Usability evaluation of voiceprint authentication in automated telephone banking: Sentences versus digits

Usability Evaluation of Voiceprint Authentication in Automated Telephone Banking: Sentences versus Digits

Nancie Gunson[a,*], Diarmid Marshall[a], Fergus McInnes[a], Mervyn Jack[a].

[a] Centre for Communication Interface Research, The University of Edinburgh, Alexander Graham Bell Building, King's Buildings, Edinburgh, EH9 3JL, UK.

* Corresponding author. Tel.: +44-131-651-7120; fax: +44-131-650-2784; E-mail address: Nancie.Gunson@ccir.ed.ac.uk (N. Gunson).

Keywords: voiceprint; authentication; verification; usability; biometrics; dialogue design; automated telephony.

Abstract

This paper describes an experiment to investigate the usability of voiceprints for customer authentication in automated telephone banking. The usability of voiceprint authentication using digits (random strings and telephone numbers) and sentences (branded and unbranded) are compared in a controlled experiment with 204 telephone banking customers. Results indicate high levels of usability and customer acceptance for voiceprint authentication in telephone banking. Customers find voiceprint authentication based on digits more usable than that based on sentences, and a majority of participants would prefer to use digits.

1. Introduction

This paper describes an experiment to investigate the usability of voiceprints for customer authentication in automated telephone banking.

Most financial institutions worldwide now offer some form of telephone banking to their customers. Such services allow customers to carry out a range of banking tasks - from simple balance enquiries to opening a new account - over the phone from the location of their choice. The hours of availability are normally considerably longer than branch opening hours, and in many cases are 24 hours a day, 7 days a week.

The majority of these services are at least partly automated, which means callers hear a series of automated messages (typically, human speech recorded by a professional voice talent) and are invited to respond using either their voice or the keys on their telephone keypad, depending on the service design.

The level of automation varies between services. Some use automation solely for call routing or customer identification and verification purposes. Customers are then transferred to the appropriate bank staff in call centres. An increasing number of banks, however, now offer fully automated systems, in which customers can carry out simple banking tasks such as balance enquiries and funds transfers over the phone without the need to speak to a human agent. Access to this type of service is based on knowledge-based authentication ("what you know"), typically a secret password or PIN known only to the customer.

Employed properly, PINs and passwords are a powerful tool in the security of services (O'Gorman, 2003). However, for users who have to remember multiple passwords and PINs across a number of applications, the cognitive burden of remembering each can become significant. A common solution amongst users is to use the same password for a number of applications (Adams and Sasse, 1999; Dhamija and Perrig, 2000) which has obvious security risks. Users are also known to

choose memorable (and therefore low security) passwords (Adams and Sasse, 1999; Bishop, 2005; Yan *et al*, 2004) and even when given careful instructions on how to choose a secure password a significant proportion do not comply, presumably for the sake of convenience (10% of students in a study by Yan *et al*, 2004, together with 32% of those in the control group who were given 'standard' instructions).

Studies suggest that PINs can be more difficult to remember than passwords (Sasse *et al*, 2001), and anecdotal evidence of customers using their date of birth as their PIN or writing it down is common.

Increasingly, therefore, alternative or supplementary security mechanisms are being sought, particularly within the financial services industry where recent years have seen a growth in remote fraud (Hiltgen *et al*, 2006). 'Two-factor' authentication using physical tokens such as card readers ("what you have") in addition to the traditional password/PIN are one possibility, since any potential fraudster must be in possession of both the relevant password *and* the device (Weir *et al*, 2009). However, for the true user the disadvantage is the potential inconvenience of having to have the device with them when they want to access the service. Moreover, the token can be physically mislaid or lost.

Another possibility is the use of *biometrics,* a range of technologies that use a distinguishing physical or behavioural feature to identify or verify an individual within automated systems. Examples of physical features include voice (Naini *et al*, 2009; Yoma *et al*, 2008), fingerprints (Tan and Schuckers, 2010; Yamagashi *et al*, 2008) and iris patterns (Arivazhagan *et al*, 2009). Behavioural features include gait (Nandini and Ravi Kumar, 2008) and handwriting (Nanni and Lumini, 2006; Scheidat *et al*, 2009).

Biometric traits ("who you are") have the advantage over both knowledge and token-based authenticators in that that they cannot be mislaid or forgotten. In addition, although no biometric system is 100% accurate, they are also considered difficult to copy or forge (Jain *et al*, 2006). This combination of factors has led to substantial interest in the technology in the financial services industry. Examples include the use of iris recognition at the ATM (Coventry *et al*, 2003a) where research found good usability and user acceptance, although lack of business benefits and high costs ultimately prevented the deployment of the technology (McClue, 2003). More recently, a major Australian bank adopted a voice verification system for their telephone banking service. In this system, once callers have completed enrolment they can have their identity verified by speaking their ID number and date of birth over the phone "without the need to remember cumbersome passwords and PINs"[1].

In a telephone banking service voiceprints are the obvious choice of biometric. Advantages include that there is no requirement for specialist equipment beyond the microphone and speaker inherently present, capture of the biometric is physically non-intrusive (in contrast to, for example, iris or fingerprint scans) and, moreover, the interaction has the potential to be highly intuitive since it is based on the familiar act of spoken conversation.

The implementation of voiceprints for authentication involves two main stages: *enrolment* and *verification*. Enrolment is a one-off process in which speech samples

[1] http://www.nab.com.au/wps/wcm/connect/nab/nab/home/Personal_Finance/21/Speech+Security/?ncID=ZBA, last accessed March 2010

are collected from the (validated) user in order to create a biometric template of their identity. This template is then stored for subsequent comparison against during future *verification* attempts, in which speech samples are again captured 'live' from the user (although less data are required at this stage than during enrolment).

Two types of voice verification are available; *text dependent*, where the user is required to speak a specific phrase or token during the verification process (one that was used during the enrolment process) or *text independent*, where there is no such restriction on the input, but achieving high accuracy is more difficult (Bimbot *et al*, 2004; Reynolds, 2002). Considerably more data is required from the user to achieve the flexibility of text-independent verification, during both enrolment (where most commercial systems require around 30 seconds minimum) and verification. This can be tedious for users and in combination with a tendency to poorer accuracy weighs against use of this mode in commercial applications.

A potential drawback of the text-dependent approach, however, is the risk of 'spoofing' i.e. fraudster attack using recordings of a real customer using the system, either through playback or synthesis[2]. However, aside from the not inconsiderable difficulty of covertly recording someone using the system (e.g. concealing a microphone in a suitable location at a suitable time), a number of methods are now available which reduce the risk of this type of attack significantly. These are primarily based on *liveness* detection and include signal processing techniques, which analyse the spoken input for acoustic effects that indicate it is a recording. However, the ability to use ever more sophisticated techniques to manipulate recordings and disguise their source means other methods are increasingly common. One approach is to compare the current verification bid with N previous (accepted) bids in addition to the comparison with the voiceprint created during enrolment. Since some variability is inherent in speech, if a verification bid is found to be *too* perfect a match to any of the stored samples, the suspicion must be that this is a spoofing attack. Another form of liveness detection involves the use of *challenge-response* protocol, sometimes known as *text-prompted* mode, in which the user enrols using several different voiceprint tokens and may be asked for any one of these during verification, reducing the predictability of the interaction and therefore the risk of fraudsters being in possession of the correct recording (O'Gorman, 2003). Note that this method assumes the system employs speech *recognition* as well as verification software, checking that what was said matched what was prompted for. Some systems based on digit strings, moreover, increase the defence by requesting the digits used during enrolment in random sequences at the time of verification.

There are therefore substantial barriers to spoofing attacks. Consequently, in a telephone banking context where accuracy is of prime importance, a text-dependent strategy is considered more appropriate than a text-independent one provided suitable anti-spoofing measures are put in place.

Previous research relating to the use of voice recognition technology in the banking sector has tended to focus on the development of the core technology, evaluating the

---

[2] A human mimic could also be used to spoof a voiceprint but this is difficult to achieve. An individual's voiceprint reflects the size and shape of their vocal tract (extending through the throat, mouth and nose). To make the attack viable, therefore, the mimic must have physiology that is similar to that of the true user. In this context, close relatives of the same gender (in particular, identical twins) pose a greater risk than professional fraudsters.

technical performance of different verification algorithms (Khan *et al*, 2010; Naini *et al*, 2009; Yoma *et al*, 2008). These studies, reflecting current biometric evaluation standards (ISO/IEC JTC, 2007), concentrate predominantly on *false match rates* (FMR), which measure the level of incorrect acceptance of impostors, *false non-match rates* (FNMR), which measure the incorrect rejection of true users, and the trade-off between the two.

There are also several examples of papers that discuss the issues surrounding biometric technology more generally, the pros and cons of each type (including voiceprints) and possible frameworks for their implementation (Coventry, 2005; O'Gorman, 2003; Jain *et al*, 2006; Venkatraman and Delpachitra, 2008). There have also been a number of European and UK government initiatives (BioVisioN[3] and BIOSECURE[4] and Biometrics Working Group) that have sought to promote the use of biometrics, including voiceprints, and which have highlighted the importance of a user-centred approach (BioVision, 2003).

However, there are few studies available that provide actual empirical data on user reaction to voiceprint technology (or indeed the other types of biometric). This is important, since as O'Gorman (2003) notes, the usability of security measures is key for customer acceptance; if the authenticator is difficult to use it will not be used, or will not be used properly, which may present vulnerabilities.

One exception is a study by Toledano *et al* (2006) in which the usability of fingerprint, voice and signature verification was measured in a controlled experiment with 43 users. The evaluation encompassed three different aspects of usability; *effectiveness*, *efficiency* and *satisfaction*. User satisfaction in each case was measured using attitude scales in Likert format (Likert, 1932; Rossi *et al*, 1983). This is an important example of an evaluation of voiceprint technology in which the reaction of the user was considered. However, the on-screen application employed in the experiment was designed solely for the purpose of comparing the different types of biometric. Moreover, users' satisfaction with each technology was based in part on being asked to act as deliberate impostors – a non-typical experience for end-users.

This paper describes an experiment to investigate the usability of voiceprints for customer authentication within the context of a real-world application, an already established telephone banking system. Customer perceptions of two different text-dependent strategies for voiceprint authentication in telephone banking were investigated – one based on the use of Sentences as the voiceprint token and the other based on the use of Digits.

The rest of the paper is organised as follows. Section 2 gives an overview of the dialogue design with details of the two different voiceprint authentication strategies that were investigated. Section 3 describes the methodology used in the research, Section 4 the specifics of the experiment design and Section 5 its participants. Results are presented in Section 6, with main conclusions given in Section 7.

---

[3] BioVisioN - roadmap to successful deployments from the user and system integrator perspective, January 2002- May 2003. http://www.ist-world.org/ProjectDetails.aspx?ProjectId=c73fc0f5692b4ef8a79a4e0644a0211f, last accessed March 2010.

[4] Biometrics for Secure Authentication, BIOSECURE, June 2004 to September 2007. http://www.ist-world.org/ProjectDetails.aspx?ProjectId=83aa8f7b0a4b416fb19f61ed18ba7390 last accessed March 2010.

2. Dialogue Design

Each of the strategies investigated was set within the context of an already established automated telephone banking service, that of a major UK bank. The two versions of the service used in the experiment were based on the live service and differed only in the customer identification and verification (ID&V) dialogue. Both began with a welcome message, followed by capture of the caller's account number and sort code (either via speech or the telephone keypad). Using this design, once a valid account number and sort code have been given, the service can identify the customer and determine whether they have enrolled a voiceprint.

Two different types of voiceprint tokens were compared in the experiment:

Sentences: either branded or unbranded.

Digits: either random digit strings or telephone numbers.

The rationale for the choice of tokens was as follows. Commercial organisations considering the deployment of voiceprint authentication may be tempted to use branded sentences for marketing purposes. It is important, therefore, to measure any impact this may have on the usability and acceptance of the technology. The two types of digit strings were examined on the basis that random digit strings may appear more secure to users, but may be more difficult to use than telephone numbers (and therefore be less appealing overall). There may be a trade-off in this context between levels of perceived security and ease of use. Empirical data on both types of digit string is required in order to investigate this.

In order to use voiceprint authentication in either version of the service, customers must first enrol a voiceprint by providing some samples of speech. During the enrolment call, the customer is identified and verified using existing procedures. Identification involves capturing the customer's account number and sort code. Verification of the customer's identity is then carried out by requesting two (randomly selected) digits from a 6-digit secret number previously registered by the user and known only to them. ("Please say or key in the Xth digit of your secret number." followed in a separate stage by "...and the Yth digit."). Once successfully verified the customer is asked whether they would like to enrol a voiceprint.

On acceptance of this offer, the enrolment dialogue for Sentences involves saying three sentences twice each (any one of which may be requested during verification). For the Digits version, the enrolment dialogue depends on the type of number used. For random digits, as with Sentences, three different randomly generated eight-digit strings are repeated twice each (six utterances in total); verification may be on any one of these. For telephone numbers, where the verification token is fixed, slightly less data are required; three repetitions of the telephone number.

A total of 24 different sentences were employed in the experiment (12 branded and 12 unbranded). Examples of the branded sentences include "At <Bank> my voice is my password." and "My <Bank> account is secured by my voiceprint." Examples of unbranded sentences are "I'm using my voice as my password." and "My voice is my password in telephone banking." Participants were not given the sentences in written form, but were prompted for them within the service.

The authentication dialogues using voiceprints were as follows. In each case authentication based on the voiceprint *replaced* the existing, knowledge-based method of two randomly selected digits from a six-digit secret number.

Sentences: The prompt here is *"I have your voiceprint on record. In order to use your voiceprint please repeat the following sentence"* – followed by a readout of the sentence to be spoken. If authentication using the sentence is not successful, the caller is not explicitly told so but is asked to speak a second sentence; the prompt here is *"Thank you. Now please repeat this next sentence."*

Digits (random digits): The caller is asked to say a randomly generated eight-digit phrase consisting of four digits repeated, e.g. *"I have your voiceprint on record. In order to use your voiceprint please repeat the following string of numbers: one seven four eight one seven four eight."* If verification on the digit string is not successful, again there is no explicit mention of this; the caller is simply prompted for a second digit string, e.g. *"Thank you. Now please repeat this next string of numbers: zero six nine two zero six nine two."*

Digits (telephone number): The customer is prompted as follows: *"I have your voiceprint on record. In order to use your voiceprint please say your telephone number now."* Unlike the prompts in the other versions, the prompt here does not state what the number is, since it is assumed that the caller will know this. (Prompting explicitly for a specific telephone number before the caller was verified would compromise the customer's security and privacy, since it would allow anyone who had the account number and sort code to call the service and find out the customer's telephone number.) However, if the caller repeatedly gives a wrong number, the error recovery prompt attempts to clarify which number is required: *"In order to confirm your identity I need you to speak the telephone number that you used to enrol your voiceprint. Please say your telephone number now."* If verification on the telephone number is not successful, the caller is simply prompted to say the 11-digit number again: *"Thank you. Now please say your telephone number again."*

In each case, for authentication to be successful, the caller must both say the correct sentence or string of digits and have it successfully recognised, and have their voiceprint matched against that stored in the database. Successful verification leads to a variety of banking services as in the existing system e.g. account balances, recent transactions listings etc. If the caller has still not been verified after two utterances, the call is transferred to an advisor.

3. Methodology

The methodology employed in this research involves the use of a controlled experiment, in which two or more versions of the system under test, differing in some design characteristic, are experienced by participants. Typically, a repeated-measures design is employed in order to minimise the effects of between-subject variability. Participants experience each of the versions under test, with the order of presentation of the different versions balanced across the group to control for any order effects.

This experimental approach is complemented by an emphasis on achieving as much realism within the experimental setting as is possible, which helps to elicit results that are representative of real-life use. Participants are given detailed personal data as fictitious personae to use during the experiment, and are presented with realistic scenarios in which they are asked to attempt a task with the service that is typical of real-life use. This involves the use of a fully functional prototype.

Following each experience of the services, participants are asked to complete a usability questionnaire. The questionnaire employed in this research is a tool for assessing users' attitudes towards automated telephone services which was developed

and tested over a number of experiments (Dutton *et al*., 1993; Jack *et al*., 1993; Love *et al*., 1992) and has been widely used and adapted since (Davidson and Jack, 2004; Foster *et al*., 1998; Larsen, 2003, 1999; Morton *et al*., 2004; Sturm and Boves, 2005). The questionnaire encompasses cognitive issues (e.g. level of concentration required by users, and how stressful the service was to use), the fluency and transparency of the system (e.g. ease of use and degree of complication), system performance (e.g. the efficiency of the application and users' preferences for a human agent), and issues relating to the voice of the service (e.g. politeness and clarity). It consists of a set of 20 brief proposal statements, each with a set of tick-boxes on a seven-point Likert scale (Likert, 1932; Rossi *et al*., 1983) ranging from "strongly agree" through neutral to "strongly disagree". See Appendix A for a full listing. Statements in the questionnaire are balanced, positive and negative, to counteract the problem of response acquiescence set - the general tendency for respondents to agree with the statement offered. In order to analyse the results, responses to the questionnaire are converted into numerical values from 1 (most unfavourable) to 7 (most favourable) allowing for the polarity of the statements. Thus, for example, a ''strongly agree'' response to a negative statement is converted to a value of 1. On this scale, a (normalised) score over 4.0 represents a positive attitude; scores below 4.0 represent a negative attitude, with 4.0 the neutral point. Once the polarity of the results is normalised, each participant's overall attitude to the service is measured by taking the mean of these numbers across all of the items in the questionnaire. A measure of the overall attitude to the service can then be obtained by averaging all the questionnaire results for participants who experienced that service.

As well as providing an overall attitude rating, the mean scores for individual statements can also be examined to highlight any aspects of the dialogue design which were particularly successful or which require improvement. Finally, the results can also be analysed according to demographic groupings of participants (age, gender etc.) and any significant differences between groups can then be identified.

At the end of the session participants take part in a structured interview, which provides qualitative attitude data that can be very useful in explaining why participants responded in the ways they did, as well as quantitative data on participants' preference between variants of the design. Objective measures are also recorded as part of the experiment session including, for example, recognition accuracy and time taken to complete the task. The research methodology thus allows both subjective and objective data concerning the service under evaluation to be obtained.

4. Experiment Design and Procedure

Here, a mixed within-participant and between-participant experiment design was adopted. Each participant experienced enrolment and verification using both Sentences and Digits, with the order of experience of the two voiceprint types balanced across the cohort. Sentence and Digits type were varied between participants, so that approximately the same number of participants (a quarter of the cohort) experienced each possible combination of sentence type (branded or unbranded) and number type (random digits or telephone number)[5].

[5] Each participant experienced *either* the branded or unbranded versions of the sentences in one version, and *either* random digits or use of 'their' telephone number in the other version.

Each participant was given a fictitious persona, equipped with all the details required to enrol a voiceprint and to use the service. Before making any calls, participants were given a letter from the Bank introducing the use of voiceprints and saying that the next time they called they would be offered the opportunity to register a voiceprint.

Participants then made a total of eight telephone calls: four calls to each version of the service, with voiceprint enrolment in the first call to each version and verification by voiceprint in the other three calls. The task in the enrolment call was simply to enrol a voiceprint; in each of the other calls the task was to find out the balance of the account (which was varied automatically between calls). A usability questionnaire was completed after each of the eight calls. At the end of the session participants took part in a structured de-briefing interview.

Note that although the verifier engine[6] was running during participants' calls, for reasons of experimental control its decisions were at times overridden to ensure parity of outcome across the cohort. In a real service using voice authentication there would be some failures due to rejection by the verifier, because of insufficiently close correspondence to the customer's voiceprint (false non-match as described earlier). However, in order to set the decision thresholds appropriately extensive data capture is required. Well-designed systems are carefully tuned using data from real customers to balance the risk of false match and false non-match errors at a level appropriate to the application in question – typically, for example, a low false match rate being of prime importance in a banking system, outweighing the inconvenience of false non-match experienced by some users as a result. This type of analysis and tuning, however, was not the focus of this research. Here, the outcome of the verification process was controlled to ensure all participants experienced *both* success and failure during the course of their calls, enabling their attitude towards each to be measured in a controlled manner. The system was designed such that participants experienced verification 'success' in the majority of their calls, as they would in real life. The exception was in participants' final call to each version of the service, where verification 'failed' on the first attempt and a second utterance was requested from the user in all cases.

Participants could also, at any stage, experience rejection or misrecognition by the recogniser[7], which having already had its thresholds tuned on data from the existing service, was live during the experiment as it is in the real service (thus ensuring callers were saying the actual phrases they were prompted for). Rejection by the recogniser could be on the grounds of no or incorrect/incomplete input by the user, or could be a false rejection of a valid utterance.

[6] Nuance Verifier v3.0

[7] Nuance Recognizer v8.0

5. Participants

A cohort of 207 participants was recruited, from which a complete set of data were obtained from 204 (104 in Enfield and 100 in Nottingham, both UK). All were customers of the Case Bank. Three participants were unable to complete a full set of calls because of repeated misrecognition of digits (one in the telephone number group, two in the random digits group) and thus were excluded from the main analysis. These are discussed further in Section 6.6

A breakdown of the final data set by age group and gender is given in Table 1. The age groups chosen were designed to reflect the profile of the Bank customers represented in the recruitment database. The breakdown by age group and gender in each location was very similar.

	Age group 1 (18-44 years)	Age group 2 (45+ years)	Total
Male	51	36	87
Female	59	58	117
Total	110	94	204

Table 1: Participant Cohort by Gender and Age Group

6. Results

The mean usability scores obtained from the questionnaire for the enrolment and verification calls using each type of voiceprint are shown in Table 2. Results for both Sentences and Digits are shown broken down by type (branded or unbranded, random digit strings or telephone numbers), with each type experienced by approximately half of the 204 participants. Scores are also shown computed over the full set of 204 participants ('All Sentences' / 'All Digits'). Scores are on a scale from 1 (least favourable) to 7 (most favourable), with 4 as the neutral point.

Voiceprint Type	Enrolment Call	Verification Calls: First	Second	Third	Verification Mean
Branded Sentences (N=105)	5.24	5.66	5.69	5.61	5.65
Unbranded Sentences (N=99)	5.34	5.62	5.66	5.64	5.64
All Sentences (N=204)	*5.29*	*5.64*	*5.68*	*5.62*	*5.65*
Random Digits (N=103)	5.55	5.79	5.80	5.79	5.79
Telephone Number (N=101)	5.86	5.92	5.96	5.92	5.93
All Digits (N=204)	*5.70*	*5.85*	*5.88*	*5.85*	*5.86*

Table 2: Mean Usability Scores by Voiceprint Type and Call Number

Encouragingly, usability scores for both types of voiceprint were high, at above 5.0 on a 7-point scale. Separate statistical analyses were performed on the scores for the enrolment call and on the scores for the verification calls – as detailed in the subsections below.

6.1. Usability of the Enrolment Process

A repeated measures analysis of variance (ANOVA) was applied to the mean usability scores for the enrolment calls, with *voiceprint type* (Sentences or Digits) as the within-participants factor, and *age group*, *gender*, *location*, *order of presentation of versions* (Sentences first or Digits first), *sentence type* and *digits type* as between-participants factors. This showed a very highly significant main effect of voiceprint type ($p<0.001$), with enrolment using digits found to be more usable than enrolment using sentences. The interactions of voiceprint type with sentence type and with digits type were also highly significant ($p=0.004$ and $p<0.001$ respectively): the difference in scores between the two types of voiceprints (in favour of Digits) was greater amongst participants who had branded sentences than amongst those who had unbranded sentences (Figure 1), and greater for participants using telephone numbers than for those using random digit strings (Figure 2).

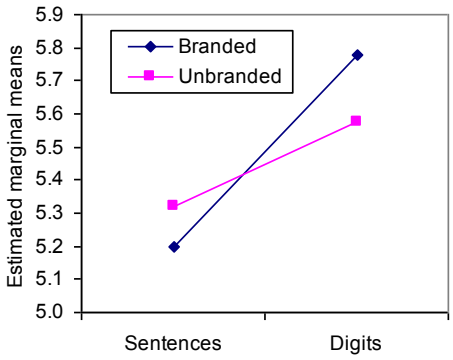

Figure 1: Enrolment usability - interaction between voiceprint type and sentence type

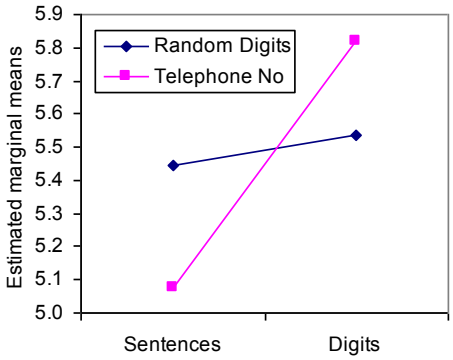

Figure 2: Enrolment usability - interaction between voiceprint type and digits type

The original repeated measures ANOVA (on enrolment usability for both types of voiceprints) was rerun for each of the 20 specific attributes in the usability questionnaire. The main effect of voiceprint type was significant at the 0.05 level for 17 of the 20 usability attributes; highly significant ($p<0.01$) for 13 of them. In each case enrolment using Digits was scored more highly than enrolment using Sentences. Particularly large differences (all of which were very highly significant; $p<0.001$) were found on some of the cognitive and fluency attributes, with enrolment using Sentences requiring much more *concentration* (Sentences 4.01, Digits 5.11) making customers feel more *flustered* (Sentences 5.00, Digits 5.63), *stressed* (Sentences 4.91, Digits 5.70) and *frustrated* (Sentences 5.34, Digits 5.82), and being found more *complicated* (Sentences 5.29, Digits 6.07) than enrolment using Digits. Large and very highly significant differences (again all at $p<0.001$) were also found for *ease of use* (Sentences 5.60, Digits 6.10), *willingness to use the service again* (Sentences 5.41, Digits 5.92), *need for improvement* (Sentences 4.87, Digits 5.45) and *enjoyment* (Sentences 4.86, Digits 5.38).

The interaction of voiceprint type and sentence type was significant ($p<0.05$) for eight attributes. In each case the pattern was similar to that found for the overall mean; participants who had unbranded sentences tended to give higher scores to the Sentences enrolment and lower scores to the Digits enrolment than those who had branded sentences. Similarly, the interaction of voiceprint type and digits type was significant ($p<0.05$) for 16 attributes. Participants using telephone numbers tended to give higher scores for Digits and lower scores for Sentences than those using random digits.

In view of the large differences in enrolment usability found between participants with random digits and those with telephone numbers, and (to a lesser extent) between those with branded and unbranded sentences, some further analysis of the mean usability scores was performed within the subsets of participants who were given particular types of digits and sentences, with the following results.

- The main effect of voiceprint type was found to be highly significant ($p<0.01$) in both sentence type groups i.e. participants rated the usability of Digits enrolment more positively than that based on Sentences, regardless of the Sentence type they experienced.
- In contrast, the main effect of voiceprint type (Digits>Sentences) was very highly significant ($p<0.001$) for participants who enrolled using telephone numbers, but not significant ($p=0.154$) for those using random digit strings. Participants who enrolled using random digit strings rated Digits enrolment similarly to that based on Sentences.
- Breaking down by both sentence type and digits type, the main effect of voiceprint type (Digits>Sentences) was very highly significant ($p<0.001$) in both telephone number groups (i.e. those with both branded and unbranded sentences). It was also significant, although less so ($p=0.011$), in the group who experienced random digit strings versus branded sentences. However, for those who experienced random digit strings with *unbranded* sentences, there was no significant difference in usability between the two. The mean within-participant difference in enrolment usability (Digits – Sentences) for the four sub-groups was 0.94, 0.52, 0.19 and -0.02 respectively.

In summary, the experiment showed that the enrolment process based on a telephone number was substantially more usable than one based on either branded or unbranded sentences, but enrolment using random digit strings was only slightly more usable than enrolment on sentences. In particular there was very little evidence of a difference in usability between the enrolment processes using unbranded sentences and random digits.

6.2. Usability of the Service with Voice Authentication

A repeated measures ANOVA was applied to the mean usability scores for the calls with voice authentication, with *voiceprint type* (Sentences or Digits) and *call number* (first, second or third call to this version of the service) as the within-participants factors. The between-participants factors included were the same as in the original ANOVA on the enrolment scores, i.e. *age group*, *gender*, *location*, *order of presentation of versions* (Sentences first or Digits first), *sentence type* and *digits type*.

Again, the main effect of voiceprint type was found to be very highly significant ($p<0.001$), with higher scores for Digits than for Sentences. There was also a highly significant interaction of voiceprint type with digits type ($p=0.002$), which was similar in form to that found for the enrolment call; participants who enrolled using telephone numbers gave higher scores for Digits and lower scores for Sentences than those who used random digit strings. However, the interaction of voiceprint type with sentence type did not approach significance on the verification calls ($p=0.738$), in contrast to the results on the enrolment calls reported in the previous section. The difference in scores amongst those with branded and unbranded sentences was weaker here, in calls involving only one or two sentence utterances, than in the enrolment call, which involved saying three sentences twice each.

There was no main effect of call number (p=0.075), although effects of call number were found on a few individual attributes. On *knowing what to do*, there was an upward trend across the three calls to each version of the service, as might be expected in view of the participant's increasing familiarity with the service over the sequence of calls – though the effect of call number was weak and was not quite significant at the 0.05 level overall. On *control*, there was a drop in scores from the second call (where verification was completed using just one utterance) to the third call (where an additional utterance was needed to complete verification). *Enjoyment* showed both a significant increase from the first to the second call (with increasing familiarity with the service) and a significant decrease from the second call to the third (where the additional utterance was required). On the whole, however, the results showed little difference across the sequence of calls to a given version, indicating little perturbation due to the 'failed' verification attempt in call 3.

Analyses similar to the first ANOVA on the mean usability scores were run on the scores for all the individual usability attributes. The main effect of voiceprint type was highly significant (p<0.01 – and in most cases p<0.001) on 12 of the 20 usability attributes, and significant (p<0.05) on two of the remaining attributes. The results here on the calls with voiceprint verification were similar in some ways to those on the enrolment calls, described in the previous section: again scores were generally higher for Digits than for Sentences, and the largest and most highly significant differences occurred on cognitive and fluency attributes, particularly *concentration*, *flustered*, *stress*, *frustration* and *complication*, plus a few of the other attributes including *need for improvement* and *enjoyment*. However, the differences between the two types of voiceprints were generally smaller on the verification calls than on the enrolment call.

The interaction of voiceprint type and digits type was significant (p<0.05) on eight attributes (*concentration, flustered*, *stress*, *frustration, control*, *complication*, *ease of use* and *needs improvement*). As in enrolment, participants using telephone numbers rated the Digits version of the service further above the Sentences version on these attributes than those who were using random digit strings. In the verification calls, however, no significant interaction of voiceprint type and sentence type was found for any of the attributes. Participants with branded sentences rated the difference in usability between Digits and Sentences similarly to those with unbranded sentences.

In summary, on the calls with voiceprint verification the comparative usability results were generally similar in form to those found for enrolment. They were, however, weaker (particularly with regard to any differences between the branded and unbranded sentence groups) perhaps because the voiceprint utterances made up a smaller proportion of each call during verification; only one or two sentence inputs in the verification calls, compared to three sentences spoken twice each during the enrolment call. In particular a significant difference in the relative scores (Sentences minus Digits) was found between branded and unbranded sentences on the enrolment call, where there were three sentences spoken twice each in the same call, but not on the verification calls, where each call involved only one or two sentence inputs.

6.3. Explicit Preference

As part of the de-briefing interview, participants were asked which of the two versions of the service they preferred overall (based on all four calls to each, including both enrolment and verification, and prior to being given the details of how

the versions differed from each other). The results are presented in Table 3. The majority of participants stated that they preferred the Digits version of the service (69.1%). A binomial test, omitting the 'no preference' responses, showed that the majority preference for Digits over Sentences was very highly significant (p<0.001).

Sentences	Digits	No preference
50 (24.5%)	141 (69.1%)	13 (6.4%)

Table 3: Votes for Most Preferred Version (N=204)

Table 4 shows the preference votes broken down by which type of sentences (branded and unbranded) and which type of numbers (random digit strings or telephone numbers) the participant had experienced. In every participant group, a majority preferred Digits over Sentences, but this majority preference was stronger amongst participants given branded sentences than amongst those with unbranded sentences, and stronger in the telephone numbers group than in the group given random digit strings. Binomial tests within the groups (ignoring the 'no preference' responses) showed that the majority preference for each type of digits over branded sentences was highly significant (p<0.001), and the preference for telephone numbers over unbranded sentences was significant (p=0.016), but the preference for random digits over unbranded sentences did not attain statistical significance (p=0.144).

	Random Digits			Telephone Number			Total		
	Pref S	Pref D	Nopref	Pref S	Pref D	Nopref	Pref S	Pref D	Nopref
Branded	12 (22%)	38 (70%)	4 (7%)	6 (12%)	43 (84%)	2 (4%)	18 (17%)	81 (77%)	6 (6%)
Unbranded	18 (37%)	29 (59%)	2 (4%)	14 (28%)	31 (62%)	5 (10%)	32 (32%)	60 (61%)	7 (7%)
Total	30 (29%)	67 (65%)	6 (6%)	20 (20%)	74 (73%)	7 (7%)	50 (25%)	141 (69%)	13 (6%)

Table 4: Votes for Most Preferred Version by Sentence Type and Digits Type

6.4. Ratings - Overall Quality and Security

Participants were asked to rate the overall quality of the two versions by placing markers on a scale from 0 (worst) to 30 (best). They were also asked to rate on the same scale a service where the security procedure is that of the existing telephone banking service (recall of two digits from a six-digit secret number). The mean ratings for each of the three versions are shown in Table 5.

Sentences	Digits	Secret Number (SN)
17.5	22.9	19.7

Table 5: Mean Ratings - Overall Quality (N=204)

A repeated measures ANOVA was run on the quality ratings, with *verification method* (Sentences, Digits or SN) as the within-participants factor, and *age group*, *gender*, *location*, *order of experience*, *sentence type* and *digits type* as between-

participants factors. The main effect of verification method was found to be very highly significant ($p<0.001$), with Digits very significantly above both Sentences and use of a Secret Number (pairwise least-significant difference $p<0.001$ in each case) and use of a Secret Number significantly above Sentences ($p=0.013$).

Separate analysis within the telephone numbers group, and within the random digits group, showed that the difference in ratings between Sentences and Digits was very highly significant ($p<0.001$) in each case, with each type of digits being rated above sentences. The random digits version was also rated very significantly above the SN version (mean ratings 23.5 and 19.1; $p<0.001$), but the difference between telephone number and SN did not attain significance (mean ratings 22.2 and 20.4; $p=0.115$).

Participants were then asked to rate the two voiceprint versions and the Secret Number version on the scale again, this time in terms of security only (Table 6).

Sentences	Digits	Secret Number (SN)
22.2	20.9	20.0

Table 6: Mean Ratings - Security (N=204)

Here, the mean rating for voice authentication using Sentences was slightly higher than for the other two methods. A repeated measures ANOVA was run on the security ratings, with the same factors as above. This showed a significant effect of verification method ($p=0.031$); pairwise comparisons (least-significant difference) yielded a significant difference between Sentences and Secret Number ($p=0.012$) and a nearly significant difference between the two voiceprint methods ($p=0.056$). Importantly, a significant interaction occurred between verification method and digits type ($p=0.045$) as illustrated in Figure 3; participants using telephone numbers (right hand column of data points in Figure 3) rated the Digits method as *less* secure than any of the other methods (including Secret Number) whereas participants using random digits (left hand column of data points in Figure 3) rated the two voiceprint methods similarly and placed both above the Secret Number method. Averaged across the different types, therefore, the mean score for Digits was slightly lower than for Sentences.

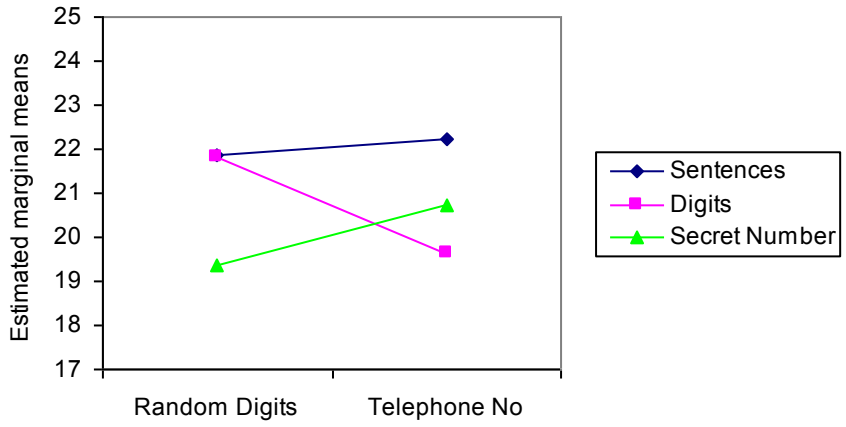

Figure 3: Security Ratings by Verification Method and Digits Type

6.5. Qualitative Data

Enrolment

Participants were asked to comment on the voiceprint enrolment process that they experienced for each of the two versions. The results for the different types of Sentences and Digits were similar, and hence are shown combined.

The most frequent comment concerning the Sentences enrolment process was that it was fine or OK (mentioned by 66 participants); 11 participants said that it was good or easy. Negative comments on this enrolment included that it required concentration or was difficult (38), that it was long-winded or laborious (38), that there was a lot to remember to do (18), and that the participants disliked it (11).

For the Digits enrolment process the most frequent comment was again that it was fine or OK (mentioned by a larger number, 91 participants); here, 48 commented that it was good or easy, whilst 10 mentioned that it was quick. Negative comments (of which there were substantially fewer than for the Sentences version) included that it required concentration or was difficult (15), that the participants disliked it (10) and that it was long-winded or laborious (9).

Sentences and digits as voiceprint tokens

Participants were next asked to comment on the Sentence type they had experienced in the experiment (branded or unbranded) and then on the other type.

The majority of comments about the branded sentences were that they were OK or fine (mentioned by 71 participants); 7 participants said that they liked them. Negative comments included that they were too long or long-winded (29), that the participants disliked them (28), that "the Bank" part of the sentence was disliked either because it felt like advertising for the bank or because by saying the name of the bank it could inform people around you which bank you use or that you are calling your bank (14), that the sentences were corny or silly (13), that saying the sentences was complicated or difficult (7), and that the sentences were not secure (6).

The majority of comments about the unbranded sentences were that they were OK or fine (mentioned by a larger number, 102 participants); 6 participants said that they liked them, 7 that they preferred saying these than those with "the Bank" in the wording. Negative comments about the unbranded sentences (of which there substantially fewer than for the branded equivalents) included that the participants disliked them (20), that they were too long or long-winded (17), that they were not secure (11), and that the sentences were corny or silly (6).

For each type of sentence participants were also asked if they would be happy to say this kind of sentence aloud when using telephone banking in private and in public. 94.2% of participants said that they would be happy to say a branded sentence aloud in a private place such as in their home; however, only 27.0% would be happy to do so in a public place such as in their office or in a train. For unbranded sentences the figures were 95.1% (in private) and 41.2% (in public).

Participants were then asked which of the two Sentence types they would prefer to use as a voiceprint. 63.7% of participants stated that they would prefer to use the unbranded sentence type; 24.0% said they would prefer the branded version (12.3% had no preference).

When asked similar questions regarding the two types of digit strings used in the experiment, 98.5% of participants said that they would be happy to say aloud a random string of digits in a private place; 65.0% would be happy to do so in public. For telephone numbers the figures were 95.1% (in private), but just 21.7% in a public setting.

When asked which of the Digits types they would prefer to use as a voiceprint 77.0% of participants stated that they would prefer to use a random string of digits; just 21.1% stated that they would prefer to use their telephone number (2.0% had no preference).

Concerns about voiceprint technology

Participants were then asked if they had any concerns about the Bank using voiceprint technology. A total of 55 participants (27.0%) stated that they had no concerns. Participants who raised a concern commented on the reliability of the technology or whether it was indeed secure (39 participants), on whether someone else could copy your voice or if similar-sounding family members could gain access using a voiceprint (38), on what would happen when they had a cold or illness which affected their voice (20), on concerns about identity theft (16), and that the customers were unsure of the technology or how it works.

Participants were also asked how they would feel about the Bank storing their voiceprint data in order to use it in future calls. The majority of participants stated that it was fine to do so (mentioned by 142 participants). Other participants commented that it would be fine as long as the data is secure or not shared with third parties (40). Some participants stated that they thought it was not secure or there was a risk of fraud (11) or that they would not be happy about the Bank storing their voiceprint data (9).

Whilst the numbers of participants happy to have their voice data stored is encouraging, these results indicate the importance of explaining the technology to new users in order to address the identified concerns and provide reassurance.

Use of voiceprints in real life

Participants were asked whether they would prefer to be verified using their voiceprint only, some secret information only (like digits from a Secret Number), or a 2-factor method of both voiceprint and secret information. A clear majority, 67.2% of participants, stated that they would prefer the 2-factor method, 25% said that they would prefer to be verified using just their voiceprint and 6.4% said that they would prefer to be verified using just the secret information (1.5% no preference).

The most common reason given for preferring a 2-factor method of verification was added security (mentioned by 97 participants). Another frequently cited reason was that one method could be used as a backup to the other method (10).

Participants were also asked if they would be happy to use voiceprints in real life when calling the Bank. 84.3% of participants said that they would be happy to use voiceprints. Only 9.8% said that they would not be happy using voiceprints (5.9% 'did not know').

6.6. Performance Data

Results were obtained for both types of voiceprint with regard to the success of enrolment and verification calls (based on participants' behaviour and recogniser

performance in each case – not, as explained earlier, on verification outcomes) and call timings.

Enrolment on sentences

Data from the 204 participants' enrolment calls were first analysed in terms of participant accuracy in saying the required phrase. This was determined by comparing the prescribed sentence against transcription of the recorded utterance.

Enrolment for Sentences required the customer to say three different sentences twice each. For branded sentences the rate of correct inputs at the first dialogue stage (first utterance of the first sentence) was 83.8%, and the overall average across the six stages was 87.0%. For unbranded sentences the figures were 81.8% and 88.3% respectively. As expected the highest error rate (across both types of sentence) was on the first utterance of the first sentence; however, there was little evidence of a consistent, decreasing trend across the sequence of utterances i.e. almost as many mistakes were made by participants in subsequent sentences as were made repeating the first.

In total, just 57.8% of participants spoke their sentences correctly at the first attempt in *all six* stages of the enrolment dialogue (54.3% of those being asked to repeat branded sentences, 61.6% of those echoing unbranded sentences).

Many of the incorrect utterances were, in fact, *almost* correct – differing from the intended sentence only in one short word (e.g. "I access telephone banking via my voiceprint" instead of "I access telephone banking using my voiceprint"), or by a minor substitution such as "I am" for "I'm". Others had an error in a more important word, such as "voicemail" for "voiceprint", or in a whole phrase, e.g. "I access telephone banking via my bankprint". In some cases the participant stumbled and repeated part of the sentence, e.g. "<Bank> uses my tele- my voiceprint for telephone banking." In other cases the participant apparently could not remember the sentence, and either broke off part-way through or made something up, e.g. "<Bank> uses my eh my bank details for telephone voiceprinting." Certain sentences seemed particularly prone to errors: for instance the word "only" in "My voice lets only me access my account" was often misplaced, either before "lets" or after "me", and there were various errors, from minor variants to completely forgetting the wording, on the sentence "<Bank> uses my voiceprint to confirm it's me."

There were a very few occurrences of silence, speaking too early or DTMF input, which in fact has the advantage in comparison to inaccurate speech input, in that it will never be accepted in error by the recogniser.

The total number of input attempts made by participants at the six stages of the sentence enrolment dialogue ranged from 6 (no extra attempts) up to 14 – not counting any inputs during failed call attempts, which occurred for a few participants (see the section 'Enrolment call failures'). The majority of participants (84.3%) had no extra attempts, i.e. their first input was accepted at every stage – despite the fact that only 57.8% gave exactly correct inputs throughout the dialogue. Many of the inputs with minor errors, and some of those with larger errors, were accepted as valid by the recogniser. The average total number of input attempts per participant was 6.33. The branded sentences gave rise to more extra attempts than the unbranded sentences (mean numbers of attempts 6.47 and 6.19 respectively).

Analysis of the transcriptions showed that the final set of accepted enrolment utterances was error-free for only 65.2% of participants (63.8% in the branded sentence group, and 66.7% in the unbranded group). The other 34.8% of participants consisted of 26.5% with errors in one of their three sentences, 7.8% with errors in two of the three, and 0.5% (i.e. one participant) with errors in all three sentences. This may indicate that the recogniser parameters need to be set more strictly when collecting enrolment utterances, so as to reject more of the utterances with errors. On the other hand, since many of the errors were minor, it may be that their effect on the speaker verification performance is acceptable for practical purposes.

Enrolment on digits

For Digits, the enrolment dialogue varied according to the type of number. Participants using telephone numbers were asked to say a single 11-digit number three times, whereas those using random digits went through an enrolment sequence similar to that for Sentences, in which three different digit strings were requested twice each.

For random digits, the rate of correct inputs at the first dialogue stage (first utterance of the first digit string) was 88.3%; rates at the subsequent five stages were all above 97%, and the overall average across the six stages was 97.2%. All of the 12 errors at the first stage consisted of saying the four digits only once instead of twice e.g. saying "five nine zero four" when prompted to say "five nine zero four, five nine zero four", probably as a result of misinterpreting the prompt; in each case the participant corrected this at the second attempt, and there were no further errors of this type at any of the later dialogue stages. Almost all the correct inputs were accepted by the recogniser. Accordingly, the total number of extra input attempts was small, and the average number of attempts per participant across the six dialogue stages was 6.19. Only one incorrect input (actually correct except for a filled pause "eh" between digits) was accepted for use in creating a voiceprint, and so 102 out of 103 participants (99.0%) had error-free enrolment data, in contrast to the figure of 65.2% reported above for Sentence enrolment.

For telephone numbers, the rate of correct inputs at the first stage of enrolment (first utterance of the telephone number) was 99.0%; at the second stage, 100%; and at the third stage, 97.0%. Thus the mean correct first-attempt input rate was 98.7%. The small number of errors were a mixture of errors in the number itself and the format in which it was given. All were detected and led to reprompts, at which the participants gave the number successfully. One correct telephone number input was misrecognised, but again the number was given and recognised successfully at the reprompt. The average number of inputs required was 3.05, against the possible minimum of 3, and the accepted set of enrolment utterances was error-free for all 101 participants (100%) in the telephone numbers group.

Enrolment call failures

In a small number of cases more than one attempt at the enrolment call was required. Four of the 204 participants failed at their first attempt at enrolment due to repeated rejection[8] of their (branded) sentence utterances. One participant failed to enrol twice

[8] Note that 'rejection' here means rejected by the speech recogniser because the correct phrase was not recognised confidently. In this experiment there was no requirement of consistency in the speaker's voice between utterances in the enrolment set, which would tend to increase the rejection rate.

using the telephone number dialogue as a result of giving their own number instead of the fictitious persona's. All succeeded on a subsequent attempt.

In addition to the main set of 204 participants, as mentioned earlier, three customers were unable to complete the experiment procedure due to repeated misrecognition of digits. One failed to complete enrolment using random digits due to repeated misrecognition of the word "six" as "two", despite making four attempts at the call. Two others completed Digits enrolment but were repeatedly misrecognised during a subsequent verification call, thus failing to progress through the experiment and therefore experience the Sentences version.

Counting these three, the failure rate on the first attempt at the enrolment dialogue was 4/205 = 2.0% for Sentences, and 2/207 = 1.0% for Digits. Given that one of the two Digits enrolment failures was due to confusion between real and fictitious telephone numbers, which would not occur in real life, the results might suggest that enrolment using Sentences is more likely to fail than enrolment using Digits; but as very few participants failed in either type of enrolment dialogue the evidence is very weak. It is worth noting, however, that as with other forms of biometric identification failure to enrol may be an issue for a minority of voiceprint users[9], even with best-case performance as employed in the reported experiment, and that appropriate strategies for the handling of such failures should be put in place.

Verification calls using sentences

Results showed that fewer participants made errors in saying the sentences during the verification dialogue than in the enrolment dialogue. For branded sentences the rate of correct inputs at the first attempt at verification was 92.4%. The average across the four different verification stages (three separate calls with two bids in Call 3) was 93.2%. For unbranded sentences the figures were similar; 89.9% accuracy on first attempt at verification and a mean accuracy of 92.6% across all four verification stages.

The errors in speaking the sentences were mostly minor, with very few instances of forgetting the sentence. As in the enrolment calls, many utterances with minor errors were accepted by the recogniser. Overall, only 1.5% of first-attempt sentence utterances were rejected by the recogniser (and hence required retries); in most cases the second attempt was accepted, and the mean total number of sentence utterances across the four verification dialogue stages per participant was 4.07. More extra attempts were required at the branded sentences (mean total 4.12 attempts) than at the unbranded ones (mean total 4.02).

Verification calls using digits

The random digit strings were also spoken a little more accurately during verification than during enrolment. First-attempt accuracy was 98.1%, with a mean of 98.5% correct first time in verification, against 88.3% and 97.2% in enrolment as reported earlier). For telephone numbers, in contrast, the mean error rate was higher in the verification process than in the enrolment call, with 97.0% correct against the 98.7% reported for enrolment (although first-attempt accuracy was the same at 99.0%).

---

[9] Coventry *et al* (2003b), for example, report an 8.5% failure to enrol rate for fingerprints at the ATM. (average age of user 34.6 years). Amongst older users (average age 65.7 years) a considerably higher figure of 33% was found (Riley *et al,* 2007).

On the random digit strings, two participants spoke the four digits only once instead of twice on their first encounter with the verification prompt (Call 1); the other errors were a mixture of DTMF input, speech-too-early, insertion of an extra digit and transposition of digits in the string, all leading to reprompts. A few correct utterances were misrecognised or rejected, also leading to reprompts, and the average total number of inputs required across the four verification stages was 4.14.

Errors on the telephone number occurred in 14 calls (distributed across the four different verification stages) by 13 different participants. All but four of the errors appeared to be accidental slips in giving the correct number; the others included errors in the format of the number and self-corrections. Four utterances containing errors were accepted as correct numbers by the recogniser; the other 10 required reprompts. In addition, three participants had correct numbers misrecognised, leading to reprompts. Overall, for those using telephone numbers, the mean total number of utterances required was 4.13.

Verification call failures

For each version of the service, a large majority of participants required only one attempt at each of the three calls. Only one failure occurred at the verification stage in the main cohort of 204 participants, with one participant repeatedly transposing two digits of the random digit string. (When given another attempt the participant did not make this error.) In addition, two of the excluded participants had repeated verification call failures using Digits (one with random digit string, one with a telephone number); one because their pronunciation of the digit "oh" in the telephone number was persistently recognised as "four", the other because "two" was misrecognised as "eight" in the random digit string. There were no failed calls due to problems with Sentence verification.

Enrolment timings

Average times for the enrolment section of the call (excluding ID&V) are shown in Table 7 (ignoring any unsuccessful call attempts).

Voiceprint Type	Enrolment Duration
Branded Sentences (N=105)	108.3
Unbranded Sentences (N=99)	97.1
All Sentences (N=204)	*102.8*
Random Digits (N=103)	105.2
Telephone Number (N=101)	62.8
All Digits (N=204)	*84.2*

Table 7: Mean Duration (in Seconds) for Enrolment

Enrolment based on telephone numbers took the shortest time on average. A repeated measures analysis of variance (ANOVA) was applied to the durations for the enrolment stage, with *voiceprint type* as the within-participants factor, and *age group*, *gender*, *location*, *order of experience*, *sentence type* and *digits type* as between-participants factors. This showed a very highly significant effect of voiceprint type ($p<0.001$), with enrolment using Digits taking less time than enrolment using

Sentences. There were also very highly significant effects of sentence type and digits type ($p<0.001$ in each case), with branded sentences taking longer than unbranded sentences, and random digits taking longer than telephone numbers – all in line with expectations based on the differing numbers and lengths of utterances required in the different enrolment dialogues.

Separate analysis on each subset of participants who had a particular combination of sentence type and digits type (about a quarter of the cohort in each case) showed that the difference in enrolment time between telephone numbers and either type of sentence was very highly significant ($p<0.001$ in each case). The difference between random digit strings and unbranded sentences was also highly significant ($p<0.001$). However, enrolment based on branded sentences took almost as long as using random digit strings.

Verification timings

The mean times taken to complete ID&V (ignoring any unsuccessful call attempts) in the three verification calls to each version of the service are shown in Table 8.

Voiceprint Type	ID&V Duration			
	Call 1	Call 2	Call 3	Mean
Branded Sentences (N=105)	41.4	40.9	55.6	46.0
Unbranded Sentences (N=99)	39.0	38.9	50.7	42.9
All Sentences (N=204)	*40.3*	*39.9*	*53.2*	*44.5*
Random Digits (N=103)	41.0	40.4	52.9	44.8
Telephone Number (N=101)	38.3	38.5	49.4	42.1
All Digits (N=204)	*39.7*	*39.4*	*51.2*	*43.4*

Table 8: Mean Duration for ID&V (in Seconds) with Verification

A repeated measures ANOVA was run on the ID&V durations, with *voiceprint type* and *call number* as within-participants factors, and *age group*, *gender*, *location*, *order of experience* (Sentences first or Digits first), *sentence type* and *digits type* as between-participants factors. The main effect of voiceprint type was found to be statistically significant ($p=0.014$), with Digits taking less time than Sentences. Sentence type also had a significant effect ($p<0.01$ for both the main effect and its interaction with voiceprint type), with branded sentences taking longer than unbranded ones. Digits type, too, had a significant effect ($p=0.026$ for the main effect and $p<0.001$ for its interaction with voiceprint type), with random digit strings taking longer than telephone numbers – presumably because the prompts were longer.

Separate analysis on each subset of participants who had a particular combination of sentence type and digits type (about a quarter of the cohort in each case) showed that in the verification calls, telephone numbers were on average significantly shorter than sentences in the branded sentence subgroup ($p=0.002$), but not in this case the unbranded subgroup (although the mean was lower in both cases). In enrolment, the difference was significant in both cases. However, as in enrolment, again random digits and branded sentences resulted in similar durations, and verification using

random digits was significantly longer than that based on unbranded sentences (p<0.001).

The main effect of call number was very highly significant (p<0.001), with highly significant pairwise differences (p<0.001) between Call 3 (which required an extra attempt at verification in all cases) and both the previous calls, but no significant difference between Call 1 and Call 2 (pairwise least-significant difference comparisons in each case).

6.7. Summary of Main Results

Table 9 summarises the key results from the experiment with regards to the overall comparison between Sentences and Digits. Note that in the "Significant Differences" column, "**>**" means 'significantly better than', and that for most of the metrics, larger values are better, but for *duration* the opposite applies.

	Sentences	Digits	Significant Differences
Usability: Enrolment	5.29	5.70	Digits **>** Sentences
Usability: Verification	5.65	5.86	Digits **>** Sentences
Preference Votes	24.5%	69.1%	Digits **>** Sentences
Quality Rating	17.5	22.9	Digits **>** Sentences
Security Rating	22.2	20.9	-
Duration: Enrolment	102.8s	84.2s	Digits **>** Sentences
Duration: Verification	44.5s	43.4s	Digits **>** Sentences

Table 9: Summary of Main Results[10]

The results consistently show significantly better results for Digits than for Sentences. Voiceprint authentication based on Digits was found to be significantly more usable, was preferred by significantly more participants and was rated significantly higher in terms of overall quality than that based on Sentences. Enrolment and verification also took on average significantly less time when based on Digits. The only metric for which there the difference between Sentences and Digits was only approaching significance (p=0.056) was perceived *security*. Here, the two voiceprint types were rated similarly when random digits were used; telephone numbers, however, were perceived as *less* secure than Sentences. Averaged across the different types, therefore, the mean score for Digits was slightly lower than for Sentences.

7. Conclusions

Usability results from the experiment show some interesting differences between voiceprint types which, together with participant preferences expressed in the interview, can inform decisions on the use of voiceprints in a dialogue system and specifically a telephone banking service.

[10] Usability scores are reported here as sample means on a 7-point response scale. Quality and Security Ratings are reported here as sample means on a 30-point response scale.

In terms of user performance, participants' accuracy in saying the requested phrases was lower for sentences than for digits, although many of the errors in saying the sentences were minor (and would not necessarily result in poorer speaker verification performance compared to digits). Both enrolment and verification took longer using random digits or sentences (especially the branded sentences) than using telephone numbers – partly because the prompts were longer (since they explicitly stated the sentence or digit string to be spoken, rather than simply saying "your telephone number"), and partly because the enrolment involved three different sentences or digit strings rather than only a single telephone number. However, the differences in time required for verification were small (three or four seconds), and the larger differences occurring in enrolment call duration would be less important in practice since enrolment is a one-off process for each customer.

In terms of users' attitude towards the usability of the service, data from the experiment consistently showed that participants found the service with telephone numbers the most usable, both at the enrolment stage and at the verification stage. Participants also found the version with random digit strings more usable than a version involving either branded or unbranded sentences.

Initial preference results between the voiceprint types experienced in the experiment also indicated that whichever sentence type and digits type had been experienced, the majority preference was for Digits over Sentences, but the preferences were stronger in favour of telephone numbers than in favour of random digits, and stronger against branded sentences than against unbranded sentences – a pattern similar to that in the usability scores overall.

On ratings of overall quality, the version using Digits was marked significantly above the version using Sentences, regardless of which type of number was used. On security, in contrast, there was no significant difference between sentences and random digits, and telephone numbers were rated as significantly *less* secure than sentences. Voiceprints based on sentences or random digits were also considered more secure than the current verification method (two digits from a Secret Number), but voiceprints based on telephone numbers were not.

These are interesting results, especially in respect of the comparison between telephone numbers and the other voiceprint types. Telephone numbers were considered more *usable* than all other voiceprint types, but less *secure* – while the measure of *overall quality* gave intermediate results (telephone numbers marginally below random digits, but both digits types above sentences).

Further preference results obtained later in the interview, when participants were asked to explicitly compare the two sentence and digits types, added to the picture. There was a strong majority preference for unbranded over branded sentences (63.7% against 24.0%), which is consistent with the usability and initial preference data. More interestingly, there was also a strong majority preference for random digits over a telephone number (77.0% against 21.1%). This suggests that security and privacy considerations (note that far fewer would be happy to say their telephone number than a random digit string in a public place) outweigh usability considerations in determining participants' preference for voiceprint type.

The main conclusion is that customers would be happy with the use of voiceprints based on random digit strings, but less happy with any of the other types of voiceprints considered here. Sentences were rated as less *usable* than digit strings, and

came out below them in the preference votes. Telephone numbers, though highly usable, were much less acceptable in respect of *security and privacy*.

A second conclusion is that customers acknowledge the need for security, possibly even at the expense of convenience when it comes to verification methods for telephone banking. This is shown in the sizeable majority of participants (67.2%) expressing a preference for 2-factor verification (voiceprint plus secret information) over the use of either a voiceprint or secret information alone. However, since participants did not directly experience a 2-factor process in this research, to investigate this further a second experiment was carried out in which the use of voiceprints (based on digits) was compared in both a 2-factor and single factor approach. Results of this experiment are currently in preparation for publication.

On a more general level, results from the experiment are encouraging in that they indicate customers are happy to accept voiceprint technology for telephone banking in real life (84.3% said they would be happy to do so). Usability scores for both types of voiceprint, although significantly different, were high at above 5.0 on a 7-point scale. Moreover, privacy concerns relating to the storage of their voice data, often cited as a potential issue in the deployment of biometric technology, do not appear to be a barrier in this context (69.6% said they had no concerns at all, a further 19.6% said they had no concerns provided the data were not shared with any third parties). These data provide quantitative support for the view that users are "*more open to the concept* [of biometrics] *where there is clear security need in their personal life e.g. banking*" (Sasse, 2004). Interview results highlighted the importance of explaining the technology to new users, and indeed the (positive) results reported here may have been coloured by participants' lack of knowledge and expertise on the subject.

Conclusions from the experiment are, of course, based on customers' experience of one particular form of enrolment and verification dialogue for each voiceprint type (designed to reflect realistic possible uses of the technology in the live service). It would be difficult to investigate the full range of design possibilities in a single experiment, but it is worth bearing this factor in mind when interpreting the results. It is possible, for example, that the preference for digits may be reduced in a scenario where more extended enrolment is used to allow increased protection against spoofing (e.g. repetition of all the digits from zero to nine), although longer enrolment was not found to be a determining factor in the existing experiment. (Moreover, it could be argued that if this were the objective, enrolment with the other voiceprint types would also have to be altered and extended.)

Additional errors due to verification failure (false non-match) may also impact the results since the level of errors may vary for the different voiceprint types. In fact, while the amount of speech data collected in this experiment was not sufficient for a definitive comparison of verification error rates, some offline tests were run on the utterances from the experiment, and these yielded equal error rates[11] on sentences (branded 0.95%, unbranded 0.86%) that were intermediate between those for

[11] The Equal Error Rate (EER) is the percentage rate of false matches or false non matches obtained when the verifier's decision threshold is set so that these two percentages are equal. The false acceptance rates for this purpose were computed using all speakers of the same sex as the target speaker who had spoken the same number or sentence during the experiment. It must be emphasised that the number of speakers here (approximately 100 for each voiceprint type) was not sufficient to provide an accurate measure of verification performance, and therefore the stated error rates are only order-of-magnitude estimates.

telephone numbers (0.33%) and random digit strings (1.34%) - low in each case. Allowing multiple attempts at verification, as in the experiment, would further reduce the rate of false non-matches, with attitude results suggesting that users were not perturbed by a 'failed' verification attempt in their third call to the service (although they were not explicitly told the reason for the retry). This experiment did not examine the situation in which a caller repeatedly experiences a false non-match and is passed to an agent as a result. Although likely to be relatively rare, it is likely that this will have considerable impact on the individual user's attitude towards the technology and requires sensitive handling on the part of the Bank.

These qualifications aside, this research represents a detailed evaluation of voiceprint technology from a user perspective that is rarely evident in other research in this area, or indeed the field of biometrics in general (Toledano *et al*, 2006). Moreover, its setting within the context of an already established telephone banking system strengthens understanding of how voiceprints can be used in real-life scenarios and provides practitioners considering use of the technology with valuable empirical data on which to base their design decisions.

Further research work being considered in this field as suggested by these results, includes a investigation of user perceptions of voiceprints technology on the basis of a more prolonged exposure to the technology in a longitudinal study; investigation of the role of text-independent voiceprint technology (in contrast to the text-dependent approach reported here): and a study of the impact of real world usage environments on technology performance and user perceptions - as opposed to the laboratory setting used to derive the results reported here.

Acknowledgements

The authors wish to acknowledge the generous support of Nuance Communications Inc. in this research.

Appendix A. Items in Usability Questionnaire

Statements were presented in a randomised order for each participant.

Q1 I thought the service was too complicated.

Q2 When I was using the service I always knew what I was expected to do.

Q3 I thought the service was efficient.

Q4 I liked the voice.

Q5 I would be happy to use the service again.

Q6 I found the service confusing to use.

Q7 The service was friendly.

Q8 I felt under stress when using the service.

Q9 The service was too fast for me.

Q10 I thought the service was polite.

Q11 I found the service frustrating to use.

Q12 I enjoyed using the service.

Q13 I felt flustered when using the service.

Q14 I think the service needs a lot of improvement.

Q15 I felt the service was easy to use.

Q16 I would prefer to talk to a human being.

Q17 I thought the voice was very clear.

Q18 I felt that the service was reliable.

Q19 I had to concentrate hard to use the service.

Q20 I did not feel in control when using the service.

References

Adams, A. and Sasse, M. (1999). "Users are not the enemy." Communications of the ACM 42 (12), pp.41–46, reprinted (2005) in: Cranor and Garfinkel (Eds), Security and Usability, O'Reilly, pp.639-649 [chapter 32].

Arivazhagan, S., Ganesan, L. and Srividya, T. (2009). "Iris recognition using multi-resolution transforms." International Journal of Biometrics, vol 1(3), pp.254-267.

Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T, Ortega-García, J., Petrovska-Delacrétaz, D. and Reynolds, D.A. (2004). "A tutorial on text-independent speaker verification.", EURASIP Journal on Applied Signal Processing, issue 4, pp.430-451.

BioVision, 2003. "Roadmap for biometrics in Europe to 2010." Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.99.6354 (last accessed March 2010).

Bishop M. (2005). "Psychological acceptability revisited." In: Cranor and Garfinkel (Eds), Security and usability. O'Reilly, pp.1–11 [chapter 1].

Coventry, L. (2005). "Usable biometrics." In: Cranor and Garfinkel (Eds), Security and usability. O'Reilly, pp.175–198 [chapter 10].

Coventry, L., De Angeli, A., and Johnson, G. (2003a). "Usability and biometric verification at the ATM interface." In: Proceedings of the SIGCHI Conference on Human factors in Computing Systems (CHI '03), pp.153-160.

Coventry, L., De Angeli, A., and Johnson, G. (2003b). "Biometric verification at a self service interface." Proceedings of the British Ergonomic Society Conference, Edinburgh.

Davidson, N., McInnes, F.R. and Jack, M.A. (2004). "Usability of dialogue design strategies for automated surname capture." Speech Communication, 43(2), pp.55-70.

Dhamija, R. and Perrig, A. (2000). "Déjà vu: a user study using images for authentication." In Proceedings of the Usenix Security Symposium, August 2000.

Dutton, R.T., Foster, J.C., Jack, M.A. and Stentiford, F.W.M. (1993). "Identifying usability attributes of automated telephone services." In Proc. EUROSPEECH'93, pp.1335-1338.

Foster, J.C., McInnes, F.R., Jack, M.A., Love, S., Dutton, R.T., Nairn, I.A. and White, L.S. (1998). "An experimental evaluation of preferences for data entry method in automated telephone services." Behaviour and Information Technology, 17(2), pp.82-92.

Jain, A.K., Ross, A. and Pankanti, S. (2006). "Biometrics: a tool for information security." IEEE Transactions on Information Forensics and Security, June, vol 1(2), pp.125-143.

Hiltgen A, Kramp T and Weigold T. (2006). “Secure internet banking authentication.” IEEE Security and Privacy, vol 21(9), pp.21-29.

ISO/IEC JTC (2007). International Standardisation Organisation / International Electrotechnical Commission Joint Technical Committee “ISO/IEC 19795-2:2007 Biometric performance testing and reporting. Part 2: Testing methodologies for technology and scenario evaluation.”

Jack, M.A., Foster, J.C. and Stentiford, F.W.M. (1993). “Usability analysis of intelligent dialogues for automated telephone services.” In Proc. Joint ESCA/NATO workshop on Applications of Speech Technology, pp.149-152.

Khan, L.A., Baig, M.S. and Youssef, A.M. (2009). “Speaker recognition from encrypted VoIP communications.” Digital Investigation, In Press, Corrected Proof, Available online 10 November 2009.

Larsen, L.B. (2003). “Assessment of spoken dialogue system usability - what are we really measuring?” In Proc. EUROSPEECH'03, pp.1945-1948.

Larsen, L.B. (1999). “Combining objective and subjective data in evaluation of spoken dialogues.” In Proc. ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, Irsee, Germany, pp.89-92.

Likert, R. (1932). “A technique for the measurement of attitudes.” Archives of Psychology, 140.

Love, S., Dutton, R.T., Foster, J.C., Jack, M.A., Nairn, I.A., Vergeynst, N.A. and Stentiford, F.W.M. (1992). “Towards a usability measure for automated telephone services.” In Proc. Institute of Acoustics Speech and Hearing Workshop, vol 14(6), pp.553-559.

Morton, H., McBreen, H.M. and Jack, M.A. (2004). “Experimental evaluation of the use of embodied conversational agents in eCommerce applications.” In Ruttkay, Z. and Pelachaud, C. (Eds.), From Brows till Trust: Evaluating Embodied Conversational Agents, Kluwer, ISBN1-4020-2929-X, pp.592-599.

O'Gorman, L. (2003). “Comparing passwords, tokens and biometrics for authentication.” Proceedings of the IEEE, vol 91(12), December, pp.2021-2040.

McClue, A. (2003). “Nationwide Ditches Iris and Fingerprint Biometrics”, 23 September, http://www.silicon.com/software/security/0,39024655,10006129,00.htm last accessed March 2010.

Naini, A.S., Homayounpour, M.M. and Samani, A. (2009). “A real-time trained system for robust speaker verification using relative space of anchor models”, Computer Speech & Language, In Press, Corrected Proof, Available online 15 July 2009.

Nanni, L and Lumini, A. (2006). “Human authentication featuring signatures and tokenised random numbers.” Neurocomputing, vol 69(7-9), March, pp.858-861.

Nandini, C. and Ravi Kumar, C.N. (2008). “Comprehensive framework to gait recognition.” International Journal of Biometrics, vol 1(1), pp129-137.

Reynolds, D.A. (2002). “An overview of automatic speaker recognition technology.” In: Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP’02), pp.300-304.

Riley, C., McCracken, H. and Buckner, K. (2007). “Fingers, veins and the grey pound: accessibility of biometric technology.” In: Proceedings of the 14th European conference on Cognitive ergonomics (ECCE '07), pp.149-152.

Rossi, P.H., Wright, J.D. and Anderson, A.B. (1983). “Handbook of survey research.” New York, Academic Press, ISBN 0125982267.

Sasse, M. A. (2004). “Usability and trust in information systems.” In: Trust and Crime in Information Societies, edited by Robin Mansell and Brian S Collins, pp. 319-348. Edward Elgar. ISBN 1 84542 177 9.

Sasse, M.A., Brostoff, S. and Weirich, D. (2001). “Transforming the ‘weakest link’: a human-computer interaction approach to usable and effective security”, BT Technology Journal 19(3): 122-31.

Scheidat, T., Vielhauer, C. and Dittman, J. (2009). “Handwriting verification - comparison of a multi-algorithmic and a multi-semantic approach.” Image and Vision Computing, vol 3, February, pp.269-278.

Sturm, J. and Boves, L. (2005). “Effective error recovery strategies for multimodal form-filling applications.” Speech Communication, 45(3), pp.289-303.

Tan, B. and Schuckers, S. (2010). “Spoofing protection for fingerprint scanner by fusing ridge signal and valley noise” Pattern Recognition, In Press, Accepted Manuscript, Available online 10 March 2010.

Toledano, D.T., Fernández Pozo, R., Hernández Trapote, A and Hernández Gómez, L. (2006). “Usability evaluation of multi-modal biometric verification systems”. Interacting with Computers, vol 18(5), September, pp.1101-1122.

Venkatraman, S. and Delpachitra, I. (2008). “Biometrics in banking security: a case study.” Information Management and Security, vol 16(4), pp.415-430.

Weir, C.S., Douglas, G., Carruthers, M. and Jack, M.A. (2009). “User perceptions of security, convenience and usability for eBanking authentication tokens”, Journal of Computers and Security, Volume 28(1), February, pp.47-62.

Yamagishi, M., Nishiuchi, N. and Yamanaka, K. (2008). “Hybrid fingerprint authentication using artifact-metrics.” International Journal of Biometrics, vol 1(2), pp160-172.

Yan, J., Blackwell, A., Anderson, R. and Grant, A. (2004). "Password memorability and security: empirical results", IEEE Security and Privacy, Vol. 2, Issue 5, pp 25-31.

Yoma, N., Garretón, C. Molina, C. and Huenupán, F. (2008). “Unsupervised intra-speaker variability compensation based on Gestalt and model adaptation in speaker verification with telephone speech.”, Speech Communication, vol 50 (11-12), November-December, pp.953-964.