bits to 40-bits)alongside the size of the participants enrolled for the study.

**Ersin Uzun *et al.***[9] exclusively reviewed comparison methods. This is performed in the context of "Secure device pairing" and thus some comparison schemes are tested that are not relevant in a fingerprint comparison context, for example "Choose-and-Enter" (A participant chooses a passphrase from an available list and enters it into the other device). Thus, only relevant comparison scheme and their results have been extracted and compared.

The relevant comparison methods tested were, "Compare-and-Confirm", "Compare-and-Select" and "Compare-and-Enter". These are almost exactly the same encoding schemes assessed in the relevant research literature. The paper also had two rounds of experiments including 40 participants each where changes were made to UX and format between these studies. Due to the "Secure device pairing" context, consideration needs to be made into the actual strings being encoded. Due to secure device pairing commonly utilising Short Authentication Strings (SAS) the paper only reviewed encodings of 4 digit long strings. This, therefore, limits the applicability of the results in comparison to other fingerprint encoding literature. Between the two rounds, the authors dropped Confirm-and-Enter due to its low usability and over similarity to a similar scheme called "Copy" (used for passphrase verification). The low rating for Confirm-and-Enter is in contrast with research performed by Kainda *et al.*[7] where they ranked Confirm-and-Enter as best method overall. With changes to the UX of Compare-and-Confirm backed up by external research [10][11] the second round of research was performed with a new set of participants. The alterations took Compare-and-Confirm from 20% to 0% fatal error rate (FER). The same trend for Confirm-and-Select with it reducing its FER from 12.5% to 5%. The authors, therefore, recommended "Compare-and-Confirm" as the overall best choice for comparison methods. Comparing this research to the work done by Kainda *et al.*[7]; this is consistent with their findings where they ranked "Compare-and-Confirm" as superior to that of "Compare-and-Select" in all regards.

Overall, the paper provides a sizeable insight into suitable methods of comparison, this is due to the cross over between fingerprint comparison and secure device pairing. This paper is limited in terms of the very short strings compared, the different overall research aim and the size of the number of participants used in the studies. In terms of overall recommendations, the results were consistent with that of alternative research with the exception of recommendations regarding "Compare-and-Enter". This, however, may be due to the differing use-cases of the comparison methods in the two respective studies.

**Dechand *et al.***[6] empirically investigated the usability of 4 distinct textual representations evaluated with an experiment involving a total of 1047 participants. The textual representations were: Alphanumeric, Numeric, Words and Sentences. They assessed the number of attacks missed with each scheme alongside results from a questionnaire on the participants preferred scheme and perceived usability.

The paper touches upon issues with decentralised methods of identification such a PGP's Web of Trust and the problems these solutions have with user adoption. These points are made in an attempt to validate the requirement for manual comparison of key based fingerprints. This is a common theme that appears in the majority of the reviewed papers. The paper has defined the upper and lower bound costs of the attacker's resources and strength as $610K to $16B, with an ability to control 80-bits of the fingerprint. This in comparison to other

papers is high and is almost encroaching into the realm of a highly sophisticated attacker. Therefore, the lack of consideration for the lower-end of the attack resource spectrum can be considered a limitation.

Summarised findings from the paper state that conventional encodings such as Hexadecimal and Base32 perform worse than all other alternatives in a realistic threat model with over 10% of users failing to detect an attack on these encoding schemes. The recommendations of the authors are to replace these encoding schemes with Words or Sentences due to their very high success rate and high usability scores. The performance of the Alphanumerical encoding schemes is relatively consistent with the literature, however, it is hard to say conclusively due to the small number of total schemes assessed making it hard to place alphanumerical encodings in an overall rating.

**J Tan *et al.***[12] work is very similar to that of the research already discussed. Eight distinct fingerprint representations were tested with over 661 participants. Each representation was tested using "Compare-and-Confirm" and "Compare-and-Select". The small difference with this paper is the inclusion of two novel graphical encodings Vash[2] and Unicorns[3]. The compared schemes were: Hexadecimal, Alternating vowel/consonants, Words, Numbers, Sentences, OpenSSH visual host key, Vash and Unicorns.

The target security level chosen was an entropy of 160-bits. This is a very high level of security in comparison to alternative literature. Alongside this, the paper defined attack strengths of an average $2^{60}$ with other specific use-cases for lower and higher attacker strength. Users were recruited via MTurk and assigned an experiment attempting to emulate a real-life condition (Comparison of a fingerprint on a business card). The real-world accuracy was one unique element that this paper concentrated on, this was portrayed through the authors' attention to detail when designing the experiment.

Overall, results from the experiments emulated previously discussed literature. Textual representations faired effectively well with Words and Sentences again being some of the best options overall. Performance of graphical representation was mixed with OpenSSH having performance matching that of a strong textual mapping while Unicorns having the worst performance overall. This performance with graphical representations is consistent with the available literature apart from the positive performance of OpenSSH. The results of different modes of comparison were highly consistent with other studies where again the recommendation was to not use "Compare-and-Select" as its performance across the board was below even basic levels of security.

**M. Shirvanian *et al.***[13] produced further work in the context of secure messaging pairing. This paper was unique for a number of reasons. First was to consideration for "remote-vs-proximity" pairing where, this is the first consid-

[2]https://github.com/thevash/vash
[3]https://unicornify.pictures/

eration of this aspect found in the literature. Second, was their studies aim into investigating fingerprint security and usability in a end-to-end encryption context. The study compared Numerical, Visual and Auditory representations. This is a standard setup for similar studies in the literature. QR code verification was also added to the proximity tests. They split the variance of attack severity into three levels; one character change, one block change and changes to the entire fingerprint.

The findings from the paper showed a high false negative rate for all the schemes, this is an aspect also missing from the literature. Alongside this, results for usability were lower in a remote setting for all the schemes where the author comments on the expected nature of this result.
The paper measured the number of successful attacks in the False Accept Rate (FAR). With this in mind images were the most secure method of authentication in the remote setting, but voted as the method with the worst usability. This is highly inconsistent with previous work in this area. However, this could be due to the unique setting of remote verification resulting in distinct results from user studies. Without further work in this area it is difficult to conclusively validate these results. In the proximity setting the results were negligible where at most 2.67% of attacks were successful. In summary, this paper provides a unique look into the real-world aspects of fingerprint verification in a end-to-end context providing a unique viewpoint on the usability and possible security of the main fingerprint encoding schemes.

### 1.1.1 Topic conclusion

To provide a visual comparison the accuracy results of the papers have been provided in Table 1. Each paper used a unique, and, therefore, have been translated into overall accuracy. Due to some of the differing characteristics of the papers ultimatly affecting results, Table 2 has also be provided. This table is useful to compare the health of the study and chosen sizes that may affect the accuracy and reliability of the results.

**TODO:** Usability table

In conclusion for this section, there has been extensive work in the investigation of the usability of textual and visual encoding schemes alongside methods of comparison. There are mixed opinions about visual representations, however, the majority defines them as easy to use but severely lacking in security. Research on textual representations, however, is in agreement; natural language based encodings such as words or sentences provide some of the highest usability and security over the baseline performance of alphanumerical encodings. The literature also concurs on the use of Compare-and-Confirm as the best method of comparing fingerprints overall, with Compare-and-Select being highlighted for its poor security and usability.
The starkest limitations of the literature are the range of participant and en-

| **Scheme** | Hsu-Chun[1] | Kainda[7] | Dechand[6] | Tan[12] | M. Shirvanian[13] |
|---|---|---|---|---|---|
| Hexadecimal | | | 89.56% | 79.00% | |
| Numerical | | 100.00% | 93.66% | 65.00% | 97.33% |
| Base32 | 86.00% | 90.00% | 91.50% | | |
| Words | 63.00% | 100.00% | 94.25% | 94.00% | |
| Scentences | | 100.00% | 97.01% | 94.00% | |
| Alternating[4] | | | | 83.00% | |
| Chinese Symbols | 59.00% | | | | |
| Japanese Symbols | 57.00% | | | | |
| Korean Symbols | 54.00% | | | | |
| Random Art | 94.00% | | | | |
| Flag | 50.00% | | | | |
| T-Flag | 85.00% | | | | |
| Flag Ext. | 88.00% | | | | |
| OpenSSH‘ | | | | 90.00% | |
| Unicorns | | | | 46.00% | |
| Vash | | | | 88.00% | |

Table 1: Accuracy comparison of textual and graphical encoding schemes

| | Hsu-Chun[1] | Kainda[7] | Dechand[6] | Tan[12] | M. Shirvanian[13] |
|---|---|---|---|---|---|
| Attacker Strength[5] | $\sim 2^{28}$ | $\sim 2^{40}$ | $2^{80}$ | $2^{60}$ | $\sim 2^{242}$ |
| Entropy Range | 22-28 bits | 20-40 bits | 122 bits | 128 bits | 160-256 bits |
| No° Participants | 436 | 30 | 1047 | 661 | 25 |

Table 2: Paper attribute comparison

coding entropy, where at its worst only 22-bits of entropy were tested. Another limitation is the scant amount of research solely investigating modes of comparison, where the literature only contains two papers with this aim. This is, therefore, an area that requires further research. A potential research avenue could be the analysis of comparison methods in unique scenarios, one example of this could be the idea of "remote-vs-proximity" proposed by M. Shirvanian *et al.*[13]. Even with these limitations, as a whole the literature on this topic has provided a relatively comprehensive and coherent view on the usability of textual and visual hash encoding schemes with relatively clear recommendations on methods of comparison and suitable encoding schemes.

## 1.2 Fingerprint representation schemes

Another area of research is investigations into the actual physical encodings of the hash digest. This section will briefly discuss the current research available on the creation and security of actual encoding schemes. The actual details of the operation of the schemes are outside the scope of this literature review, therefore, minimal attention will be allocated to these details.

Some of the oldest preliminary work into visual encoding schemes was performed by **Adrian Perrig *et al***[2]. in the creation of their scheme "Random Art" in 1999. The motivation for creating such a scheme was the perceived flaws in the ways humans verify and compare written information. As mentioned in previous sections visual encoding schemes have been shown to have mixed success, with low security being one of their most alarming flaws. This research laid the foundation for further work in analysing the security of visual encoding schemes.

Further research into the creation of unique visual hash schemes have been performed by **C Ellison *et al.*** [3] (Flag), **Yue-Hsun Lin *et al.***[4] (T-Flag) and work by **M. Olembo *et al.***[14]. Each publication has provided a new way to visually represent a key fingerprint. Alongside the academic literature, there are more informally presented methods of visual fingerprints such as Unicorns[6] and Robots[7]. This list is by no means exhaustive but is used to depict the amount of research and work invested into graphical hash representations.

One paper of note is the preliminary work performed by **D Loss *et al.***[15] in their *"An analysis of the OpenSSH fingerprint visualization algorithm"* where their aim was to spur on further research with their initial findings into the security of the OpenSSH scheme. The authors claim that the use of the algorithm in OpenSSH is only heuristically defined and there is a need for a formal proof of its security.
The paper proposed a number of ways to generate similar fingerprints. The

[6] https://unicornify.pictures/
[7] https://github.com/e1ven/Robohash

methods proposed were: Naive brute force, Graph Theory, and brute force of a full visual set. They were only able to produce only very basic results and have proposed a large amount of potential further work. Since the paper's publication in 2009, there seems to have been no research building on the work of the authors. This highlights a current gap in the available literature.

Minimal research has also focused on basic textual fingerprint representations and their respective security. Work by **A. Karole** and **N. Saxena**[16] looked into ways to improve the security of a textual representation. This research aim was to improve the secure device pairing process of comparing two numerical values. The devices used (Nokia 6030b; Mid-range devices at the time of publication) and the SAS compared results in findings that are not directly applicable in a fingerprint comparison context.

A more specific subsection of textual fingerprints is the use of words and sentences to encode hash digests. Some of the first work in this area was produced by **Juola** and **Zimmermann** [17] and their work in generating a word list where phonetic distinctiveness was prioritised. Each word is mapped to a single byte. The unique aspect of the word list is the separation of "even" and "odd" words where "even" byte positions are sample from the even-list and "odd" from the odd-list. This effectively creates two sub-word lists. The maximisation of linguistic distinctiveness of these word lists were maximised through the use of a Genetic algorithm. The paper also includes a study on effective measures of "linguistic distances" of words and provided an in-depth discussion into these areas.

Overall the paper provides a foundation for formalising the creation of effective wordlists. A limitation is the lack of empirical data gathered on the performance. However, this was later evaluated in work by Dechand *et al.* [6] and shown to be an effective encoding scheme.

Other research of note is work by **M. Goodrich *et al.***[18] called *Loud and Clear: Human-Verifiable Authentication Based on Audio.* As the name suggests the authors were researching ways to improve current methods of secure device pairing. The unique aspect of this work is the use of a Text-to-Speech system reading out syntactically correct English sentences. The sentences are based on a MadLibs[8] where static placeholders were replaced with potential words.
The work into a potential wordlist can be seen as an extension to the work performed by Juola and Zimmermann[17] as they aimed to emulate the techniques used in PGPfone. The paper's finding are limited by the lack of empirical data backing up claims made by the author as the systems performance and security are only theoretically assessed.

Aside from this research, there have been further informal implementations of fingerprint encodings. The first being by **Michael Rogers**[9]. Rogers' implemen-

[8] https://en.wikipedia.org/wiki/Mad_Libs
[9] https://github.com/akwizgran/basic-english

tation is a program designed to map fingerprints to pseudo-random poems. This implementation was again, empirically evaluated by Dechand *et al.*[6]. Older work by **N. Haller** with the S/KEY[19] shows the implementation of a system designed to represent a hash as a series of six short words. However, this system is designed for a one-time-password purpose and only provides word mappings for basic human usability of the password and not within a fingerprint verification context. Therefore, the wordlist has not been designed with pronounceability in mind.
A very recent implementation of a word list can be found in Pretty Easy Privacy (pEp) implementation of TrustWords[10]. The unique aspect of TrustWords is its mapping of a single word to 16-bits. In comparison to other literature, this is the highest number of bits-per-word seen. Full mappings (no duplication of words) would, therefore, require $2^{16}$ words in the dictionary and arguably is higher than most users vocabulary. this deviation from the norm has not been currently backed up by research. Moreover the main RFC documentation still remains in a draft stage and states *"It is for further study, what minimal number of words (or entropy) should be required."*. These aspects clearly highlight on a gap in the current literature.

### 1.2.1 Topic conclusion

In conclusion to this topic, the current research has primarily focused on the research and creation of visual representations. Research for textual fingerprints is fragmented and incomplete with work Juola and Zimmermann [17] and M. Goodrich *et al.*[18] providing meaningful research to build upon in terms of word a sentence based encodings. The fragmentation of this research leaves room for further work into this topic area. Alongside this, findings from the previous sections research shows that human language based encodings provided the best usability and, therefore, should be a target for further research looking to improve upon their security and usability.

## 1.3 Fingerprint representation attacks

This area of research studies ways to physically execute attacks on fingerprint encoding schemes. This differs from previously examined work due to papers discussing the performance and fallibility of encoding schemes simulated the attack without consideration for how the attack would be performed. Research in this area is scant, with lots of research attention being directed towards the security of Man-in-the-Middle (MITM) attacks and not the encoding schemes themselves.

Research in 2002 by **Konrad Rieck**[20] is the first formalisation of attacks on fingerprint representations. The paper titled *"Fuzzy Fingerprints Attacking*

[10] https://tools.ietf.org/html/draft-birk-pep-trustwords-03

*Vulnerabilities in the Human Brain"* aimed to look into ways users check hexadecimal encoded OpenSSH fingerprint representations. The author created an elegant way to 'weight' more important chunks of the digest. The bytes furthest to the right and left of the digests provided the highest weight. The weight was the smallest in the centre of the digest. This provides a way to score digests and determine the best partial collisions found. For example with the target fingerprint: `9F23` a partial match `9313` is given a score of 45% even though only two characters were matching. This is due to the weightings.

The paper contains an implementation with a "1.2GHz CPU" being able to obtain 130,000 H/s (With MD5). In comparison to this, a mid-range Intel i5-3320M CPU can today obtain 111,700,000 MD5 H/s. This shows that the results obtained from the paper are significantly outdated. However, even with the low hash rate, the author was able to obtain some promising results. Figure 1 contains the best example used.

```
TARGET: d6:b7:df:31:aa:55:d2:56:9b:32:71:61:24:08:44:87
MATCH:  d6:b7:8f:a6:fa:21:0c:0d:7d:0a:fb:9d:30:90:4a:87
```

Figure 1: Best match obtained after a few minutes of hashing

Overall the paper shows an interesting way to create partial fingerprint matches but is not quantified by any empirical evidence gathered on real world users. This, therefore, highlights on gaps in the coverage of this literature.

The only other relevant research on this topic is the work by **M Shirvanian *et al.***[21] and their paper *"Wiretapping via Mimicry: Short Voice Imitation Man-in-the-Middle Attacks on Crypto Phones"*. Further research in the area of "human voice impersonation" has received lots of attention [22][23][24]. This paper was chosen over other alternatives due to is specific use of encoding schemes in its evaluation.

In this paper, the authors develop a way to impersonate users when authenticating Short-authentication-Strings (SAS) in pairing of Crypto-phones. To achieve this impersonating they propose two methods: "Short voice reordering attack" where an arbitrary SAS string is recreated by re-ordering snipped obtained from eavesdropping a previous connection and "Short voice morphing attacks" whereby the use of previously eavesdropped audio snippets the attacker can morph their own voice to match that of the victim. With these methods, they aimed to attack encodings of Numbers, PGP word list (previously discussed work by Juola and Zimmermann [17]) and MadLib (M. Goodrich *et al.*[18] work also previously discussed). The effectiveness of these attacks were evaluated with a study involving 30 participants.

Results from the paper show the effectiveness of these methods. Compared to the baseline of the attacker's voice replacing the victim where this performed with a ∼18% success rate. Morphing gained an overall success rate of 50.58%

and Reordering a very impressive 78.23% success rate. Showing that these attacks provide an improvement on top of the naive implementation.

One of the biggest limitations addressed by the authors was the reduction in success rates as the size of the authentication string grew. The morphing and reordering attacks become increasingly ineffective as the user has more time to detect imperfections. This is not quantified by the author and the extent of this degradation is never empirically discussed. Therefore, the results from this study are only effective and applicable in a SAS context.

### 1.3.1 Topic Conclusion

Overall the literature for this subtopic remains sparse and incomplete. Further suggested work could look into the feasibility of generating partial collisions for all textual representations alongside quantified effectiveness on users. With the possibility to concentrate on a few selected implementations. The work would aim to focus on the various physical methods used and their feasibility. This is one area the previous literature has failed to cover and has only theoretically quantified attacker strength without consideration for the actual real-world cost of these attacks.

# 2 Overall Summary

**TODO:** Create an overall summary of all the gaps identified

# 3 Research Questions

**TODO:** - Backup choice of questions up using my previous discussion.

# References


[1] H.-C. Hsiao, Y.-H. Lin, A. Studer, C. Studer, K.-H. Wang, H. Kikuchi, A. Perrig, H.-M. Sun, and B.-Y. Yang, "A study of user-friendly hash comparison schemes," in *2009 Annual Computer Security Applications Conference*. IEEE, 2009, pp. 105–114.

[2] A. Perrig and D. Song, "Hash visualization: A new technique to improve real-world security," in *International Workshop on Cryptographic Techniques and E-Commerce*, 1999, pp. 131–138.

[3] C. Ellison and S. Dohrmann, "Public-key support for group collaboration," *ACM Transactions on Information and System Security (TISSEC)*, vol. 6, no. 4, pp. 547–565, 2003.

[4] Y.-H. Lin, A. Studer, Y.-H. Chen, H.-C. Hsiao, L.-H. Kuo, J. M. McCune, K.-H. Wang, M. Krohn, A. Perrig, B.-Y. Yang *et al.*, "Spate: small-group pki-less authenticated trust establishment," *IEEE Transactions on Mobile Computing*, vol. 9, no. 12, pp. 1666–1681, 2010.

[5] M. Blum, "Coin flipping by telephone," *Proc. of COMPCON, IEEE, 1982*, 1982.

[6] S. Dechand, D. Schürmann, K. Busse, Y. Acar, S. Fahl, and M. Smith, "An empirical study of textual key-fingerprint representations," in *25th {USENIX} Security Symposium ({USENIX} Security 16)*, 2016, pp. 193–208.

[7] R. Kainda, I. Flechais, and A. Roscoe, "Usability and security of out-of-band channels in secure device pairing protocols," in *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, 2009, p. 11.

[8] J. Sauro and E. Kindlund, "A method to standardize usability metrics into a single score," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005, pp. 401–409.

[9] E. Uzun, K. Karvonen, and N. Asokan, "Usability analysis of secure pairing methods," in *International Conference on Financial Cryptography and Data Security*. Springer, 2007, pp. 307–324.

[10] J. Palmer, "Attentional limits on the perception and memory of visual information." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 16, no. 2, p. 332, 1990.

[11] R. Hammer, T. Hertz, S. Hochstein, and D. Weinshall, "Category learning from equivalence constraints," *Cognitive Processing*, vol. 10, no. 3, pp. 211–232, 2009.

[12] J. Tan, L. Bauer, J. Bonneau, L. F. Cranor, J. Thomas, and B. Ur, "Can unicorns help users compare crypto key fingerprints?" in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 3787–3798.

[13] M. Shirvanian, N. Saxena, and J. J. George, "On the pitfalls of end-to-end encrypted communications: A study of remote key-fingerprint verification," in *Proceedings of the 33rd Annual Computer Security Applications Conference*. ACM, 2017, pp. 499–511.

[14] M. M. Olembo, T. Kilian, S. Stockhardt, A. Hülsing, and M. Volkamer, "Developing and testing a visual hash scheme." in *EISMC*, 2013, pp. 91–100.

[15] D. Loss, T. Limmer, and A. von Gernler, "The drunken bishop: An analysis of the openssh fingerprint visualization algorithm," 2009.

[16] A. Karole and N. Saxena, "Improving the robustness of wireless device pairing using hyphen-delimited numeric comparison," in *2009 International Conference on Network-Based Information Systems*. IEEE, 2009, pp. 273–278.

[17] P. Juola, "Whole-word phonetic distances and the pgpfone alphabet," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 1. IEEE, 1996, pp. 98–101.

[18] M. T. Goodrich, M. Sirivianos, J. Solis, G. Tsudik, and E. Uzun, "Loud and clear: Human-verifiable authentication based on audio," in *26th IEEE International Conference on Distributed Computing Systems (ICDCS'06)*. IEEE, 2006, pp. 10–10.

[19] N. Haller, "The s/key one-time password system," 1995.

[20] K. Rieck, "Fuzzy fingerprints attacking vulnerabilities in the human brain," *Online publication at http://freeworld. thc. org/papers/ffp. pdf*, 2002.

[21] M. Shirvanian and N. Saxena, "Wiretapping via mimicry: Short voice imitation man-in-the-middle attacks on crypto phones," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 868–879.

[22] D. Mukhopadhyay, M. Shirvanian, and N. Saxena, "All your voices are belong to us: Stealing voices to fool humans and machines," in *European Symposium on Research in Computer Security*. Springer, 2015, pp. 599–621.

[23] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen, "You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2017, pp. 183–195.

[24] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.