Usability of Dialogue Design Strategies for Automated Surname Capture

Usability of Dialogue Design Strategies for Automated Surname Capture

Nancie Davidson*, Fergus McInnes, Mervyn A. Jack
Centre for Communication Interface Research
The University of Edinburgh, U.K.

*Corresponding author

Telephone: +44 (0)131 651 7120

Fax: +44 (0)131 650 2784

E-mail addresses: Nancie.Davidson@ccir.ed.ac.uk , Fergus.McInnes@ccir.ed.ac.uk, Mervyn.Jack@ccir.ed.ac.uk

Abstract
Surname capture via automatic speech recognition over the telephone has many commercial applications, including automated directory assistance and travel reservation services. This paper presents a usability evaluation of three different dialogue designs for automated surname capture, within the context of a flight reservation service. The three designs explored were: a Speak Only strategy, in which callers simply say the surname; a One Stage Speak and Spell strategy in which callers speak and spell the surname in a single utterance; and a Two Stage Speak and Spell strategy in which callers speak and spell the surname in two separate dialogue stages. The methodology employed in the research provides both quantitative user attitude data and performance results for each of the strategies, based on an empirical study with a cohort of 95 participants. The results show a clear distinction between strategies. User attitude towards the dialogues that involve both speaking and spelling the name is high. User attitude towards the Speak Only strategy is significantly less positive. Task completion rates are also significantly higher in the two strategies that involve spelling the name, at around 80% compared to just over 50% in the Speak Only strategy. The data underline the importance of user testing, demonstrating the value of the evaluation methodology used, and provide encouraging results for the strategies that involve both speaking and spelling the name.

Keywords
Usability; Name recognition; Spelling; Dialogue design.

1 Introduction

The problem of proper name recognition has received a great deal of attention from the speech research community in recent years. There are many potential applications, including automated directory assistance (Lehtinen et al., 2000; San-Segundo et al., 2002; Schramm et al., 2000) and the identification of city names for travel services (Lamel et al., 2000).

In principle, proper names can be recognised like any other words if their pronunciation is added to the dictionary of a speech recogniser. In practice, there are two main problems associated with this. The first is the large set of names involved in many applications, ranging from a few thousand names to over a million in some cases. The second is the lack of standardised pronunciations for many names; each can have multiple valid pronunciations, which further increases the difficulty of the recognition task. Given the large number of names involved, automating the process of generating their pronunciations for use in recognition is desirable. Some work has been done on this (Schmidt and Jack, 1994). However, the grapheme-to-phoneme rules involved are extremely complex. It is difficult to construct rules that accommodate fully the high variability in the pronunciation of proper names, and in practice, manual augmentation of the pronunciation dictionary is often required. More recently, a few data-driven grapheme-to-phoneme conversion techniques have been proposed to tackle the problem of automatic pronunciation generation. The decision-tree technique employed by Font Llitjos and Black (2001) for example, produced a word accuracy of 62% on a set of 56 000 names when features based on the language of origin were included in the model. Galescu and Allen (2002) investigated a data-driven joint n-gram method, reporting 68% word accuracy for spelling-to-pronunciation conversion on a similar number of names.

Proper name recognition is therefore an extremely challenging task. Previous reported work has explored a variety of approaches. The simplest in terms of the user interface is to recognise the fluently spoken name without the aid of any other information. Several studies have focused on developing recognition algorithms that achieve acceptable levels of performance using this approach. For example, Béchet et al. (2001) examined a method in which recognition was guided by canonical representation of the name, allowing alternative pronunciations by dynamically generating these in a re-scoring phase. The best result obtained was 69% accuracy on 128 000 names. Sethy and Narayanan (2002) reported a syllable-based recognition system, comparing it to one based on more commonly used context-dependent phones. Their results showed a substantial improvement in name recognition accuracy using the syllable-based recogniser, with a final accuracy of 75% on a word list of 10 000 names. Gao et al. (2001) also investigated various techniques for the improvement of large vocabulary name recognition algorithms, such as weighted speaker clustering, “massive adaptation” of the acoustic models based on data from a pool of calls rather than a single speaker, and various forms of unsupervised utterance adaptation, including Maximum Likelihood Linear Regression (MLLR) and a modified version of Maximum-a-Posteriori Linear Regression (MAPLR). They reported collective gains in accuracy of about 28% relative to their baseline system.

Other methods have also been considered. It has been established that the recognition of spelled names is more accurate than that of spoken names (Kamm et al., 1995;

Meyer and Hild, 1997; Neubert et al., 1998; Seide and Kellner, 1997). Some studies have focused on the use of spelling alone as a means of communicating proper names over the telephone (Hild and Waibel, 1996; Jouvet et al., 1993; Jouvet and Monné, 1999; Mitchell and Setlur, 1999). However, whilst achieving higher accuracy, simply spelling the name without saying it may not seem intuitive to the user. Other work has sought to use spelling in *combination* with the spoken name. Bauer and Junkawitsch (1999), Córdoba et al. (2001) and San-Segundo et al. (2002) investigated the use of spelling as a fallback strategy when problems occur with the fluently spoken name. In Bauer and Junkawitsch (1999) isolated letter recognition with prompting for each letter was initiated for names rejected by the recogniser. The spelling process was then aborted as soon as the name was identified. In Córdoba et al. (2001) and San-Segundo et al. (2002) spelling was invoked only if the top two recognition hypotheses based on the fluently spoken name were rejected by the user, although in this case continuous spelling was used. In all three studies the addition of names recognised correctly at the spelling stage meant a substantial increase in the number of names captured successfully overall.

Other authors have attempted to combine the recognition of spoken and spelled names more explicitly. In Meyer and Hild (1997) and Neubert et al. (1998) a joint recognition approach was investigated in which the name was spoken and spelled in a single utterance. Both calculated the final recognition score of each hypothesis via a weighted combination of the spoken and spelled components, with greater emphasis placed on the spelled part. The result was a recognition accuracy of 90% in Neubert et al. (1998) on a database of around 8 000 names. In Meyer and Hild (1997) the accuracy was 97% on a smaller set of approximately 1 300 names. Both sets of authors report that the spelling was the main source of information, with use of the spoken name producing a slight improvement in the accuracy found using spelling alone.

Meyer and Hild (1997) also investigated joint recognition of the spoken and spelled name when these were two separate recordings. Two separate N-best lists were generated, and only afterwards combined via a weighted addition of matching entry scores. Again, the best result was obtained when the spelling was weighted more heavily than the fluently spoken utterance (98% accuracy). Schramm et al. (2000) explored a similar method, although in this case equal weighting was given to both the spoken and spelled hypotheses. Similar levels of accuracy (92.5% first-best and 97.3% three-best) were obtained on a large inventory of names (approximately 190 000).

Schramm et al. (2000) also examined an alternative method of combining the two separate utterances, in which the spelling of the name was employed as the first step in the dialogue and the subsequent active vocabulary for the spoken part restricted to the candidates identified in the spelling stage. This was found to offer slightly higher accuracy than the previous method (generating two separate N-best lists and only afterwards combining them) but with the added advantage of being computationally more efficient. It follows on from the work of Seide and Kellner (1997) where this approach was used and found to be more accurate than spelling alone. Both of these studies are part of a larger body of work carried out within the context of directory assistance applications where other information relevant to the fluently spoken name is available (its spelling, the city name, street name etc.) and can be used in a

hierarchical combination, reducing the search space with every dialogue turn based on recognition in the previous step (Attwater and Whittaker, 1996; Kaspar et al., 1995). This is a useful approach where such information is available. However, in the case where spelling is the only additional information it may not be intuitive for the user to give this as the first item of dialogue input.

It is from the perspective of the user that the research in this area is weakest. Few studies of the name recognition problem have made any attempt to assess callers' reaction to the various strategies investigated. Much of the work described above involves evaluations of recognition accuracy based on databases of pre-recorded speech (including all of the studies on joint recognition of spoken and spelled names). In some cases the speech was collected in a relevant context (e.g. via recordings of calls made to a live directory assistance service); however more frequently, the recordings were part of a larger corpus of speech collected by asking callers to read aloud a selection of vocabulary items, as in SpeechDat[1]. This is important since previous research has shown that various aspects of speech such as segmental duration and fundamental frequency characteristics are different for read and spontaneous speech (Eskénazi 1993; Laan 1997), and that recognition performance is poorer for spontaneous speech in comparison to read speech (Saraçlar et al., 2000; Weintraub et al. 1996). Data collected within a realistic dialogue context are more valuable and are more likely to produce results that are representative of real-life performance.

Some field trials have been carried out. In San-Segundo et al. (2002) recognition results for the spelling recognizer were considerably poorer in the field evaluation than in the authors' previous laboratory tests. The authors suggest this was the result of operating in difficult conditions since in the field evaluation spelling was only used when the fluently spoken name recognition had failed, indicating the presence of significant background noise, low energy speech signals or callers unused to talking to automatic systems. However, since these are realistic conditions typical of a live environment this simply underlines the importance of evaluating in a field setting.

In other field trials, reported in Lehtinen et al. (2000), participants were recruited to carry out a predefined task using an automated directory assistance system. Here, in addition to recognition accuracy, successful transaction rates and mean task durations were also measured. This is an important step since the effectiveness of an automated dialogue system cannot be judged on the recognition accuracy alone. However, little emphasis was placed on users' reactions to the system.

In Lennig et al. (1995) a customer acceptance survey was used to determine user reaction to a directory assistance service involving increased levels of automation. However, only a small proportion of the research was concerned with automated recognition of the listing name, and no results specific to this are presented.

In Córdoba et al. (2001) volunteers were asked to use an automated directory assistance service to find listings for ten private and ten company entries. The dialogue in this case used spelling as a fallback mechanism. Following this experience each participant then completed a satisfaction questionnaire, the results of which are

[1] For more information on the SpeechDat project visit www.speechdat.org

reported together with recognition accuracy and query success rate. This is one example of an experiment in which user reaction was considered. However, there has been very limited work published which examines the issue of proper name recognition from a user perspective, in particular with respect to the joint recognition of spoken and spelled names. This paper attempts to redress this, in presenting the results of an experiment in which 95 members of the public experienced three different strategies for automated surname capture over the telephone within the context of a flight booking service. In the Speak Only strategy callers simply say the surname. In the One Stage Speak and Spell strategy callers speak and spell the surname in a single utterance. In the Two Stage Speak and Spell strategy callers speak and spell the surname in two separate dialogue stages. In each approach, surname recognition accuracy necessarily plays an important part in the user experience. However, it forms only part of the overall quality judgement. Other factors such as the way in which the system prompts the caller for the required information, and the way in which any recognition errors are handled, also contribute to the interaction. The objective of this study therefore, was to evaluate the impact of the different strategies on the user experience as a whole, an approach that distinguishes this work from the previous research described above. The paper presents quantitative and qualitative data on user attitude towards each of the strategies in addition to objective measures of performance. This provides a measure of the relative effectiveness of the different strategies within a realistic context that is particularly relevant to designers interested in deploying a live application in the near future.

The rest of the paper is organised as follows. Section 2 gives an overview of the dialogue design, with details of the three different strategies examined. Section 3 describes the system implementation, and Section 4 details the experiment. In Sections 5 and 6 the results are presented, with main conclusions given in Section 7.

2 Dialogue design

2.1 Overall structure

Each of the surname capture strategies investigated was set within the context of a flight booking service. Whilst this offered a realistic service, its scope was limited in order to focus on the problem of surname capture in the experiment. A hypothetical scenario was created in which the airline had chosen to give away free flights on a particular route for a particular date. This meant the dialogue consisted only of the capture of passenger name details.

Figure 1 shows a top-level view of the service dialogue.

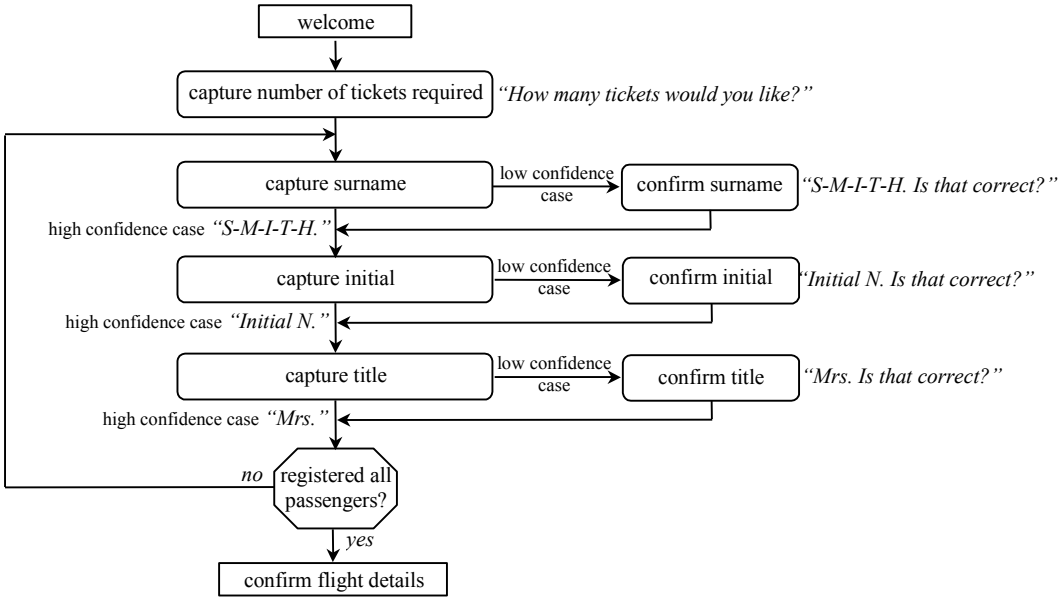

Figure 1. Dialogue call flow

The service was deliberately designed to be fully system-driven in order to provide maximum support for the speech recogniser.

Strict regulations within the airline industry mean that it is vital passenger details are transcribed correctly. As a result, items with a low recognition confidence were played back to the caller for explicit confirmation, as in "*S-M-I-T-H. Is that correct?*" Items that were recognised confidently were simply echoed back to the caller with the confirmation question omitted. The dialogue in this case proceeded immediately to the next request for information, as in "*S-M-I-T-H. And your first initial?"* The use of this approach was intended to speed up the interaction and reduce the monotony of repeated confirmation. Surnames were spelled out to the caller using concatenated recordings of letters, since it is not practical to record all the surnames possible in an application of this type, and previous work has shown strong user preferences for concatenated speech over text-to-speech synthesis (McInnes et al. 1999).

All of the above features were common to all three design variants used in the experiment. Details of how the individual strategies differed within the context of this service are given in the following sections. The three strategies are illustrated below.

(Speak Only) *Please say your surname:* "Smith"

(One Stage Speak and Spell) *Please say then spell your surname: "*Smith, S-M-I-T-H"

(Two Stage Speak and Spell) *Please say your surname:* "Smith"
How do you spell that? "S-M-I-T-H"

2.2 Speak Only Strategy

In this variant of the service callers were simply asked to say the surname and other details as required. The obvious appeal of this strategy is its simplicity, since it is both natural and intuitive for the caller.

However, in a context such as the one used here, where an accurate orthographic transcription of the name is required and other disambiguating information is not available, the question of how to deal with homonymous surnames becomes an issue. The solution employed here was to offer each alternative spelling in succession until the correct surname was read out or the list was exhausted, in which case, depending on the number of errors already made, callers were given the opportunity to say the surname again or the call ended with a recorded message informing the caller that at this point in the real service they would be passed to an agent to complete their reservation.

2.3 One Stage Speak and Spell Strategy

In this variant callers were asked to say and spell the surname in a single utterance, with recognition carried out on the whole.

From a user perspective, the advantage of including spelling information in this way is that it avoids the problems posed by homonymous surnames. A potential drawback is that it may be cognitively more difficult for callers to give both pieces of information at once, and may appear unnatural to those with common, unambiguous surnames (e.g. Jones) who are not normally asked to spell their name.

2.4 Two Stage Speak and Spell Strategy

In this version a joint recognition approach was used in which callers were asked to say their surname and then in a separate stage were asked to spell it, generating separate N-best lists which were only afterwards combined.

In the recognition system employed in the research each item in the N-best list is associated with an acoustic *confidence score*. This provides a measure of the likelihood that the recognition hypothesis matches the actual utterance. Confidence scores computed by the recogniser range from 0 to 100 with the higher the score, the greater the degree of confidence. The N-best list in each case is ranked from highest to lowest confidence.

In order to combine the two N-best lists generated by the separate speak and spell stages therefore, the confidence scores of hypotheses that appeared in both lists were summed, and the list was then reordered according to the new overall confidence.

Items that appeared in only one of the N-best lists were excluded. However, if *no* matches were found, the surname with the most confident spelling was selected as the recognition candidate, since the literature and early testing prior to the experiment indicated that this was the more accurate of the two stages. Testing also indicated that the maximum length of the N-best list should be increased to 30 for the Speak stage (where 10 is the default value used in other stages) in order to increase the likelihood of a match.

As for the One Stage Speak and Spell strategy, the use of spelling resolves the problem of homonymous surnames. Moreover, this may be a more natural and cognitively simpler way for the caller to give the spelling. It does however involve an extra dialogue stage.

3 System implementation

Speech recognition in the experiment was implemented using a commercially available large-vocabulary speaker-independent HMM recogniser capable of recognising both fluent speech and continuous spelling. Due to its commercial nature, full details of the recognition mechanisms employed by the system are not obtainable, however the core approach used is summarised below.

The system employs context-dependent phonemes as its unit of recognition. Some 45 context-dependent phonemes are used to represent the sounds of UK English, together with a separate set of models for digits and letters of the alphabet. The acoustic models are based on the Gaussian mixtures approach, and have been optimised for telephone-quality audio. Decoding is implemented using the Viterbi algorithm and pruning is realised via the beam search method. A proprietary technique known as *phonetic pruning* is also used, which performs additional computation based on the last phoneme analysed at any given time during recognition.

As well as providing speech recognition the system also provides facilities for prompt recording and playback, natural language understanding (NLU) and dialogue management.

The language model employed in the research was a finite-state grammar. In the NLU module of the recognition system used, grammars of this type are hand-coded as an allowable sequence of words and phrases, with NLU implemented by associating an appropriate feature-value pair with each path in the grammar. A database of 11 926 British surnames, all of which had been transcribed or inspected by a trained phonetician, was used to create the system dictionary and grammars. The One Stage Speak and Spell grammar was restricted to matching pairs of fluently spoken names and their corresponding spelling. Users were allowed to link the fluent spoken name with the spelling using either the word “spelt”, as in “Smith spelt S-M-I-T-H”, or “that’s” as in “Smith that’s S-M-I-T-H”. Moreover, in all of the spelling grammars the use of the word “double” was allowed for surnames with two identical letters in sequence e.g. “H-A-double-L”.

Barge-in was disabled for the majority of the dialogue, the main exception being confirmations, where barge-in was allowed during the final part of the prompt (usually the question *“Is that correct?”*).

4 Usability experiment

4.1 Experiment design

In order to measure the relative usability of the three approaches to surname capture, a repeated-measures balanced order experiment design was adopted. Participants were asked to make three telephone calls, one to each version of the service. In each call they were given the same task – to book themselves and a “friend” on the free flight being offered by the airline. Details of the “friend” were supplied by the researcher, from a set of 95 personae created by random selection from the telephone directory. A different “friend” was supplied for each call.

The use of the participant's own surname in the experiment reflects the most likely scenario in real life, and was considered most likely to elicit a natural speaking style, whilst the addition of a "friend" provided performance data on less familiar surnames.

Some of the participant and personae surnames were found to be missing from the original dictionary. The automatic acquisition of unknown names is an ongoing research problem (Chung et al., 2003; Chung and Seneff, 2002). However, the problem of out-of-vocabulary names was outside the scope of this investigation. As a result the missing names were added to the dictionary manually before the relevant experiment session.

Some 45.3% of participants were found to have homonymous surnames based on this dictionary. The sample selected from the telephone directory contained a similar proportion (49.0%).

After each telephone call participants were asked to complete a usability questionnaire to assess their attitude towards the interface. The results were used to compare participants' attitudes towards the three different strategies. A de-briefing interview was also carried out, at the end of the experiment, in order to provide detailed qualitative data on users' responses.

A total of 95 volunteers took part in the research**,** in a group that was balanced for age and gender. Participants received a small honorarium payment. The age groups examined were 18-35 years, 36-49 years and 50 years plus. Participants represented a broad range of socio-economic groups, and all were native speakers of English.

4.2 Key measures

The experiment was designed to provide both subjective and objective data on each of the different strategies. This data may be summarised in terms of three key measures.

4.2.1 Mean attitude score

The first key measure is the *mean attitude score* for each strategy, derived from the usability questionnaire that participants were asked to complete after each telephone call. This questionnaire is a tool for assessing users' attitudes towards automated telephone services that has been developed and refined by the authors and their colleagues over a number of such experiments (Dutton et al., 1993). It consists of a set of proposal statements which are short and simple, each with a set of tick-boxes along a seven-point Likert scale (Likert, 1932) ranging from "strongly agree" through neutral to "strongly disagree". The wording of the statements in the questionnaire is balanced, positive and negative, to counteract the problem of response acquiescence set - the general tendency for respondents to agree with an offered statement. 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. 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 (this is the *mean attitude score*).

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.

4.2.2 Explicit preference

The second key measure is participants' *explicit preference* between the three variants of the dialogue. This was obtained as part of the de-briefing interview, where participants were first asked which version of the service they *preferred*, followed by which version they liked *least*.

4.2.3 Task completion rate

The third key measure is the *task completion rate*. This is the proportion of participants in each strategy who succeeded in booking two passengers onto the flight. Surname recognition accuracy plays an important part in this, however task completion also encompasses other factors such as the system's ability to elicit valid responses from the user, and to handle successfully any errors that occur. As such, it is an important objective measure of the effectiveness of the dialogue as a whole.

5 Results

Table 1 summarises the results for each strategy on each of the three defined measures.

Strategy	Mean attitude score	Explicit preference		Task completion
		Most preferred	Least preferred	
Speak Only	4.57	13.7%	63.2%	51.6%
One Stage Speak and Spell	5.18	46.3%	10.5%	80.0%
Two Stage Speak and Spell	5.17	37.9%	17.9%	77.9%

Table 1. Key results for each strategy

In each case the Speak Only strategy performed or was rated the poorest. Details are provided in the following sections.

5.1 Mean attitude score

All three strategies were rated better than neutral. However, the mean attitude score for the two strategies that involved both speaking and spelling the name was considerably more positive than that of the Speak Only version.

To establish the significance of these results, a repeated-measures analysis of variance (ANOVA) was carried out using the mean attitude scores for each strategy. The within-subject factor was *strategy*, with *age group*, *gender* and *order of presentation* of the three versions as the between-subject factors. The result demonstrated a very highly significant effect of strategy on attitude ($p<0.001$).

Post-hoc pair-wise comparisons showed there was no significant difference in the mean attitude score when comparing the two strategies which involved both speaking and spelling the name to each other. There was one significant difference between the two when examining *individual* issues: participants were significantly more positive

towards the Two Stage Speak and Spell version with regard to the level of concentration required (Two Stage Speak and Spell mean 4.67, One Stage Speak and Spell mean 4.38, p=0.021). On the whole however, participants rated the two spelling strategies very similarly. Both were rated positively throughout, with only one exception: *preference for a human operator*. All three strategies actually scored below neutral on this point, indicating that participants would prefer to talk to a human regardless of the strategy employed by the automated service - a result often encountered in previous research with other telephone-based services.

In contrast to the two spelling strategies however, the Speak Only version was also rated below neutral on several other issues. For example participants did not *enjoy* using this version of the service, they found it *frustrating*, and felt that it required a lot of *improvement.*

Moreover, even when scoring above neutral the Speak Only version was judged to be consistently worse than either the One Stage Speak and Spell or the Two Stage Speak and Spell version. The differences in attitude were found to be significant for a large number of issues, resulting in a very highly significant difference in the mean attitude score in both cases ($p<0.001$). In total, the Speak Only version was rated significantly lower than the One Stage Speak and Spell version on fifteen of the twenty core usability issues, and significantly lower than the Two Stage Speak and Spell version on sixteen of these issues.

There were several usability attributes for which the effects were particularly strong. Participants felt significantly more *frustrated, stressed* and *flustered* when using the Speak Only version in comparison to either of the other two versions. They also found it less *reliable,* less *efficient* and more in *need of improvement*. They *enjoyed* using it less and were significantly less *happy to use it again.* All of these differences were highly significant ($p<0.001$).

Attitudes towards the three strategies converged on only three issues. Participants did not find any of the versions *too fast* or *too complicated*, and all three were considered *friendly*.

The overall pattern to emerge therefore was that user reaction to both spelling strategies was positive, and both were rated significantly higher than the Speak Only version.

5.2 Explicit preference

Figures for explicit preferences are given in Table 1. The Speak Only version was the least preferred option for the majority of participants (63.2%). A chi-square test confirmed that this distribution of responses was unlikely to occur by chance ($p<0.001$).

Based on the responses to the question of most and least preferred version, an absolute ranking was calculated for each of the three versions, for each participant. Pair-wise comparisons on these rankings were then carried out using the Binomial test. The Speak Only result was very highly significant when compared to each of the other two versions ($p<0.001$). However, there was no significant difference between the two strategies which involved both speaking and spelling the name.

When asked for their reasons for their choice, most of the group who selected the Speak Only version as their least preferred option (81.7%) said this was the result of *trouble being understood*.

More than half of the participants who chose the One Stage Speak and Spell version as their most preferred strategy mentioned spelling in their reasons. Some 28.9% of those who selected this version said that being allowed to spell the passenger details influenced their decision. A further 28.9% were more specific, citing being able to speak and spell the details *at the same time*. Better recognition performance was also given as a reason, by 33.3% of this group.

Of those who preferred the Two Stage Speak and Spell version, 36.1% said better recognition performance was their reason. The other main reasons mentioned were that it was easier (25%), quicker (13.9%) and did not ask the caller to say and spell information at the same time (16.7%).

When questioned as to what they thought of the ways in which they were asked to give surnames, a total of 50.5% of participants mentioned spelling as a positive feature; 15.8% of this group specified that they liked being asked to say and spell the surnames at the same time, 8.4% expressed a preference for the two stage process and 26.3% were non-specific. Those in the non-specific group generally liked spelling because it improved recognition performance and/or they were in the habit of spelling their name over the telephone. Those who expressed a preference for giving the surname and its spelling at the same time did so generally because they perceived this to be quicker. The group who preferred the two stage process felt it was more natural.

5.3 Task completion rate

Observing the figures in Table 1 it is clear that task completion was much higher in the two spelling strategies than it was in the Speak Only version. More than three quarters of all participants succeeded in achieving their goal using the two spelling strategies, compared to only just over half in the Speak Only version.

The pattern of results was very similar to that observed in the attitude and interview data. The effect of strategy on task completion was very highly significant (Cochran's Q $p<0.001$). Pair-wise comparisons then showed that the differences between the Speak Only version and each of the other two versions were very highly significant (McNemar $p<0.001$), whilst there was no significant difference between the two spelling strategies.

5.3.1 Reasons for task failure

There were two main reasons for task failure in this application: the registration of incorrect passenger details or breakout to an agent as a result of dialogue failure.

The former could occur either as the result of a confident mis-recognition on the part of the system, or as a result of participants explicitly confirming incorrect information.

The latter could also occur for one of two reasons. Firstly, as a result of three successive failures to recognise a valid response from the user, either because they

were silent or gave an out-of-grammar response, or because the recogniser was unable to produce a recognition hypothesis. Secondly, breakout could occur as a result of repeated failure on the part of the system to recognise valid information correctly. Callers were asked to give each piece of information up to a total of five times. If after five attempts the system failed to recognise it correctly, breakout was initiated.

Table 2 summarises the incidence of each type of task failure in the experiment.

Strategy	Incorrect details registered (% participants)	Breakout (% participants)
Speak Only	9.5%	38.9%
One Stage Speak and Spell	9.5%	10.5%
Two Stage Speak and Spell	8.4%	13.7%

Table 2. Summary of task failures.

The number of failures due to the registration of incorrect details was very similar in each of the strategies. Strategy had no effect on the number of participants who failed as a result of this problem (Cochran's Q).

Breakout was the most common cause of task failure in all three strategies. However, the level of breakout was significantly higher in the Speak Only version than in either of the other two strategies (McNemar $p<0.001$). This was largely as a result of the number of breakouts at the surname stage. Some 29.5% of participants broke out during this stage in the Speak Only version, compared to 4.2% in the One Stage Speak and Spell version and 5.3% in the Two Stage Speak and Spell version. This contrasts with other stages in the dialogue where all three strategies produced a similar level of breakout.

The breakout figures for the surname entry stage broken down by participants' own surname and that of the second passenger are shown in Table 3.

Strategy	Own surname (% participants)	Other surname (% participants)
Speak Only	11.6%	17.9%
One Stage Speak and Spell	1.1%	3.2%
Two Stage Speak and Spell	4.2%	1.1%

Table 3. Summary of breakouts at surname stage

Strategy had a significant effect on the breakout rate when participants were giving their own surname (Cochran's Q $p=0.005$). Pair-wise comparisons showed that the breakout rate in the Speak Only version was significantly higher than in the One Stage Speak and Spell version (McNemar $p=0.006$). This was the only pair-wise comparison that produced significant results.

Results for participants' own surnames can be tested in this way since the same set of 95 participants attempted this stage in all three strategies. However, not all participants attempted the other passenger surname in all three versions of the service (as a result of breakouts earlier in the dialogue). Different sets of participants attempted this stage across the different strategies and as a result statistical comparisons between them are invalid.

5.4 Surname recognition accuracy

Surname recognition accuracy played an important part in the level of breakout (and therefore task completion) observed in the experiment. Analysis showed that 92.9% of the breakouts at the surname stage in the Speak Only version were the result of five failed attempts to recognise the information correctly.

Table 4 shows the average recognition accuracy experienced by users, when giving their own surname and that of the other passenger, for all participants who attempted these stages and gave an in-grammar response.

Strategy	Own surname	Other surname
Speak Only	63.9%	55.4%
One Stage Speak and Spell	96.0%	92.2%
Two Stage Speak and Spell	91.6%	89.3%

Table 4. In-grammar surname recognition accuracy

The results are comparable to other work in the field. The One Stage Speak and Spell strategy achieved an accuracy of over 90% for both surnames. Performance in the Two Stage Speak and Spell strategy was only slightly lower, falling to just under 90% for the other surname. Both performed considerably better than the Speak Only strategy, where the average recognition accuracy was as low as 55.4% for the other passenger's surname.

5.4.1 Own surname

A repeated-measures ANOVA was carried out on data from the 83 participants who provided an in-grammar response at this stage in all three versions of the service. *Strategy* was the within-subject factor, with *age group*, *gender* and *order of presentation* included as between-subject factors.

The result was a very highly significant effect of strategy on surname recognition accuracy ($p<0.001$). The Speak Only strategy performed significantly worse than either of the other two strategies ($p<0.001$) although there was no significant difference between the two spelling strategies. Both of the spelling strategies performed well, achieving accuracies of over 90%.

5.4.2 Other surname

To allow some comparisons between the different strategies to be made, data for this stage were restricted to the 44 participants who completed all three calls, and gave an in-grammar response in each.

Based on this group the average recognition accuracy experienced by participants was 67.2% in the Speak Only version, 94.9% in the One Stage Speak and Spell version and 86.7% in the Two Stage Speak and Spell version.

A repeated-measures ANOVA was carried out on the results, with *strategy* as the within-subjects factor and *age* and *gender* as the between-subjects factors. *Order of presentation* was omitted as a factor in this case since it was found to have no effect, and the reduced sample size meant its inclusion created empty cells in the analysis.

Based on this, strategy was found to have a highly significant effect on the recognition accuracy for the other passenger surname ($p=0.008$). The performance in the Speak Only version was significantly poorer than that in the One Stage Speak and Spell strategy ($p=0.001$), although this was the only significant difference in the pair-wise comparisons.

Closer inspection revealed that gender had a significant effect on these results ($p=0.030$). On average, women experienced poorer recognition accuracy than men when giving the second passenger's surname. Table 5 shows the results for each strategy broken down by gender.

Strategy	Male (N=19)	Female (N=25)
Speak Only	78.9%	58.3%
One Stage Speak and Spell	97.4%	93.0%
Two Stage Speak and Spell	89.5%	84.7%

Table 5. Surname recognition accuracy by gender (other surname)

Analysing the two groups separately it was found that strategy had no significant effect on the recognition accuracy experienced by men. The results for women, on the other hand, followed the pattern found in previous analyses i.e. strategy had a highly significant effect ($p=0.004$) and the Speak Only version performed significantly worse than either of the other two strategies ($p<0.05$). There was no significant difference in the performance of the two spelling strategies.

5.4.3 Bias in the results

Removing participants who broke out of the dialogue from the data set means that those with the greatest recognition difficulties were excluded from the analysis. As a result, the figures given for the other passenger's surname will tend to exhibit a positive bias.

An estimate of the degree of bias introduced can be obtained by calculating the accuracy for participants' *own* surname based on the sub-group who did not break out and comparing it with the figure already calculated for the whole sample.

Table 6 shows both sets of figures. Of the 47 participants who completed calls to all three strategies, 40 provided an in-grammar response in each when asked for their own surname.

Strategy	Own surname (N=83)	Own surname (N=40)
Speak Only	62.9%	74.3%
One Stage Speak and Spell	96.8%	98.5%
Two Stage Speak and Spell	91.6%	92.9%

Table 6. Effects of removing participants who did not complete three calls

The bias introduced was greatest in the Speak Only version, which is to be expected since this was the strategy with the highest level of breakout.

A repeated-measures ANOVA on the reduced data set again showed that strategy had a significant effect on the recognition accuracy for participants' own surnames ($p<0.001$). Recognition accuracy in the Speak Only version was significantly poorer than in either of the two versions ($p<0.005$), although there was no significance in the

difference between the two spelling strategies. Thus, even excluding those participants who broke out as a result of recognition difficulties, the Speak Only performance was significantly worse.

5.4.4 Own vs other passenger surname

In all three strategies recognition performance on participants' own surname was slightly better than on the other passenger's surname, suggesting that familiarity with the name had a positive effect. However, comparing only those participants who made an in-grammar attempt at both surnames, the effect was not found to be significant in any of the strategies.

5.5 Other results

5.5.1 Out-of-grammar responses

One area in which the Speak Only strategy was not the poorest performer was in its ability to elicit in-grammar responses at the surname stage. Table 7 shows the proportion of input attempts that were in-grammar at this stage, for each strategy.

Strategy	Own surname	Other surname
Speak Only	92.8%	91.8%
One Stage Speak and Spell	80.0%	82.7%
Two Stage Speak and Spell: Say Surname	90.2%	87.7%
Spell Surname	94.1%	91.7%

Table 7. Surname in-grammar rates by strategy

The level of out-of grammar responses was highest in the One Stage Speak and Spell strategy. Table 8 shows a breakdown of the various types of out-of-grammar responses provided during surname capture, for each of the different strategies.

	Speak Only	One Stage Speak and Spell	Two Stage Speak and Spell Say Surname	Two Stage Speak and Spell Spell Surname
Additional speech	5	13	7	0
No spelling	n/a	15	n/a	0
Added spelling	6	n/a	4	n/a
Spelling only	5	0	10	n/a
Filled pause/false start	2	5	0	0
End pointing	1	5	0	4
Speech too early	0	2	0	8
Other	9	6	3	5
Total input attempts	421	251	242	254

Table 8. Breakdown of out-of-grammar surname utterances by strategy

The analysis showed that omission of the spelling was not the principal reason for the higher level of out-of-grammar responses in the One Stage Speak and Spell strategy, as might have been expected. In fact, there were a similar number of cases of participants including the spelling in stages where it was *not* requested (either providing it together with the fluently spoken name or providing it in place of the spoken name). Instead, the main reason for the higher level of out-of-grammar responses was the inclusion of additional speech. All three versions produced replies where the correct response was embedded in extraneous speech (due to the already vast size of the surname grammars, this was not allowed). However, the incidence of

this was higher in the One Stage Speak and Spell strategy. The nature of the additional speech was roughly divided between preamble such as "My surname is…" and the inclusion of the title and/or the first name together with the surname, as in for example "Simon Moffat M O F F A T T". There was also a higher incidence of filled pauses and false starts in the One Stage Speak and Spell strategy, possibly as a result of the greater complexity of the input task.

In stages where the spelling of the surname was requested there was some occurrence of end-pointing errors, which meant that participants were interrupted mid-spelling. Increasing the length of the end-of-speech timeout for these stages may help to alleviate this problem. The value used in the experiment was 1.5 seconds.

The way in which out-of-grammar utterances were handled by each strategy was then examined. It was found that in terms of rejecting out-of-grammar utterances the One Stage Speak and Spell strategy was the most effective. Of the 34 out-of-grammar responses to the top level prompt in this strategy for example, 24 were rejected by the recogniser, which meant that error recovery was initiated. The corresponding figures for the Speak Only strategy and the individual speak and spell stages of the Two Stage Speak and Spell strategy were 9 out of 23, 3 out of 20 and 4 out of 16 respectively. A higher proportion of out-of-grammar responses were falsely accepted as valid input in these strategies, resulting in an incorrect recognition hypothesis in the majority of cases. Careful design of the error recovery prompts in the One Stage Speak and Spell strategy also meant that the majority of the out-of-grammar responses that resulted in error recovery were subsequently converted to an in-grammar response at either the second or third level (18 out of 24). Thus, although the number of out-of-grammar responses was higher in the One Stage Speak and Spell strategy, in most cases these were successfully detected and recovered from.

5.5.2 Call length

The inclusion of an extra dialogue stage in the Two Stage Speak and Spell strategy did not have a significant effect on call duration. The average call length, based on the 47 participants who completed all three calls, was 142 seconds in the Speak Only version, 130 seconds in the One Stage Speak and Spell version, and 140 seconds in the Two Stage Speak and Spell version. The One Stage Speak and Spell version showed a tendency to be fastest, however none of the differences were found to be statistically significant.

6 Relationship between results

6.1 Relating mean attitude score to explicit preference

The ability of the usability questionnaire to predict participants' explicit preferences was assessed by comparing the predicted preference (based on the difference between the questionnaire scores for the three versions of the service) with the expressed preference for each participant. The prediction accuracy was scored as 1 if the predicted and expressed preference agreed, or 0 if they disagreed outright. Cases where either the predicted or expressed preference was neutral were excluded from this part of the analysis (e.g. if the mean attitude score was the same for two or more versions).

For the question of participants' least preferred strategy the overall prediction accuracy was 80.7%, which is considerably better than the 33% accuracy that would be expected for random or uniformly neutral prediction. Correlation analysis confirmed that the departure from chance was very highly significant (Cramer's V 0.596, $p<0.001$).

Similarly, the overall prediction accuracy for the most preferred strategy was 71.2%. Again, the correlation was very highly significant (Cramer's V 0.504, $p<0.001$).

The usability questionnaire was therefore a fairly reliable indicator of participants' preference between versions.

6.2 Relating mean attitude score to task completion

Table 9 shows the mean attitude score for each strategy broken down by task completion.

Strategy	Mean attitude score (Task failure)	Mean attitude score (Task success)
Speak Only	4.04	5.08
One Stage Speak and Spell	4.46	5.36
Two Stage Speak and Spell	4.75	5.29

Table 9. Mean attitude score by task completion

Unsurprisingly perhaps, it appears that participants who were successful in completing the task had a more positive attitude towards the interface, in all three strategies. Unrelated-samples t-tests confirmed the effect was very highly significant ($p<0.001$) for the Speak Only and One Stage Speak and Spell versions and significant ($p=0.021$) for the Two Stage Speak and Spell version.

7 Conclusions

In this paper the results of a usability experiment that examined three different dialogue strategies for automatic surname capture in a flight reservations context have been presented. The three strategies were Speak Only, One Stage Speak and Spell and Two Stage Speak and Spell.

The objective of the study was to examine the impact of the different strategies on the user experience as a whole, measuring user attitudes and task completion rates as well as recognition accuracy. From the results it is concluded that the Speak Only strategy was the least effective, in terms of all of the key measures of performance. Both strategies involving spelling performed significantly better, although there was no substantial difference between the two.

Participants had a positive attitude towards both spelling strategies (scoring them very similarly at 5.17 and 5.18 on a 7-point scale) but were significantly less positive towards the Speak Only strategy (4.57).

The Speak Only strategy was also the least preferred option for the majority of participants (63.2%) and was ranked top by fewest participants (13.7%). Opinion was more divided on the two spelling strategies - both were ranked top by a roughly equal

number of participants. In either case, qualitative data showed that participants were happy to use spelling as part of the surname capture process.

Objective measures of performance also yielded positive results for the two strategies that involved spelling, with high task completion in each (80.0% in the One Stage Speak and Spell strategy and 77.9% in the Two Stage Speak and Spell strategy). The results for the Speak Only strategy were significantly poorer, with only just over half of all participants succeeding in their goal using this strategy (51.6%).

The combination of poor results for the Speak Only strategy suggests that this design approach is not ready for commercial deployment in an application of this type. A substantial improvement in both user attitude and objective performance is required before this strategy should be considered for use in a live service.

The results for the two strategies that involve spelling, on the other hand, are encouraging. The research has demonstrated that, through the use of spelling information, high levels of user satisfaction and task completion are achievable using a commercially available speech recognition system.

Acknowledgements
This work was carried out as part of EC IST Project Spotlight. The authors also wish to acknowledge Nuance Communications for use of their recognition software, and bmi British Midland for their contribution in providing the context for the research.

References

Attwater, D. J., Whittaker, S.J., 1996. Issues in large-vocabulary interactive speech systems. BT Technology Journal 14(1), pp.177-186.

Bauer, J., Junkawitsch, J., 1999. Accurate recognition of city names with spelling as a fall back strategy. In Proc. EUROSPEECH, pp.263-266.

Béchet, F., de Mori, R., Subsol, G., 2001. Very large vocabulary proper name recognition for directory assistance. In Proc. IEEE ASRU Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio, Italy.

Chung, G., Seneff, S., Wang, C., 2003. Automatic acquisition of names using speak and spell mode in spoken dialogue systems. In Proc. HLT-NAACL, pp.32-39.

Chung, G., Seneff, S., 2002. Integrating speech with keypad input for automatic entry of spelling and pronunciation of new words. In Proc. ICSLP, pp.2053-2056.

Córdoba, R., San-Segundo, R., Montero, J.M., Colás, J., Ferreiros, J., Macías-Guarasa J., Pardo, J.M., 2001. An interactive directory assistance service for Spanish with large-vocabulary recognition. In Proc. EUROSPEECH, pp.1279-1282.

Dutton, R.T. Foster, J.C., Jack, M.A. and Stentiford, F.W., 1993. Identifying usability attributes of automated telephone services. In Proc. EUROSPEECH pp.1335-1338.

Eskénazi, M., 1993. Trends in Speaking Styles Research. In Proc. EUROSPEECH, pp.501-509.

Font Llitjos, A., Black, A.W., 2001. Knowledge of language origin improves pronunciation accuracy of proper names. In Proc. EUROSPEECH, pp.1919-1922.

Galescu, L., Allen, J.F., 2002. Pronunciation of proper names with a joint n-gram model for bi-directional grapheme-to-phoneme conversion. In Proc. ICSLP, pp.109-112.

Gao, Y-Q., Ramabhadram, B., Chen, J., Erdogan H. and Picheny, M., 2001. Innovative approaches for large vocabulary name recognition. In Proc. ICASSP, pp.333-336.

Hild, H., Waibel, A., 1996. Recognition of spelled names over the telephone. In Proc. ICSLP, Volume 1, pp.346-349.

Jouvet, D., Lokbani, M.N., Monné, J., 1993. Application of the N-best solutions algorithm to speaker-independent spelling recognition over the telephone. In Proc. EUROSPEECH, pp.2081-2084.

Jouvet, D., Monné, J., 1999. Recognition of spelled names over the telephone and rejection of data out of the spelling lexicon. In Proc. EUROSPEECH, pp.283-286.

Kamm, C.A., Shamieh, C.R., Singhal, S., 1995. Speech recognition issues for directory assistance applications. Speech Communication 17, pp.303-311.

Kaspar, B., Fries, G., Schuhmacher, K., Wirth, A., 1995. FAUST – A directory assistance demonstrator. In Proc. EUROSPEECH, pp.1161-1164.

Laan, G.P.M., 1997. The contribution of intonation segmental durations and spectral features to the perception of a spontaneous and a read speaking style. Speech Communication 22, pp.43-65.

Lamel, L., Rosset, S., Gauvain, J.L., Bennacef, S., Garnier-Rizet, M., Prouts, B., 2000. The LIMSI ARISE system. Speech Communication 31, pp.339-353.

Lehtinen, G., Safra, S., Gauger, M., Cochard, J.L., Kaspar, B., Hennecke, H., Pardo, J.M., Córdoba, R., San-Segundo, R., Tsopanoglou, A., Louloudis, D., Mantakas, M., 2000. IDAS: Interactive directory assistance service. In Proc. VOTS-2000 Workshop, Belgium.

Lennig, M., Bielby, G., Massicote, J., 1995. Directory assistance automation in Bell Canada: Trial results. Speech Communication 31, pp.227-234.

Likert, R., 1932. A technique for the measurement of attitudes. Archives of Psychology, Vol.140.

McInnes, F.R., Attwater, D.J., Edgington, M.D., Schmidt, M.S., Jack, M.A., 1999. User attitudes to concatenated natural speech and text-to-speech synthesis in an automated information service. In Proc. EUROSPEECH, pp.831-834.

Meyer, M., Hild, H., 1997. Recognition of spoken and spelled proper names. In Proc. EUROSPEECH, Vol 3, pp.1579-1582.

Mitchell, C.D., Setlur, A.R., 1999. Improved spelling recognition using a tree-based fast lexical match. In Proc. ICASSP, vol. 2, pp.597-600.

Neubert, F., Gravier, G., Yvon, F., Chollet, G., 1998. Directory name retrieval over the telephone in the Picasso project. In Proc. of the IEEE IVTTA Workshop, Torino, Italy.

San-Segundo, R., Colás, J., de Córdoba, R., Pardo, J., 2002. Spanish recognizer of continuously spelled names over the telephone. Speech Communication 38, pp.287-303.

Saraçlar, M., Nock, H., Khudanpur, S., 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer Speech and Language, 14(2), pp.137-160.

Schmidt, M.S., Jack, M.A., 1994. A multi-language pronunciation dictionary for proper names and place names: ONOMASTICA. In Proc. Language Engineering Convention, pp.125-128.

Schramm, H., Rueber, B., Kellner, A., 2000. Strategies for name recognition in automatic directory assistance systems. Speech Communication 31, pp.329-338.

Seide, F., Kellner, A., 1997. Towards an automated directory information system. In Proc. EUROSPEECH, pp.1327-1330.

Sethy, A., Narayanan, S., 2002. A syllable based approach for improved recognition of spoken names. In Proc. ISCA Pronunciation Modeling and Lexicon Adaptation Workshop, Denver, USA.

Weintraub, M., Taussig, K., Hunicke-Smith, K, Snodgrass, A., 1996. Effect of speaking style on LVCSR performance. In Proc. ICSLP, Addendum, pp.16-19.