Usability assessment of text-to-speech synthesis for additional detail in an automated telephone banking system

Usability Assessment of Text-to-Speech Synthesis for Additional Detail in an Automated Telephone Banking System

Hazel Morton[a], Nancie Gunson[a,*], Diarmid Marshall[a], Fergus McInnes[a], Andrea Ayres[b], Mervyn Jack[a].

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

[b] Lloyds Banking Group, Canons House, Canons Way, Bristol, BS99 7LB, 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: text-to-speech; TTS; usability; dialogue design; automated telephony.

Abstract

This paper describes a comprehensive usability evaluation of an automated telephone banking system which employs text-to-speech (TTS) synthesis in offering additional detail on customers' account transactions. The paper describes a series of four experiments in which TTS was employed to offer an extra level of detail to recent transactions listings within an established banking service which otherwise uses recorded speech from a professional recording artist. Results from the experiments show that participants welcome the added value of TTS in being able to provide additional detail on their account transactions, but that TTS should be used minimally in the service.

1. Introduction

Speech applications have two primary options for speech output: natural speech prompts, recorded from human voice actors, and synthesised speech. Many early uses of synthesised speech, or text-to-speech (TTS) synthesis, were in systems for accessibility, for example reading systems for blind or sight impaired computer users, and mainstream usage of TTS was "severely limited by its quality" (Taylor, 2009: p.2). However as the quality of TTS systems improves, where quality defined in terms of the intelligibility of the system and the naturalness of the voice, TTS becomes more common in everyday applications.

In the creation of speech systems there is a trade-off to be made between the quality and expense of recorded prompts against the flexibility of synthesised prompts. Recorded prompts, which although having the benefit of sounding natural, can be expensive to create as they require the recording time of a voice actor. Synthesised speech may sound less natural, but has the advantage of being more flexible as the service designer can create new prompts as and when required without having to visit the recording studio. The use of TTS could be particularly beneficial therefore in services that require to output dynamic information, such as place names or company names, where recording such a diverse set of prompts would be unfeasible. Importantly, the use of TTS in such a case can potentially add value to a system that otherwise would be more limited in the information it can provide.

Previous research has investigated the usability and effectiveness of synthesised speech in a variety of applications, for example in a flight information system

(McInnes et al., 1999), in a personal information management application (Gong and Lai, 2003), in tutoring applications (Baylor et al., 2003; Forbes-Riley et al., 2006) and in a smart-home system (Möller et al., 2006).

Research which investigated users' perceptions of the personality of a synthesised voice compared with a recorded voice (on which the TTS system was modelled) found that the synthesised voice is associated with more negative personality characteristics than the recorded voice (Love et al., 2000). However other research which investigated synthesised speech in comparison to a number of recorded speech samples in a smart-home system found that synthesised speech prompts do not necessarily receive more negative ratings than recorded speech (Möller et al, 2006). Further, investigation was made of a combined recorded and synthesised voice compared to a fully synthesised voice. It was found that the combined recorded and synthesised version scored significantly higher than the fully synthesised version on overall quality, voice adequacy and voice pleasantness. However, no significant differences were found for listening effort. This study recommends that, as much as possible, recorded voices should be used and supplemented with synthesised when required, rather than opting for a fully synthesised system.

In the evaluation of TTS systems, many empirical evaluations focus on the acceptability, naturalness and comprehensibility of the systems (Stern et al., 1999; Stevens et al., 2005; Viswanathan and Viswanathan, 2005). Such research focuses on the comprehension or acceptability of TTS as a speech solution, that is, assessing TTS system prompts solely from a quality perspective. However, even if it can be assumed that the quality of TTS speech prompts are not as good as recorded prompts, the use of TTS in a dialogue system can be beneficial to its users by providing additional information that would be not be viable as a recorded prompt solution. Thus it is important to evaluate the use of TTS as a speech output solution from a usability perspective, within the context of a real-world application. The four studies described in this paper detail the evaluation of the usability of TTS within an already established dialogue system.

2. Evaluation of TTS: four studies

Four studies are presented here which investigate the use of TTS in an automated telephone banking service from a usability perspective. The service used in the experiments is a mirror copy of a telephone banking service from a major UK bank, referred to here as the Case Bank. The existing service at the time of the research utilised service prompts recorded from a human voice-talent actor (a female Southern British English voice) and the system functionality allows users to access their bank accounts, find out balance and transaction information and complete simple banking tasks such as transferring money between their accounts. The system utilises a speaker independent commercial speech recogniser[1] so that users can interact with the system using speech input without any prior training of the system; dual tone multifrequency (DTMF) input is also available. Natural language understanding (NLU) is implemented in the system via a finite-state grammar, in which allowable sequences of words and phrases are hand-coded. Each path in the grammar is associated with an appropriate feature-value pair in order to extract the meaning of the utterance.

[1] Nuance v8.0.0 www.nuance.com

Each of the experiments described here focused on the transaction listings within the banking service which provides a number of pieces of information on recent banking transactions such as the date and the amount of the transaction (e.g. "on the 12th of January, a debit for £30"). These prompts are concatenated from a library of recorded speech prompts. However, the system was unable to provide more detailed information, such as the exact location of a cash withdrawal transaction or the retailer to which a debit card payment was made. In employing TTS in the system development, such information could be used and the relevant speech output created dynamically in order that the information be passed on to the user (e.g. "on the 12th of January, a debit for £30 to *The Gift Shop*").

This series of experiments investigates customer perceptions of the use of TTS in the automated telephone banking system for this purpose. The first experiment investigates the usability of both a fully and a partially synthesised system for recent transaction readouts in comparison to a fully recorded system. The second experiment investigates the use of minimal TTS (for less frequently occurring names only) in providing additional detail. The third experiment investigates the optimal location of the TTS component within a recorded speech prompt. The fourth experiment investigates the readout of company names in the transaction lists, where some names may not be accurately represented in the transaction records.

In each case, where TTS prompts were employed, a commercially available TTS engine was used with one of their standard female (Southern) British English voices[2]. The system is based on concatenative synthesis[3] and allows for customisation of pronunciation and intonation, however, this was not required in the experiments. In the first experiment, all other prompts in the service were also pre-recorded by the voice talent on which the TTS voice is based, so as to minimise the change in voice between natural and synthesised speech. In subsequent experiments the non-TTS prompts were pre-recorded by the voice talent employed in the live service, also a female Southern British English voice.

Results from these four experiments show that TTS in offering additional information adds value to an already established telephone banking system but that the use of TTS should be kept to a minimum when providing this additional detail to the user.

The following table (Table 1) summarises the main features of the four experiments. It should be noted that the different versions relate to only the transaction listings part of the automated telephone banking service, and that all other aspects of the service (e.g. the identification and verification process, main menu content etc.) remained the same across all versions in each experiment and were identical to those in the live service.

---

[2] Nuance RealSpeak Telecom (Serena) www.nuance.com

[3] Concatenation-based synthesisers employ a database of segments that have been extracted from corpuses of recordings of human speakers. Different systems use different types of segment (in some cases multiple instances of each from different prosodic contexts) e.g. diphones, triphones and/or a combination, together with a range of different methods for selecting the appropriate segment. However, due to the commercial nature of the TTS system used in this research full details of its implementation are not obtainable.

Experiment	**Experiment 1** Extent of TTS Inclusion	**Experiment 2** Minimal Use of TTS	**Experiment 3** Location of Additional Detail	**Experiment 4** Readout of Company Names
Number of Versions	3	2	2	2
Versions Compared	A: All recorded speech as rest of banking service B: Recorded speech for carrier phrase + TTS additional detail C: TTS carrier phrase + TTS additional detail	A: All recorded speech, additional detail for Top companies only B: Recorded speech for carrier phrase, additional detail for all companies - Top companies recorded, all others TTS	A: Recorded speech for carrier phrase, additional detail – sentence final B: Recorded speech for carrier phrase, additional detail – sentence medial	A: Recorded speech for carrier phrase, additional detail in TTS – exact company name B: Recorded speech for carrier phrase, additional detail in TTS – malformed company name
Transactions on which Additional Detail offered	A: none B: all (TTS only) C: all (TTS only)	A: Top companies only (recorded only) B: all (Top recorded + others TTS)	A: all (Top recorded + others TTS), sentence final B: all (Top recorded + others TTS), sentence medial	A: all (TTS only) B: all (TTS only)

Table 1: Overview of Four TTS Experiments

Taken together these four experiments detail a comprehensive evaluation of the inclusion of TTS in an already established telephone banking system.

3. Experiment approach

The experiment approach involves a contrastive study where two or more versions of the dialogue system, differing in some design characteristic, are experienced by the participants. Participants are given detailed personal data as fictitious personae to use during the experiment and are asked to perform tasks typical of real-life use within the dialogue system. The results obtained from this procedure are considered to approximate the responses the service would generate in a real world context of use.

In this approach, a repeated-measures design is largely used to ensure maximum control over between-subject variability and a rich set of data is collected based on both performance measurements and subjective attitudes to the experiences of using the different versions of the service.

Participants' attitudes are measured using questionnaires completed after experiencing each version of the service. The approach uses attitude questions having a Likert format (Likert, 1932) where each usability attribute to be measured is presented to the participant in the form of a stimulus statement followed by an agree-disagree scale. The advantages of this format are described in Coolican (1994):

- Participants prefer the Likert scaling technique because it is "more natural" to complete and because it maintains their direct involvement in the process.
- The Likert technique has been shown to have a high degree of validity and reliability.
- The Likert scale has been shown to be effective in measuring changes over time.

The Likert format has been employed in previous research seeking to develop a general-purpose tool for the assessment of users' attitudes towards spoken language dialogue services or SLDSs. Hone and Graham (2000), for example, developed a prototype questionnaire (known as SASSI) initially containing 50 statements in Likert format, which they then used in four different studies involving the assessment of speech systems. Exploratory factor analysis on the data indicated six main factors in users' perceptions of SLDSs: identified as System Response Accuracy, Likeability, Cognitive Demand, Annoyance, Habitability and Speed. Evidence to support the reliability of the questionnaire was presented, but the conclusion was that further work is required on its development before it warrants general use.

The questionnaire employed in this research is a tool for evaluating users' attitudes towards automated telephone services which was developed over a number of experiments (Dutton *et al*., 1993; Jack *et al*., 1993; Love, 1997, Love *et al*., 1992, 1994). During development, salient attributes relating to the perceived usability of interactive systems were identified and a questionnaire was constructed to measure these attributes. Tests provided evidence of its reliability, validity and sensitivity (Dutton *et al*., 1993; Jack *et al*., 1993; Love *et al*., 1992) and it has been widely used and adapted since (Davidson *et al*, 2004; Foster *et al*., 1998; Larsen, 2003, 1999; Morton *et al*., 2004; Sturm and Boves, 2005).

The questionnaire contains 20 items in Likert format and covers 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). For the experiments detailed in this paper, two items were added to the questionnaire, specific to the inclusion of TTS, one referring to comprehension of the voice ("It was sometimes difficult to understand what the service was saying") and the other referring to the information provided in the system ("I thought the service provided enough information"). See Appendix A for a full listing.

In the approach used for the four experiments, 7-point Likert attitude scales were used with a balance of positively and negatively worded stimulus statements in the questionnaire. On this scale, once the responses are normalised for statement polarity, a score over 4.0 represents a positive attitude; scores below 4.0 represent negative attitudes to the identified attributes. Overall usability scores are obtained by taking the mean of all the items in the questionnaire. 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.

Statistical analysis of the data is carried out using parametric tests, since there is considerable evidence to suggest that such tests are robust to any potential violations of their underlying assumptions (Baker *et al*., 1966; Box, 1953; Kim, 1975; Labovitz, 1967, 1970; O'Brien, 1979) and they are generally held to have greater power to detect effects than their nonparametric equivalents (Field, 2000).

Performance measurements include task success rates and the time taken to complete tasks. Detailed information on participant behaviour at each stage (e.g. type of response – speech, DTMF or none etc) together with information on any errors made

by the speech recognition engine are also available where these are relevant. However, this was not the focus of this research and a result these data are not reported here. Whilst it is possible that recognition errors can affect user attitudes towards the service under test, previous experience suggests the level of recognition errors is low. Moreover, where the recognition grammars are the same across versions (as in this series of experiments) the recognition errors can be expected to be broadly evenly distributed across the different versions. Any effects on user attitude due to speech recognition errors, therefore, are averaged out across versions, allowing any differences in attitude that are due to the design differences to be successfully isolated and identified.

In addition to performance data and user attitudes, the approach also provides for the gathering of qualitative data through the use of structured interviews with participants after they have completed all their tasks. Data gathered from these interviews can be very useful in providing insights into why participants responded in the ways they did.

4. Experiment 1: Extent of TTS inclusion

The aim of Experiment 1 was to explore participants' attitudes towards the use of TTS synthesis in the recent transaction listings in the automated telephone banking system. This experiment compared versions of telephone banking where the recent transactions listings are read out with recorded speech, TTS or a mixture of both. In the cases where TTS was used, additional detail on the transaction was provided to the user; in the version with only recorded speech, no additional detail was provided on the transaction. A repeated measures design compared three different versions of the service.

4.1. Versions compared

The three versions of the service were based on the existing telephone banking service of a major UK bank, and differed only in the recent transactions section of the service.

> Version A: Fully recorded with non-specific transaction details, as in the existing service. For example:
>
> "*on the 3rd of March a debit for £55.00*"
>
> Version B: Recorded speech for the carrier phrases, dates and amounts, and using TTS for the additional detail information. For example (underlined would be played in TTS):
>
> "*on the 3rd of March a debit for £55.00 to The Gift Shop*"
>
> Version C: Fully TTS for whole transaction (carrier phrases, dates, amounts and additional detail information.) For example (underlined would be played in TTS):
>
> "*on the 3rd of March a debit for £55.00 to The Gift Shop*"

In this case, where whole phrases were read out using TTS in one of the versions, both the recorded and TTS prompts employed the same female Southern British English voice (that of the standard UK voice supplied with the TTS engine). Each participant made two calls to each of the three versions described above. The rest of the banking service was identical for all three versions; only the transaction listings differed.

As TTS allows particular additional information to be given on a transaction listing, its usefulness depends on the task, that is, on the transaction listing being looked for. Giving a task that specifies which retailer a transaction has been made to will maximise the usefulness of the additional information and may bias the results towards TTS. Giving a task that does not specify the retailer information may bias the results against TTS. In this experiment, a range of tasks were included both with and without the additional detail on the task sheet. Each participant made two calls per service design, one with a task scenario that did not specify the extra information, e.g.

> "Listen to the list of recent transactions and find out if a bill payment debit for £45.00 has come out of your account yet."

And one with a task scenario that did refer to the additional detail, e.g.

> "Listen to the list of recent transactions and find out if a bill payment debit for £45.00 you made to Vodafone has come out of your account yet."

The orders of presentation of the different designs and task types were balanced with respect to each other and to the other experimental variables of age and gender in order to achieve a fully balanced design.

4.2. Participants

A cohort of 94 participants was recruited in Edinburgh. All were customers of the Case Bank. There were 43 male participants and 51 female participants. A breakdown of the 94 participants by age group and gender is given in Table 2. The age groups chosen were designed to reflect the profile of the banking customers represented in the recruitment database.

	Age group 1 (18-44 years)	**Age group 2 (45+ years)**	**Total**
Male	21	22	**43**
Female	24	27	**51**
Total	**45**	**49**	**94**

Table 2: Participant Cohort by Gender and Age Group – Experiment 1

4.3. Procedure

This experiment adopted a mixed within-subjects and between-subjects design in which each participating customer used all three versions of the service, with the order of experience of the versions being balanced and randomised across the cohort of participants.

The inclusion of the additional detail provided in TTS is particularly useful when disambiguating two transactions of the same amount. Therefore the experiment design included a between-subjects variable of duplicate amount. Participants were allocated to one of two groups - half the cohort were given task scenarios where the transaction amount was unique in the listings, while the other half were asked, in one of their two calls to each version, to search for a transaction where there was more than one instance of the amount in the transaction listings. The use of an identical transaction amount occurred in only those task scenarios that specified the additional detail which could potentially be used to disambiguate the amount. A between-

subjects design was adopted for this variable to reflect the fact that in real life a duplicated amount is likely to occur less frequently.

Participants made two calls to each version, in each call being asked to search for a recent transaction listing (as described above, one with the additional detail specified, one without). A usability questionnaire was completed after each call and there was a structured interview at the end of the session which allowed participants the opportunity to make comments on each of the versions they experienced.

4.4. Results

The scores of each of the 22 usability attributes were averaged to obtain an overall usability score for each version. Results are shown in Table 3 (by version and call number) and in Table 4 (by version and task type).

Service Version	Call 1	Call 2	Mean Score
A (All Recorded)	5.20 (SD=0.99)	5.22 (SD=1.01)	**5.21 (SD=0.84)**
B (Additional detail, Mixed)	5.29 (SD=1.00)	5.31 (SD=1.00)	**5.30 (SD=0.79)**
C (Additional detail, All TTS)	5.24 (SD=1.00)	5.23 (SD=1.02)	**5.24 (SD=0.89)**

Table 3: Mean Usability Scores by Version and Call Number

Service Version	Task: additional detail not specified	Task: additional detail specified	Mean Score
A (All Recorded)	5.29 (SD=0.90)	5.13 (SD=0.92)	**5.21 (SD=0.84)**
B (Additional detail, Mixed)	5.28 (SD=0.84)	5.32 (SD=0.85)	**5.30 (SD=0.79)**
C (Additional detail, All TTS)	5.25 (SD=0.91)	5.23 (SD=0.96)	**5.24 (SD=0.89)**

Table 4: Mean Usability Scores by Version and Task

A repeated measures analysis of variance (ANOVA) was applied to the mean usability scores (computed on the full set of 22 attributes), with *service version* and *call number* as within-participants factors, and *version order* (any of six possible orders), *task type order* and *duplication of the target amount* as between-participants factors.

This yielded no significant main effect of service version (p=0.315) or of call number (p=0.954): the only significant main effect was that of amount duplication (p=0.017), with participants who had a duplicate transaction giving lower usability scores than those who had none. There was a significant interaction of call number, task type order and duplication (p=0.012), and the interaction of version, call number and task type order was nearly significant (p=0.055). An interaction of call number and task type order is exactly equivalent to a main effect of task type, and so these results are equivalent to a significant interaction of task type and duplication (p=0.012) and a nearly significant interaction of version and task type (p=0.055). The difference between participants with and without a duplicate transaction was greater on tasks which specified the additional detail than on those tasks that did not; this is as

expected since the duplication directly affected those tasks where the payer/payee information was known to the user. The interaction of version and task type was also as expected, with detail-specified tasks in Version A (where the transaction details given in the task were not read out in full by the service) yielding lower scores than all the other combinations.

A second ANOVA was run on the mean usability scores, with the same factors as above except that version order was omitted and the additional between-participants factors *age group* and *gender* were included. (The sample size was not sufficient for all the factors to be included in a single analysis.) Again the main effect of amount duplication was significant (p=0.013) and so was the interaction of task type and duplication (p=0.007), while the interaction of version and task type approached significance though it was weaker than in the first analysis (p=0.091). The interaction of version and age group was also significant (p=0.025), with the younger participants (aged 18-44) giving lower scores to Version C than to the other versions, but the older group (45+) giving lower scores to Version A than to Version B or C.

The scores for each individual attribute were analysed in a similar way, using the same set of factors as in the second ANOVA on the mean scores. The main effect of service version was significant for 6 of the 22 attributes. Version A (recorded speech only) was rated significantly better than Version C (full use of TTS for recent transactions) for *voice clarity* (p=0.001), *friendliness* (p=0.035) and *liking the voice* (p=0.006), but significantly poorer than Version C for *efficiency* (p=0.014), *improvement needed* (p=0.048) and *providing enough information* (p=0.001). Version B (use of TTS for proper names only) was rated significantly above Version A on *efficiency* (p=0.012), *improvement needed* (p=0.016) and *providing enough information* (p<0.001), and significantly above Version C on *friendliness* (p=0.026) and *liking the voice* (p<0.001); Version B did not score significantly below either of the others on any of the attributes.

The strongest effects of version occurred for the attributes *voice clarity* (p=0.005), *liked voice* (p<0.001) and *enough information* (p<0.001), and were in the expected directions – the voice being found less clear and being less liked when the whole of each transaction was given by TTS (Version C) than when only recorded speech was used (Version A), but the versions with TTS providing additional detail on the transactions scoring above the recorded speech version in terms of the amount of information given.

The main effect of task type was highly significant for the attributes *liked voice* (p=0.005) and *enough information* (p=0.004), and marginally significant for *confusion* (p=0.034); in each case those tasks where no payee details were stipulated yielded more favourable scores for all versions of the service than the tasks in which the payee information was specified, except for Version B on *confusion* where the results were more favourable for the tasks in which the payee information was specified.

The effect of age group was significant for 11 of the 22 usability attributes, with participants aged 18-44 (age group 1) giving generally lower scores than those aged 45 and over (age group 2). There were also significant interactions of age group and version for six attributes – the older participants giving higher scores to the versions with TTS (especially Version C) relative to Version A, than the younger group. The results suggest that the younger participants were more sensitive to the differences in

the voice, and in particular less tolerant to the extended use of TTS (Version C), than the older participants.

4.5. Quality ratings

To collect a quality rating for each of the versions experienced, participants were asked to order and rate each version by preference by placing markers on a scale marked from 0 (worst) to 30 (best). The mean rating scores (out of 30) for each of the three versions were as follows: Version B (mixed recorded and TTS) scored highest with 20.6, followed by Version C (all TTS) at 19.4. Version A (recorded throughout, with no additional detail) scored 17.1.

A repeated measures analysis of variance (ANOVA) was performed on the quality ratings, with *service version* as the within-participants factor, and *age group*, *gender*, *task type order* and *duplication of the target amount* as between-participants factors. The main effect of version was found to be significant (p=0.009), with a significant pairwise difference between Versions A and B (p=0.003) and a nearly significant difference between Versions A and C (p=0.057). The only other significant effect was a main effect of age group (p=0.016), with age group 1 (18-44 years old) giving lower scores to all three versions than age group 2 (45+).

4.6. Performance data

Results were obtained in respect of task completion and call duration. In terms of task performance, it was found that a higher proportion of participants succeeded in finding a specified transaction in the versions with the additional detail (Versions B and C) than in the version without it (Version A) – both when the task scenario gave the additional detail of the transaction and, more surprisingly, when the task gave only the transaction type and amount. When another transaction for the same amount as the target transaction was present in the list, Version A did not provide enough information to enable participants to distinguish between these, and most participants listened only to the more recent transaction; even those who listened to both transactions could not be sure whether either of them was the one they were looking for. Versions B and C, in contrast, provided the information to distinguish the duplicate transaction (same amount, different detail) from the target transaction, and in these versions 80% of participants continued listening after the duplicate transaction and thus found the transaction specified in the task.

Calls to Version A were shorter than calls to Versions B and C, partly because the transaction descriptions in Version A contained less information and therefore fewer words, and partly because participants tended to listen to fewer transactions in Version A, especially in the cases with a duplicate transaction. Calls to Version B took longer than calls to Version C, mainly because the natural speech versions of the transaction descriptions were longer (i.e. slower in pace, counting any pauses) than the TTS versions.

4.7. Discussion

This experiment compared the form of transaction readout used in an existing telephone banking service – consisting of the date, transaction type and amount, with no additional detail, given using recorded natural speech throughout – against two versions which made use of TTS synthesis in order to provide additional detail on the transaction listing. The main focus of the experiment was on whether participants

would find such additional detail worth having, given that it could not be provided using natural speech throughout and therefore required the introduction of a synthetic voice. A second aim of the experiment was to compare the two styles of speech used in the versions with additional detail: using TTS for the proper names only, with the date, transaction type and amount remaining in natural speech (Version B), and using TTS for the whole of the transaction description (Version C).

Overall scores in the usability questionnaire were highest for the service which included the additional detail with minimal TTS and lowest for the version which employed recorded natural speech only with no additional detail; however, the differences between versions were not statistically significant. On some specific usability attributes, there were significant differences, with the fully TTS version scoring significantly below the fully recorded version for *clarity of the voice* and significantly below both other versions for *liking the voice* and *friendliness*, and the fully recorded version scoring significantly poorer than the other versions (which included TTS for additional detail) for *providing enough information*, *efficiency* and *needing improvement*.

Ratings on a quality scale (given after listening to an example of the transaction announcement in each version) were significantly higher for Version B than for Version A, with the difference between Version C and Version A approaching significance. In conjunction with the usability scores, this provides some evidence that participants prefer Version B (additional detail with minimal use of TTS) over Version A (natural speech with no additional detail). This is supported by the preferences expressed during the one-to-one interview when the differences between the versions had been explained: 78.7% of participants said they would prefer additional detail given by TTS over no additional detail, against 16.0% expressing the opposite preference.

The main practical conclusion is that using TTS to provide additional detail in recent transaction listings would improve the recent transactions dialogue in a telephone banking system from the customer's point of view. It appears also that it is best to keep the use of TTS to a minimum, by using it only for proper names (personal, company or place names) rather than for the whole of the transaction description – a finding which is consistent with results of previous research (McInnes et al, 1999; Möller et al, 2006).

5. Experiment 2: Minimal use of TTS

A second experiment was conducted to explore participants' attitudes towards the use of TTS minimally in the recent transaction listings for additional detail on some transactions only. This experiment compared two versions of the telephone banking service where additional detail on transactions was available for only some transactions in one version (all recorded) and for all transactions (some recorded, some using TTS synthesis) in the other version.

Recordings of the 100 most frequently used company names (as detailed in data provided by the Case Bank) were made by the voice recording artist and were used in both versions of the service experienced in this experiment. When transactions were made with these companies, the additional detail was provided using recorded speech for both versions. For all other companies, the additional detail was provided using TTS for one version and was not available in the other version. A repeated measures design compared the two different versions of the service.

5.1. Versions compared

The two versions of the service were based on the existing telephone banking service, and differed only in the recent transactions section of the service.

> Version A: Fully recorded with specific transaction details provided for the most frequently used companies; no detail provided for others. For example:
>
> "*on the 23rd of May a debit for £39.74 to Tesco*
>
> *and on the 21st of May a debit for £49.00.* "
>
> Version B: Additional detail provided for all transaction listings using recorded speech for the most frequently used companies and TTS synthesis for all other company names. For example (underlined would be played in TTS):
>
> "*on the 23rd of May a debit for £39.74 to Tesco*
>
> *and on the 21st of May a debit for £49.00 to Brora.* "

In this case, where use of TTS was kept to a minimum the recorded and TTS prompts employed different voices, albeit both were female with Southern British English accents. Version A employed natural speech throughout using speech prompts recorded by the existing service's voice talent. Version B used recorded speech prompts as in version A, supplemented with TTS prompts in the standard UK female voice supplied with the TTS engine where necessary.

5.2. Participants

A cohort of 75 participants was recruited in Edinburgh; there were 37 male participants and 38 female participants. All were customers of the Case Bank. A breakdown of the 75 participants by age group and gender is given in Table 5.

	Age group 1 (18-44 years)	**Age group 2 (45+ years)**	**Total**
Male	18	19	**37**
Female	17	21	**38**
Total	**35**	**40**	**75**

Table 5: Participant Cohort by Gender and Age Group – Experiment 2

5.3. Procedure

Each participant made two calls to both versions in a repeated-measures design. As in the first experiment, two task types were used in this experiment; one did not specify the company information on the participants' task sheet and one did refer to this additional detail. In addition, in order to provide further exposure to the version types the experiment tasks requested that the participant find out if a number of different transactions (specifically 4 different transactions) had occurred on their account within the same telephone call.

The orders of presentation of the different designs and task types were balanced with respect to each other and to the other experimental variables of age and gender in order to achieve a fully balanced design.

Again, as before, participants completed a 22-item usability questionnaire after each call they made to the telephone banking service. Participants were given the opportunity to make comments on the versions of the service during a structured interview at the end of the session.

5.4. Results

The scores of each of the 22 usability attributes were averaged to obtain an overall usability score for each version. Results are shown in Table 6 (by version and call number) and in Table 7 (by version and task type).

Service Version	Call 1	Call 2	Mean Score
A (Detail on Top companies only - recorded)	5.28 (SD=0.94)	5.35 (SD=0.88)	**5.32 (SD=0.88)**
B (Detail on all – Top recorded, others TTS)	5.29 (SD=0.88)	5.30 (SD=0.82)	**5.30 (SD=0.80)**

Table 6: Mean Usability Scores by Version and Call Number

Service Version	Task: info not specified on task	Task: info specified on task	Mean Score
A (Detail on Top companies only - recorded)	5.38 (SD=0.88)	5.25 (SD=0.94)	**5.32 (SD=0.88)**
B (Detail on all – Top recorded, others TTS)	5.30 (SD=0.86)	5.29 (SD=0.84)	**5.30 (SD=0.80)**

Table 7: Mean Usability Scores by Version and Task Type

A repeated measures analysis of variance (ANOVA) was applied to the mean usability scores (computed on the full set of 22 attributes), with *service version* and *task type* as within-participants factors, and *version order*, *task type order*, *age* and *gender* as between-participants factors. This yielded no significant main effect of service version (p=0.746) or of task type (p=0.191).

The scores for each individual attribute were analysed in a similar way, using the same set of factors as on the overall mean scores. The main effect of service version was significant for 3 of the 22 attributes. Version A (Detail on Top companies only - recorded) was rated significantly better than Version B (Detail on all – recorded for Top companies, others in TTS) for *voice clarity* (p=0.044), and *difficulty to understand* (p=0.045), that is, Version A was significantly easier to understand than Version B. However, Version A was rated as significantly poorer than Version B for *providing enough information* (p=0.008). These results are to be expected as Version A did not include any TTS, therefore the attributes of voice clarity and ease of understanding are scored higher for this version. However, Version A only provided the additional detail for some transactions and therefore scored significantly lower on providing enough information compared with Version B.

A main effect of task was significant for the attribute *provided enough information* (p=0.021). For both versions, those tasks in which the company name was not

specified to the user prior to their call yielded more favourable scores than those which included detail of the transaction stipulated. In addition it was found that Version B scored slightly higher than Version A on the attribute *provided enough information* for those tasks where the information was not specified in advance.

The effect of age group was significant for 6 of the 22 usability attributes. For the attributes *concentration*, *stress*, *service is too fast*, *preference for a human*, and *difficult to understand* the younger participants (aged 18-44) gave higher scores than the older participants (aged 45 and over). That is, younger participants felt the services required less concentration, younger participants were less stressed when using the service, they were less likely to think the service was too fast, they were less likely to prefer a human and they were less likely to think the service was difficult to understand. In contrast, older participants gave higher scores for the attribute *liked the voice*; that is, the older participants indicated a stronger liking for the voice than the younger participants.

5.5. Quality ratings

As before, participants were asked to order and rate the two versions on a scale from 0 to 30. Average scores from the quality ratings show that the minimal TTS version scored higher at 21.5 than the other version at 19.0. A repeated measures analysis of variance (ANOVA) was performed on the quality ratings, with *service version* as the within-participants factor, and *age group*, *gender* and *task type order* as between-participants factors. The main effect of version was found to be significant (p=0.011).

5.6. Performance data

In terms of task performance, there was no difference between the two versions of the service. The experiment criteria for a successful call were that the participant heard at least one target transaction in each call they made. Taking into consideration those calls where the participant required another attempt, all participants in the experiment were able to successfully proceed through each of their four calls to the service. The task scenarios detailed four transactions to be searched for in each call. All four transactions were found in the majority of calls (87%).

Calls to Version A were shorter than calls to Version B which is to be expected given that Version B provided the additional detail on transaction listings for all transactions and Version A only for some transactions. Overall, the call duration was generally longer for calls where more transactions were heard as would be expected.

5.7. Discussion

This experiment compared two versions of recent transactions readout in a telephone banking service where additional detail on the transactions was provided. In one version additional detail on transactions was available for some transactions simulating the most frequently used companies and this company name detail was played using recorded speech. In the other version, the additional detail was provided for all transactions where the most frequently used companies' names were played using recorded speech and all other company names were played using TTS synthesis. The main focus of the experiment was on whether participants would prefer the additional detail on all transactions using TTS or an approach which avoided synthesised speech but could only provide detail on some transactions.

Although overall mean scores in the usability questionnaire were marginally higher for the fully recorded version, the difference was not statistically significant. Significant differences were found on three usability attributes with Version B (with TTS) scoring significantly above Version A on *providing enough information* and significantly below Version A for *clarity of the voice* and *difficulty to understand*.

The majority of participants, when asked during the interview to express a preference, stated they had no preference between the versions; although, of those participants who stated a preference there was a slight majority for the TTS version with detail on all transactions over the fully recorded version with detail on only some transactions. However, ratings on a quality scale (given after listening to an example of some transaction listings in each version) were significantly higher for the TTS version with detail on all transactions than for the fully recorded version with detail on only some transactions.

Thus although the overall usability scores indicated marginal differences between the versions, when the differences between the versions have been explained there is a preference for the TTS version. This therefore again supports the inclusion of TTS to provide additional detail even compared with a version of the telephone banking system which can provide the detail for some transactions.

6. Experiment 3: Location of additional detail

The aim of experiment 3 was to explore participants' attitudes towards the location of additional detail (company names) in the recent transaction listings in the automated telephone banking system. This experiment compared two versions of telephone banking where additional detail on transactions was provided at the end of the transaction listing, after the amount information, or in the middle of the transaction listings, before the amount information.

Following on from the previous experiments, the additional detail was provided using recorded speech for the most frequently used companies and using TTS for all other company names. A repeated measures design compared the two different versions of the service.

6.1. Versions compared

The two versions of the service differed only in the location of the additional detail in the recent transactions section of the service.

> Version A: Additional detail, whether recorded or TTS, given at the end of the transaction listing, after the amount information. For example (underlined would be played in TTS):
>
> "*on the 12th of June a direct debit for £44.85 to Vodafone*
>
> *and on the 14th of June a debit for £49.00 to Brora.* "
>
> Version B: Additional detail, whether recorded or TTS, given in the middle of the transaction listing, before the amount information. For example (underlined would be played in TTS):
>
> "*on the 12th of June a direct debit to Vodafone for £44.85*
>
> *and on the 14th of June a debit to Brora for £49.00.* "

As in the previous experiment, all natural speech prompts were recorded by the voice talent employed in the existing live service, while the TTS prompts were in the standard UK voice supplied with the TTS engine. Both were female with Southern British English accents.

6.2. Participants

A cohort of 66 participants was recruited in Edinburgh; there were 31 male participants and 35 female participants. All were customers of the Case Bank. A breakdown of the 66 participants by age group and gender is given in Table 8.

	Age group 1 (18-44 years)	**Age group 2 (45+ years)**	**Total**
Male	13	18	**31**
Female	17	18	**35**
Total	**30**	**36**	**66**

Table 8: Participant Cohort by Gender and Age Group – Experiment 3

6.3. Procedure

This experiment adopted a within-subjects design, in which each participating customer used both versions of the service.

In addition to the two versions being compared, the experiment design included a variable of duplicate amount in the transaction listings as the inclusion of the additional information provided on company name would be particularly useful when disambiguating two transaction types of the same amount. Each participant made two calls to each of the versions. In one of their calls to each version, participants were given task scenarios where the transaction amount they were asked to search for was unique in the listings. In the other call to each version, participants were asked to search for a transaction of a particular amount where there was more than one instance of this amount in the transaction listings.

As in the previous experiment, participants were asked to search for a number of transaction listings (specifically 4) in each call.

The orders of presentation of the different designs and task types were balanced with respect to each other and to the other experimental variables of age and gender in order to achieve a fully balanced design.

Again, as before, participants completed a usability questionnaire after each call they made to the telephone banking service. Participants were given the opportunity to make comments on the versions of the service during a structured interview at the end of the session.

6.4. Results

Participants completed two usability questionnaires for each version of the service: one after the first call to this version and one after the second call. The scores of each of the 22 usability attributes were averaged to obtain an overall usability score for each version. Results are shown in Table 9 (by version and call number) and in Table 10 (by version and task type).

Service Version	Call 1	Call 2	Mean Score
A (Sentence-Final Detail)	5.39 (SD=0.91)	5.42 (SD=0.87)	**5.41 (SD=0.85)**
B (Sentence-Medial Detail)	5.24 (SD=1.03)	5.35 (SD=0.89)	**5.30 (SD=0.92)**

Table 9: Mean Usability Scores by Version and Call Number

Service Version	Task: unique amount	Task: duplicate amount	Mean Score
A (Sentence-Final Detail)	5.37 (SD=0.95)	5.44 (SD=0.83)	**5.41 (SD=0.85)**
B (Sentence-Medial Detail)	5.31 (SD=0.99)	5.29 (SD=0.94)	**5.30 (SD=0.92)**

Table 10: Mean Usability Scores by Version and Task Type

A repeated measures analysis of variance (ANOVA) was applied to the mean usability scores (computed on the full set of 22 attributes), with *service version* and *task type* as within-participants factors, and *version order*, *task type order*, *age* and *gender* as between-participants factors. This yielded no significant main effect of service version (p=0.073) or of task type (p=0.586) and no interactions were found to be significant.

The scores for each individual attribute were analysed in a similar way, using the same set of factors as on the overall mean scores. The main effect of service version was significant for 3 of the 22 attributes. Version A (sentence-final detail) was rated significantly better than Version B (sentence-medial detail) for *concentration* (p=0.021), *stress* (p=0.038) and *complication* (p=0.023), that is, the sentence-medial version required more concentration, was more stressful to use and was more complicated than the sentence-final version.

The interaction of version and order was highly significant (p=0.002) for the attribute *concentration* where participants scored the second version of the service they tried higher than the first version. This was particularly the case for participants who experienced Version B (sentence-medial) followed by Version A (sentence-final).

6.5. Quality ratings

As before, participants were asked to order and rate the two versions on a scale from 0 to 30. Average scores from the quality ratings show that the sentence-medial TTS version scored slightly higher at 22.0 than the sentence-final TTS version at 21.2. A repeated measures analysis of variance (ANOVA) was performed on the quality ratings, with *service version* as the within-participants factor, and *age group*, *gender* and *task type order* as between-participants factors. The main effect of version was not found to be significant (p=0.240) and no other significant effects were found.

6.6. Performance data

In terms of task performance, there was no difference between the two versions of the service. The experiment criteria for a successful call were that the participant heard at least one target transaction in each call they made. Taking into consideration any calls where the participant required another attempt, all participants in the experiment were able to successfully proceed through each of their four calls to the service. The task scenarios detailed four transactions to be searched for in each call. All four

transactions were heard by participants in 67.4% of calls. In a substantial minority of calls (21.2%) only one transaction was heard, although it was found that this was usually participant specific in that if the participant chose to search for only one transaction in one of their calls, they did so in all four of their calls.

Overall, the call duration was generally longer for calls where more transactions were heard as would be expected. The duration was also slightly higher for Version B and tasks where there was a duplicate amount in the transaction listings. This was due to a slight increase in participants asking for the transactions to be 'repeated'.

6.7. Discussion

This experiment compared the location of additional detail provided using TTS in the recent transactions readout in a telephone banking service. In one version additional detail on transactions, specifically company names, was provided at the end of the transaction listing (after the amount information). In the other version, additional detail on transactions was provided in the middle of the transaction listing (before the amount information). The main focus of the experiment was on whether participants would prefer this additional detail at the end or in the middle of the transaction listing.

Overall usability scores were higher for the sentence-final version (mean score 5.41, on a scale from 1 to 7) than for the sentence-medial version (5.30); however, the difference was not statistically significant. Significant differences were found on three usability attributes with the sentence-final version being rated significantly better than the sentence-medial version for *concentration*, *stress* and *complication.* So, the sentence-medial version required more concentration, was more stressful to use and was more complicated than the sentence-final version.

The majority of participants, when asked in the interview to express a preference, stated they had no preference between the two versions; and of those participants who stated a preference there was an almost equal split between the two versions. Ratings on a quality scale (given after listening to examples of transaction listings in each version) also showed little difference between the two versions. Therefore, from these results it would suggest that the location of the detail, which was provided in TTS for some transactions, is equally usable both in a sentence-medial position and in a sentence-final position.

7. Experiment 4: Readout of company names

The aim of experiment 4 was to explore participants' attitudes towards the readout of company names as part of the additional detail in the recent transaction listings in the automated telephone banking system. The Case Bank indicated that the company information returned to their systems as part of a debit or credit transaction do not always follow a systematic or standard format. For example, when a customer completes a transaction with the company 'National Express' this information could be returned as "Nat Express E Cst". Although this same information would be provided on a customer's statements, either on paper or Internet banking, it was felt it might be difficult for the customer to comprehend this detail when hearing it via text-to-speech in the automated telephone banking service.

This experiment compared two versions of telephone banking: one in which the additional detail consisted of the exact company name, and one in which it was a malformed version of the company name, as returned to Case Bank systems following the transaction. In both versions, the additional detail was provided using TTS for all

company names. A repeated measures design compared the two different versions of the service.

7.1. Versions compared

The two versions of the service differed only in the format of the additional detail in the recent transactions section of the service.

> Version A: Exact company name in the enhanced detail. For example (underlined would be played in TTS):
>
> "*on the 12th of January a direct debit for £44.85 to Amazon*
>
> *and on the 11th of January a debit for £49.00 to Tesco.* "
>
> Version B: Malformed version of the company name in the enhanced detail. For example (underlined would be played in TTS):
>
> "*on the 12th of January a direct debit for £44.85 to Amazon SVCS EU-UK*
>
> *and on the 11th of January a debit for £49.00 to Tesco Store 2920.* "

All natural speech prompts were recorded by the voice talent employed in the existing live service, while the TTS prompts used the standard UK voice supplied with the TTS engine. Both were female with Southern British English accents.

Data from the Case Bank indicate that malformations of the company name vary considerably, and can consist of the addition of acronyms, alphanumeric or numeric codes (as in the above examples) and/or shortened or lengthened forms of the company name (e.g. "*Carphone Warehse*" in place of "*Carphone Warehouse*" or "*Waterstones Book Selle*" in place of "*Waterstones*"). Note that the TTS engine's default treatment of numeric codes is to read them out as a number ("*Tesco Store twenty-nine twenty*" in the above example) rather than as a string of digits. Acronyms are read out as a sequence of individual letters e.g. "Amazon S-V-C-S-E-U-U-K" in the above example.

In addition to the two versions being compared, two types of the recent transactions task were used in the experiment. In one of their calls to each version, participants were given a task scenario where the company name was already known and the amount of the transaction was not. An example is shown below:

> *Listen to your list of recent transactions and find out if a debit card payment to Tesco has come out of your account.*

In the other call to each version, participants were asked to search for a transaction of a particular amount where the company name was unknown. In this task type, participants were asked to write down the name of the company involved in the transaction. An example is shown below:

> *Listen to your list of recent transactions and find out to whom a debit card payment for £18.50 was made.*
>
> *This payment was made to ……*

A second task variable, 'malformation type', was also employed in the experiment. To reflect the real-life data provided by the Case Bank, the malformation of the target transaction in one call involved the addition of a numeric code to the company name

e.g. "*All Bar One 160210*" (read out as "*All Bar One, one hundred and sixty thousand two hundred and ten*"), whilst in the other the malformation was of the name itself or involved the addition of an acronym e.g. "*Claire's Access Lt*" (read out as "*Claire's Access el-t"*).

In order to achieve a fully balanced experiment design the orders of presentation of the different service designs, task types and malformation types were balanced with respect to each other and to the other experimental variables of age and gender.

7.2. Participants

A cohort of 70 participants was recruited in Edinburgh; there were 33 male participants and 37 female participants. All were customers of the Case Bank. A breakdown of the 70 participants by age group and gender is given in Table 11.

	Age group 1 (18-44 years)	**Age group 2 (45+ years)**	**Total**
Male	16	17	**33**
Female	15	22	**37**
Total	**31**	**39**	**70**

Table 11: Participant Cohort by Gender and Age Group – Experiment 4

7.3. Procedure

This experiment adopted a within-subjects design, in which each participating customer used both versions of the service, with the order of experience of the versions being balanced across the cohort of participants. Participants made two calls to each version, in each call being asked to search for recent transaction listings. Before each call participants were also talked through a written scenario that described their recent shopping activity. This scenario involved a number of different companies, including the one involved in the task (the 'target' transaction). Use of the scenario was designed to reflect real life, where customers are aware of their recent shopping behaviour and would expect to recognise any activity on their account.

In one call to each version the company name information was known to the user; in the other call to each version, only the amount was known to the user. Again, as before, participants completed a usability questionnaire after each call they made to the telephone banking service, and took part in a structured interview at the end of the session.

7.4. Results

Participants completed two usability questionnaires for each version of the service: one after the first call to this version and one after the second call. The scores of each of the 22 usability attributes were averaged to obtain an overall usability score for each version. Results are shown in Table 12 (by version and call number) and in Table 13 (by version and task type).

Service Version	Call 1	Call 2	Mean Score
A (Exact Name)	5.20 (SD=0.86)	5.18 (SD=0.89)	**5.19 (SD=0.81)**
B (Malformed Name)	4.78 (SD=1.03)	4.80 (SD=1.10)	**4.79 (SD=1.03)**

Table 12: Mean Usability Scores by Version and Call Number

Service Version	Task: company name known	Task: amount known	Mean Score
A (Exact Name)	5.06 (SD=0.91)	5.32 (SD=0.81)	**5.19 (SD=0.81)**
B (Malformed Name)	4.71 (SD=1.09)	4.87 (SD=1.05)	**4.79 (SD=1.03)**

Table 13: Mean Usability Scores by Version and Task Type

A repeated measures analysis of variance (ANOVA) was applied to the mean usability scores (computed on the full set of 22 attributes) for the two calls to the alternative services, with *service version* and *task type* as within-participants factors, and *version order* (Exact-Malformed or Malformed-Exact in the pair of calls to each version), *task type order*, *age* and *gender* as between-participants factors.

This yielded a highly significant main effect of service version ($p<0.001$). The Exact version was rated significantly more usable than the Malformed version. The magnitude of the difference in scores here is considerable.

There was also a highly significant main effect of task type ($p<0.001$). When carrying out tasks in which only the company name was known participants found the service significantly less usable than during tasks in which only the amount was known. This is consistent with data on participants' performance and behaviour, in which significant numbers of participants were found to have attempted to search by company name for those tasks which detailed the name, when no such option exists within the banking service.

The scores for each individual attribute were then analysed in a similar way, using the same set of factors as on the overall mean scores. The main effect of service version was significant for 18 of the 22 attributes, 15 of which were highly significant results ($p<0.01$). The Exact version was rated significantly better than the Malformed version in each case. The difference was particularly marked for the issues *needs improvement* and *difficult to understand*, with the Malformed version rated around neutral on both attributes.

The only four attributes for which the effect of version was *not* significant were *confusion*, *degree of control, too fast* and *polite*. Otherwise a consistent pattern of difference was found across attributes, leading to the highly significant difference in mean usability scores overall.

The main effect of task type was significant for 15 of the 22 attributes, 8 of which were highly significant ($p<0.01$). The attributes for which there was a highly significant difference were: *confusion*, *stress*, *knew what to do*, *ease of use*, *efficient*, *friendly*, *enjoyment* and *enough information*. In each case, the service was rated significantly less usable for tasks where only the company name was known than for

tasks where only the transaction amount was known, which as described earlier, is consistent with the user behaviour and performance data.

7.5. Quality ratings

As before, participants were asked to order and rate the two versions on a scale from 0 to 30. Average scores from the quality ratings show that the Exact version scored considerably higher at 23.6 than the Malformed version at 13.1. A repeated measures analysis of variance (ANOVA) was performed on the quality ratings, with *service version* as the within-participants factor, and *version order*, *age group*, *gender* and *task type order* as between-participants factors. This yielded a highly significant main effect of service version ($p<0.001$) with the Exact version rated considerably higher than the Malformed version. The magnitude of the difference here is substantial.

There was a significant between-participants effect for order and age ($p=0.018$) indicating that older participants gave lower scores overall when they experienced the order Exact-Malformed than those who experienced the order Malformed-Exact, whilst younger participants scored the services in similar way regardless of order.

7.6. Performance data

In terms of performance both versions performed equally well, with a mean task completion rate of 92.9% in the Exact version and 95.7% in the Malformed version. Mean call duration, however, was shorter in the Exact version (152s), than in the Malformed version (165s). Since participants completed the two task types in a similar way in each version, this suggests that the read-out of malformed company names, with additional numerical codes and/or acronyms, took longer than that of exact names.

7.7. Discussion

This experiment compared the readouts of company names forming the additional detail in the recent transactions listings in the telephone banking service. The additional detail was provided using TTS in both cases. In one version it consisted of the exact name of the company involved in the transaction. In the other it was a malformed version of the company name, as returned to the Case Bank systems at the point of the transaction.

In terms of performance, there was very little difference between the two versions of the service. Both versions resulted in a similar (low) number of failed calls. In all cases of failure participants were given further attempt(s) as necessary and eventually succeeded in hearing at least one transaction, thus enabling them to continue with the experiment. Task completion rates in the resulting complete calls were similarly high in both versions and for both task types, with a mean task completion rate of 92.9% in the Exact version and 95.7% in the Malformed version. Most task failures (twelve out of sixteen) were because the participant hung up before the target group had been reached during the service's read out of transactions.

Overall, call duration was lower in the Exact version (152s) than the Malformed version (165s). This can be attributed to the fact that the read-out of malformed company names, with additional numerical codes and/or acronyms, took longer than that of exact names. In calls where the task was completed the effect was found to be highly significant ($p=0.001$).

In terms of usability, the Exact version was rated significantly more usable overall than the Malformed version. The mean scores were 5.19 and 4.79 respectively (on a scale from 1 to 7), and the difference was highly significant ($p<0.001$). The Exact version was rated significantly higher than the Malformed version on 18 of the 22 individual attributes, 15 of which were highly significant results ($p<0.01$). The difference was particularly marked for the issues *needs improvement* and *difficult to understand*, with the Malformed version rated poorly on both.

There was also a highly significant main effect of task type in the overall means. The effect of task type was significant for 15 of the 22 attributes, 8 of which were highly significant ($p<0.01$). In each case, the service was rated significantly less usable for tasks where only the company name was known than for tasks where only the transaction amount was known, which is consistent with the user behaviour and performance data described above.

When asked in the interview which version they preferred, a majority of participants (56%) stated they preferred the Exact version. Just 10% said they preferred the Malformed version, with a significant minority saying they had no preference (35%). Ratings on a quality scale (given after listening to examples of transaction listings in each version) further substantiated the other results. The Exact version was rated significantly higher on the quality scale than the Malformed version, with mean ratings (on a scale of 0 to 30) of 23.6 and 13.1 respectively – a sizeable difference, which was found to be highly significant ($p<0.001$). Finally, participants were asked to rate a third possible alternative against the two versions experienced in the experiment, in which the read out of transactions did not include any company information (and therefore no TTS, as in the Case Bank's existing banking service). The mean rating for this version was 8.9 i.e. considerably lower than for either the Exact or Malformed versions.

Therefore, from these results it would suggest a system that provided additional detail in TTS, whether the readout was made from an exact orthography or whether it was malformed would be preferred to a system that did not offer the detail; however, using exact names to produce the TTS output would be preferred to a malformed version.

Certain aspects of the name read outs were limited by the rules set within the TTS engine. For example, as described earlier a malformed name listed as 'Tesco Store 2920' was read out by TTS system as 'tesco store two thousand nine hundred and twenty'. This was due to a default rule in the engine specifying four concurrent digits to be read in a full number format, rather than individual digits. Similarly, shortened terms such as 'ltd' for 'limited' were read as individual letters. This could be rectified or at least lessened to a degree by some tuning of the rules employed by the TTS engine used for the specific application. For example, a rule could be set which details that numbers listed after a proper noun be read out as individual digits (although exceptions to any rule would also have to be considered).

However, tuning the engine would entail a cost in terms of both time and money and therefore it would have to be decided whether the slight improvement made would be worth the extra cost.

8. General discussion

Usability scores from the experiments show some interesting differences across the versions which, together with participant preferences expressed in the interview,

particularly when the differences between versions have been explained, can inform recommendations for the use of TTS in a dialogue system and specifically a telephone banking service.

Results from the first experiment indicate that participants welcome the use of text-to-speech synthesis as a means of providing additional detail in their recent transaction listings. However, participants prefer TTS to be used minimally in providing the additional detail, and therefore prefer a mixed prompt system (recorded speech and TTS synthesis) rather than transaction listings which are entirely TTS. Results from the second experiment indicate that participants would prefer TTS to be used to provide additional detail for all transaction listings rather than a version which is able to provide additional detail using recorded speech (and therefore employs recorded speech only), but only for some of the transactions. This is consistent with the results found in the first experiment where participants preferred the systems which provided extra information even when this was provided using synthesised speech.

Results from the third experiment on the location of the additional detail did not indicate a preference between providing this detail at the end of the listing or in the middle of the listing. However, for tasks where there was more than one transaction of the same amount and the additional detail was therefore the disambiguating information, it was found that there was a request to 'repeat' the transaction information more frequently when the additional detail was in the middle of the transaction, rather than at the end of the transaction. Therefore, this suggests that the additional detail may be more salient if it appears at the end of the listing.

Results from the fourth experiment on company name readouts indicate that participants would prefer the detail to be exact, thus being easier to understand in the service; however, a malformed detail would preferable to no detail at all. It should also be noted that exact readouts would substantially cut down the average call duration in comparison to malformed versions.

The following table (Table 14) summarises the main findings from the four experiments.

Experiment	Experiment 1 Extent of TTS Inclusion	Experiment 2 Minimal Use of TTS	Experiment 3 Location of Additional Detail	Experiment 4 Readout of Company Names
Number of Versions	3	2	2	2
Versions Compared – preferred version in bold	A: All recorded speech as rest of banking service **B: Recorded speech for carrier phrase + TTS additional detail** C: TTS carrier phrase + TTS additional detail	A: All recorded speech, additional detail for Top companies only **B: Recorded speech for carrier phrase, additional detail for all companies - Top companies recorded, all others TTS**	**A: Recorded speech for carrier phrase, additional detail – sentence final** B: Recorded speech for carrier phrase, additional detail – sentence medial	**A: Recorded speech for carrier phrase, additional detail in TTS – exact company name** B: Recorded speech for carrier phrase, additional detail in TTS – malformed company name
Overall findings	The use of TTS for additional information is welcomed, though TTS should only be used where necessary.	The inclusion of TTS to provide additional information where necessary is preferable to a fully recorded system that can give detail occasionally.	No clear preference for location of the additional information was found, though sentence-final may be more salient.	Exact company name information is preferable to malformed names, and would result in shorter call durations.

Table 14: Overall findings of four TTS experiments

Taking into consideration the results of the four experiments, it is suggested that transaction listings in a telephone banking application can be improved by providing additional detail on all transactions, that as far as possible recorded speech should be maintained (for example, for carrier phrases etc) and that where available the additional detail should be provided in recorded speech if recordings have been made. However, for any additional detail where is it unfeasible to make voice recordings, this detail should be provided using TTS synthesis. Although both a sentence-medial and sentence-final location of the additional detail is usable, sentence-final may be more salient. Finally, prior to implementing a system using TTS for such detail as company or place names, care should be taken that the detail employed by the system uses the correct orthography so that the information provided to the user is as clear as possible.

These recommendations are, of course, based on experiments with one particular TTS voice. Ideally, a range of TTS systems and voices would be tested to further validate the results. However, the fact that consistent results were obtained when the TTS voice was used in conjunction with two different pre-recorded voices (the pre-recorded voice in the first experiment being different to that in the other experiments, albeit with the same gender and accent) is encouraging in allowing generalisations to be drawn from the data. Participants preferred the systems which provided extra information even when this was provided using synthesised speech – regardless of whether the synthesised voice matched that of the other prompts in the service.

As to whether the results can be generalised to other application areas, it is suggested that more caution is applied. It is reasonable to assume that the recommendations would apply in other task-oriented dialogue systems aimed at a broad section of users,

such as flight information systems. However, where the domain is less constrained e.g. in companion systems or technical assistants, it is less feasible to pre-record a large proportion of the system prompts and as a result the question of what constitutes 'minimal' TTS alters significantly.

Finally, although the methodological approach described in this paper attempts as much as possible to create a realistic scenario in which the user experiences the service versions, real-world use of a telephone banking service may differ from the experimental condition in the tasks that customers are attempting when they call. In the experiments described in this paper, the task in each call was to check on a set of transactions for known amounts, with or without known payee/payer names. Where the amount was always known exactly before the call, and the additional detail (payee or payer) was either known exactly or entirely unknown, the task becomes one of *spotting known information* in the transaction listing that the service read out. In real life, customers calling a telephone banking service will sometimes have only an approximate idea of the amounts and/or names that they are looking for, and the exact amount or name will sometimes be part of the information that they are seeking to obtain from the service. In this case the task is one of *matching an approximate description and extracting further information* from the spoken output of the service. It would be difficult to replicate the full range of real-world scenarios in an experiment, but it is worth bearing such scenarios in mind when interpreting the experiment results and considering the design of the service. It is likely to remain true in real life, as found in the experiments, that customers welcome the inclusion of additional details such as payee names in the transaction listing, but preferences as to other features of the listing style may vary with the task. For instance, when the amount is not already known to the customer, there may be an advantage in having the payee name in the middle of the listing and the amount at the end of it so that the customer can recognise this transaction as the one they are interested in and then listen carefully for the amount. Although in this case the results of the third experiment showed no clear preference for location of the information and in a real-world application users would come to the system with differing requests (e.g. to seek information on an amount or on a payee), it is worthwhile to note the limitations of the experimental approach, particularly when interpreting the results for real-world applications.

Despite the limitations of the experimental setting, this series of experiments represents a comprehensive usability evaluation of the use of TTS in the context of a real-world application that is not often evident in other research in this area (as suggested by Stevens et al., 2005). The results are in agreement with previous research in a computer aided language learning (CALL) application (Handley, 2009), in which a sample of seventeen language professionals (teachers and researchers) rated the adequacy and acceptability of various TTS systems for use in a variety of teaching roles (e.g. as pronunciation model, conversational partner etc). Handley concludes that the best TTS synthesis systems are ready for use in applications in which they 'add value' to the application, that is they exploit the unique capacity of TTS synthesis to generate speech models on demand. The research reported here supports this idea, but crucially extends the findings into an eCommerce domain using large numbers of end-users. Moreover, through its focus on detailed design issues it aims to increase understanding amongst practitioners of *how* to use TTS in real-life scenarios in a usable way.

This series of experiments on the use of TTS to provide additional information in recent transaction listings for an already established automated telephone banking service indicated that this level of detail would provide added value to the existing system and that the inclusion of TTS is indeed usable and beneficial to customers. Following this research, the Case Bank has successfully incorporated TTS for additional information as part of their recent transactions listings for all customers (using the TTS voice tested in this research) and are currently considering the use of TTS in other parts of the telephone banking functionality.

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 It was sometimes difficult to understand what the service was saying.
Q10 The service was too fast for me.
Q11 I thought the service was polite.
Q12 I found the service frustrating to use.
Q13 I enjoyed using the service.
Q14 I felt flustered when using the service.
Q15 I think the service needs a lot of improvement.
Q16 I thought the service provided enough information.
Q17 I felt the service was easy to use.
Q18 I would prefer to talk to a human being.
Q19 I thought the voice was very clear.
Q20 I felt that the service was reliable.
Q21 I had to concentrate hard to use the service.
Q22 I did not feel in control when using the service.

References

Baker, B.O., Hardyck, C.D. and Petrinovich, L.F. (1966). “Weak measurement vs. strong statistics: an empirical critique of S.S. Stevens’ proscriptions on statistics.” Educational and Psychological Measurement, 26, pp.291-309.

Baylor, A.L., Rye, J. and Shen, E. (2003). “The effects of pedagogical agent voice and animations on learning, motivation and perceived persona.” In *Proceedings of ED-MEDIA*, Hawaii.

Box, G.E.P. (1953). “Non-normality and tests on variances.” Biometrika, 40(3-4), pp.318-335.

Coolican, H. (1994). *Research Methods and Statistics in Psychology.* London: Hodder & Stoughton.

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.

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.

Field, A. (2000). “Discovering statistics using SPSS: advanced techniques for beginners (introducing statistical methods).” SAGE publications Ltd, ISBN 0761957553.

Forbes-Riley, K., Litman, D., Silliman, S. and Tetreault, J. (2006). “Comparing synthesized versus pre-recorded tutor speech in an intelligent tutoring spoken dialogue system.” *Proceedings 19th International Conference of the Florida Artificial Intelligence Research Society (FLAIRS)*, Melbourne Beach, Florida.

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.

Gong, L. and Lai, J. (2003). “To mix or not to mix synthetic speech and human speech? Contrasting impact on judge-rated task performance versus self-rated performance and attitudinal responses.” International Journal of Speech Technology, 6, pp.123-131.

Handley, Z. (2009). “Is text-to-speech synthesis ready for use in computer-assisted language learning?” Speech Communication, 51, pp.906-919.

Hone, K.S. and Graham, R. (2000). “Towards a tool for the subjective assessment of speech system interfaces (SASSI)”. Natural Language Engineering, 6(3-4), pp.287-305.

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.

Kim, J-O. (1975). “Multivariate analysis of ordinal variables.” American Journal of Sociology, 81(2), pp.261-298.

Labovitz, S. (1970). “The assignment of numbers to rank order categories.” American Sociological Review, 35, pp.515-524.

Labovitz, S. (1967). “Some observations on measurement and statistics.” Social Forces, 46, December, pp.151-160.

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. (1997). *The Role of Individual Differences in Dialogue Engineering for Automated Telephone Services*. University of Edinburgh, PhD thesis.

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.” *Proceedings of Institute of Acoustics Speech and Hearing Workshop*, vol.14, no.6, pp.553-559.

Love S., Dutton R.T., Foster J.C., Jack M.A. and Stentiford F.W.M. (1994). “Identifying salient usability attributes for automated telephone services.” *Proceedings of International Conference on Spoken Language Processing*, pp.1307-1310.

Love, S., Foster, J., and Jack, M. (2000). “Health warning: Use of speech synthesis can cause personality changes.” *Proceedings of the IEE Electronics and Communications Seminar on the State of the Art in Speech Synthesis*. London.

McInnes, F. R., Attwater, D. J., Edgington, M.D., Schmidt, M.S. and Jack, M.A. (1999). “User attitudes to concatenated natural speech and text-to-speech synthesis in an automated information service.” *Proceedings of 6th European Conference on Speech Communication and Technology (EUROSPEECH ‘99)*, Budapest, Hungary, pp.831-834.

Möller, S., Krebber, J. and Smeele, P. (2006). “Evaluating the speech output component of a smart-home system.” *Speech Communication, 48* (1), pp.1-27.

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'Brien, R.M. (1979). “The use of Pearson's r with ordinal data.” American Sociological Review, 44, pp.851-857.

Stern, S.E., Mullennix, J.W., Dyson, C.L. and Wilson, S.J. (1999). “The persuasiveness of synthetic speech versus human speech.” Human Factors, 4, pp.588-595.

Stevens, C., Lees, N., Vonwiller, J. and Burnham, D. (2005). “On-line experimental methods to evaluation text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference.” Computer Speech and Language, 19, pp.129-146.

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

Taylor, P. (2009). *Text-to-speech Synthesis*. Cambridge: Cambridge University Press.

Viswanathan, M. and Viswanathan, M. (2005). "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale." Computer Speech and Language, 19, pp.55-83.