Journal of Library and Information Studies 14:1 (June 2016) pp.25-39
doi:10.6182/jlis.2016.14(1).025

Users' Perceived Difficulties and Corresponding Reformulation Strategies in Google Voice Search

Wei Jeng[1], Jiepu Jiang[2,†], Daqing He[3]

Abstract

In this article, we report users' perceptions of query input errors and query reformulation strategies in voice search using data collected through a laboratory user study. Our results reveal that: 1) users' perceived obstacles during a voice search can be related to speech recognition errors and topic complexity; 2) users naturally develop different strategies to deal with various types of words (e.g., acronyms, single-worded queries, non-English words) with high error rates in speech recognition; and 3) users can have various emotional reactions when encounter voice input errors and they develop preferred usage occasions for voice search.

Keywords: Voice Search; Voice Input Errors; Query Reformulation; Google Voice

1. Introduction

With the rapid development of mobile devices, voice search has become an attractive input interface for constrained devices, such as mobile handsets. Unlike conventional search systems that require a keyboard for inputting queries, voice search systems have to engage users in much more complex interactions. Recently, there are few contemporary studies specifically focusing on user interactions in voice search (Schalkwyk et al., 2010; Shokouhi, Jones, Ozertem, Raghunathan, & Diaz, 2014). While these previous studies help deepen understanding of users' voice queries and reformulations, however, users' barriers and reactions when conducting a voice search remain unexplored.

To bridge the gap, our previous study focuses on typical query input errors and users' query reformulation behaviors (Jiang, Jeng, & He, 2013). By analyzing search logs generated by users, we found that voice input errors were prevalent in state-of-the-art voice search systems and resulted in the substantial decline of search performance. Users adopted both lexical query reformulations (e.g., query term addition, substitution, removal, and re-ordering) and phonetic query reformulations (e.g., emphasize a part of or the entire query), some of which are closely related to the previously misrecognized words (e.g., query term substitution and hyper-articulation a part of the query).

In this paper, we intend to augment our previous study by examining the data collected

[1,3] School of Information Sciences, University of Pittsburgh, USA
[2] Center for Intelligent Information Retrieval, School of Computer Science, University of Massachusetts, USA
† This work was done while Jiepu Jiang was at the School of Information Sciences, University of Pittsburgh.
* Corresponding Author: Wei Jeng, E-mail: wej9@pitt.edu

from a study and interview. Specifically, we are interested in the following new research questions:

- What are the users' perceived challenges while using voice search?
- What are the users' perceptions of the query reformulation strategies for resolving voice input errors?
- What are the users' emotional reactions when encountering voice input errors?

Our results provide better insight on voice input errors and users' interactions in current voice search systems, which will further help design a more effective and user-friendly voice search interface.

2. Related Work

2.1 Voice search and related applications

Voice search is a relatively new research topic. Among the few but existing studies, Crestani and Du (2006) conducted a user experiment comparing voice queries with written queries, but the experiment settings did not involve user interaction; Schalkwyk et al. (2010) report statistics of individual queries from Google Voice's search logs. Our prior work (Jiang et al., 2013) examines user behaviors in voice search by recruiting 20 participants, but mainly focuses on the problem of using log analysis. The results confirm that voice input errors greatly affect performance for individual search, and the probabilities of using different query reformulation strategies. Shokouhi et al. (2014) compare the use of text-to-text, voice-to-voice, text-to-voice, and voice-to-text reformulations in Bing's mobile search logs. They found that voice-to-text reformulation usually indicates the occurrence of speech recognition errors in the voice query. Narang and Bedathur (2013) developed an experiment with 13 participants, each of whom was asked to finish 20 TREC topics. They found that individuals who had higher English proficiency performed better in voice search. Jiang et al. (2015) included acoustic features of voice query reformulation in an online evaluation approach for intelligent personal assistants, such as the metaphone similarity of queries and changes in the user's speaking rate.

Another group of related studies focuses on users' responses in spoken dialog systems. For example, Swerts, Litman, and Hirschberg (2000) categorize users' responses to the recognition errors of dialog systems, including repeating, paraphrasing, adding relevant content, omission and hyperarticulation, similar to the lexical and phonetic reformulation patterns we observed in voice search. Comparable findings are reported in (Bohus & Rudnicky, 2005; Raux, Langner, Bohus, Black, & Eskenazi, 2005; Shin, Narayanan, Gerber, Kazemzadeh, & Byrd, 2002). However, spoken dialog systems significantly differ from voice search systems. The former is usually designed to handle structural query input (e.g., location and time) and solve a specific task (e.g., flight information inquiry), while the latter deals with far more diverse information needs and flexible query inputs. Overall, there are very limited studies on user interactions and query reformulation strategies in a voice search. Understanding these issues can foster the design of voice search systems.

2.2 Query reformulation

Query reformulation, in the scope of this paper, refers to the users' self-motivated behavior of

formulating a new query successive to an existing query. The relation between the two queries is the focus of many previous works, including our study.

Previous work has characterized the patterns of query reformulation in conventional search systems without voice query input (Anick, 2003; Bruza & Dennis, 1997; Huang & Efthimiadis, 2009; Jansen, Booth, & Spink, 2009; Rieh & Xie, 2006). The patterns can be characterized from a lexical aspect, such as query term addition (adding words to a query), query term deletion (removing words from a query), query term substitution (replacing a word into another with similar meaning), spelling correction, stemming, case change, and using acronyms. On the other hand, the patterns can indicate syntactic differences, such as punctuation (e.g., adding or removing a whitespace), reordering of words, and using search operators. Finally, some patterns are related to users' intentions and search tactics. These include specification (using more specific terms), generalization (using more general terms), and topic or subtopic change. Not all of these patterns are mutually exclusive of each other. For example, specification can happen by adding new query terms or by replacing a general term with a specific one.

In contrast to these studies, due to the nature of query reformulation in voice search that we previously discovered (Jiang et al., 2013), this paper focuses not only on textual changes in query reformulation, but also the variation of acoustic characteristics. Additionally, repeating a query (without any change) is also considered an important voice query reformulation strategy to deal with voice input errors.

3. Methodology

3.1 Experiment design and piloting

This study adopts a laboratory experiment design with a follow-up survey and semi-structured interview. We used the Google Voice Search on a tablet for our experiment because Google search history provides an easy method for tracking users' search and browsing history. Google Voice Search is a Google product that allows users to use its search engine by inputting speech queries.

We used three sets of text collection for collecting participants' voice queries: the TREC Robust Track 2004 (trec.nist.gov/data/robust/04.guidelines.html) and the TREC Web Track 2010 (trec.nist.gov/data/web10.html) and 2011 (trec.nist.gov/data/web2011.html). We selected Topic No. 668 from the TREC Robust track collection for training, and the topics selected for the experiment are listed in Table 1.

Table 1. Selected TREC Topics for Experiments

Datasets	Selected topics
Robust Track 2004	301, 302, 303, 307, 309, 311, 313, 314, 316, 318, 321, 322, 338, 348, 351, 356, 365, 380, 404, 406, 608, 628, 630, 637, 647, 651, 654, 668 (for training sessions only), 672, 683, 698
Web Track 2010, 2011	51, 52, 54, 56, 68, 70, 72, 73, 74, 91, 94, 100, 104, 107, 108, 110, 112, 113, 122, 141

We created an interactive PowerPoint mockup and Chrome browser environment to mimic a user's typical browsing behavior. This method has been commonly used in Human Computer Interaction experimental settings (e.g., Kim et al., 2012). Our mockup interface had three components, including an experimental instruction, a training session, and sets of TREC topics. An example is shown in Figure 1.

There were six participants in our initial pilot experiments. The participants were asked to work on TREC topics using the Google search app, with the voice search activated. The experiment was conducted at the University of Pittsburgh, a major public university in the United States of America. Each pilot participant was compensated $15 (USD) for his or her time and the experiment lasted for about 30-50 minutes. In the pilot experiment, we did not clearly instruct participants how long we expected them to interact with the system for each task. We found that most pilot participants finished the task very quickly without many reformulations. Thus, we modified our experiment protocol by instructing participants that they should stay on one topic for at least two minutes.

3.2 Experiment procedure

With our refined experiment protocol, 20 native speakers of English were subsequently recruited. Each participant was compensated $25 (USD) for their time in the experiment, which lasted for about 90 minutes. At the beginning of the experiment, each participant was trained to work on one TREC topic (Table 1) to ensure that they all knew how to use and interact with the voice search system, and to ensure they all understood the experiment requirements.

We used the Google voice search app on a tablet as our experiment system. All participants' behaviors, including spoken queries, system-transcribed queries and clicking history were recorded. Each participant worked on 25 topics selected from a pool of 50 topics in total. For each topic, the participants could freely interact with the system within a two-minute session (e.g., click and check results, reformulate voice queries), but typing on the tablet was prohibited throughout the experiment.

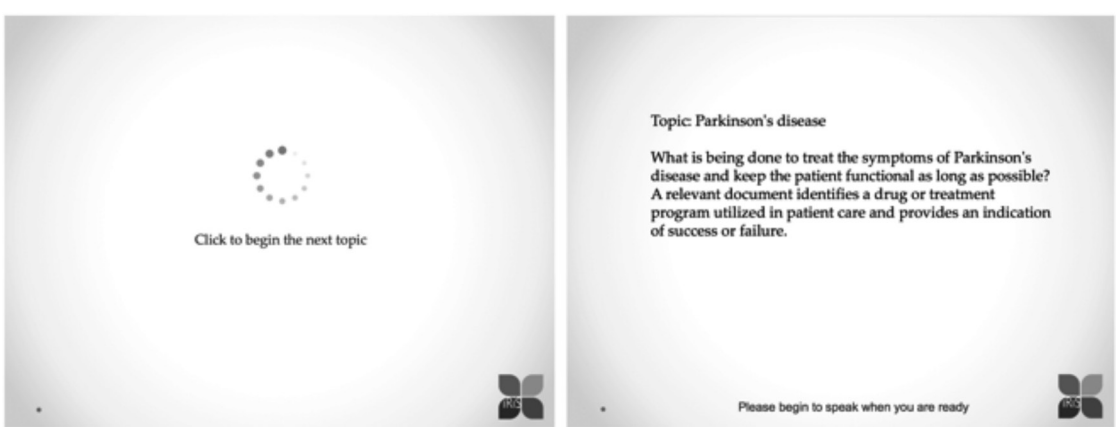

Figure 1. Screenshots of a TREC Topic

While our previous work (Jiang et al., 2013) reports quantitative results regarding the lexical and phonetic query reformulation in voice search, this paper focuses on studying users' perceptions of the difficulties and query reformulation strategies in voice search. The data used in the analysis includes:

1. Participants' background information collected at the beginning of the experiment.
2. Participants' topic ratings (collected after finishing each of the 25 topics) regarding the topic familiarity (i.e., I am familiar with this topic); easiness/difficulty of query formulation (i.e., I find it easy to form a query in this topic). We used three questions, a 6-point Likert scale (strongly disagree, disagree, somewhat disagree, somewhat agree, agree, strongly agree) to present the extent of the measured items. We later recoded the question of easiness with "1" being the most difficult and "6" being the least difficult for presenting the topic complexity.
3. Participants' answers in a semi-structured interview with six overarching interview questions at the end of the experiment. These interview questions include:
   (1) the challenges of search using voice input versus using a keyboard,
   (2) the most difficult topic(s),
   (3) types of words that are not easily recognized,
   (4) the solutions or strategies when a user encounters voice input errors,
   (5) users' affective feelings when recognition errors happened, and finally,
   (6) situations in which they prefer to use/not use voice search.

3.3 Data collection and analysis

Our dataset consists of the experimental record of 20 participants. The experimental record includes Voice Query and Transcribed Query. Voice Query represents what the user actually said, and Transcribed Query represents the automatic recognition result by Google Voice. For the purpose of this study, we focused on Voice Query, and encoded *Lexical Query Reformulation* and *Phonetic Query Reformulation*, as summarized in the Table 2. Specifically, Lexical Query Reformulation considers four actions: Addition, Substitute, Remove, and Reorder; Phonetic Query Reformulation also considers four actions: Different Pronunciation, Spelling, Partial Emphasis, and Whole Emphasis.

Two coders coded the same dataset, and the inter-rater reliability was 0.94. Coders finally came to agreements for the remaining disagreements after a discussion. The detailed procedure of this coding process can be found in Jiang et al. (2013).

4. Results

4.1 Participants

Among the 20 participants, 65% (n=13) were undergraduate students and the rest were graduate students. The average age of the participants was 23.7 (SD=4.72), and 14 were female. Ten participants' majors were in STEM fields (e.g., chemical engineering and biology) and the other ten participants were from the humanities (e.g., French or non-fiction writing) and social sciences (e.g., international affairs or education). When asked about the frequency of using search engines, 85% (n=17) reported that they use search engines on desktop or laptop computers daily, whereas

Table 2. Types of Query Reformulation (Compared with the Prior Query)

Types of query reformulation	Actions	Description (adopted from Jiang et al., 2013, p. 4)
Lexical Query Reformulation	Addition (ADD)	"add new words to the query"
	Substitute (SUB)	"replace words with semantically-related words"
	Remove (RMV)	"remove words from the query"
	Reorder (ORD)	"change the order of the words in a query"
Phonetic Query Reformulation	Different Pronunciation (DIF)	"try different pronunciations for some words (e.g., Puerto Rico)"
	Spelling (SPL)	"spell out each letter in the word"
	Partial Emphasis (PEM)	"phonetically emphasize a part of the current query that also appeared in the previous query"
	Whole Emphasis (WEM)	"phonetically emphasize the whole query that also appeared in the previous query"

only 40% (*n*=8) use search engines on mobile devices every day. Half of our participants reported that they had never used any voice search systems, neither on computers nor on mobile devices.

4.2 Users' perceived difficulties

4.2.1 Voice input errors

In our study, we define *voice input error* as the situation where the search query received and recognized by the voice search system is different from what the user intended to issue. Two types of errors were observed in our experiments (Jiang et al., 2013). 89% are speech recognition errors, i.e. the automatic speech recognition system fails to provide an accurate transcription. 11% are errors caused by improper system interruption, i.e., the user is interrupted by the voice search system before finishing articulation of the query. This happens when the system "believes" that the user has finished speaking before the user has actually finished (e.g., the user pauses for a relatively long period of time but would like to continue speaking).

In the interview, the majority of participants (*n*=12) explicitly expressed that searching via voice input was overall more challenging than conventional search: this approach requires more effort than keyboard search because of the voice input errors. For example, P16 expressed: *"I'd rather type. It takes forever for them (the search engine) to pick up what you're saying*." P14 mentioned: "*In numerous times I had to repeat. Actually, this topic right here, I didn't search for Philippines. It just sort of popped up*." This is consistent with our previous article (Jiang et al., 2013), in which voice input errors were not only responsible for a significant decline of search performance for individual queries, but also led to increased effort and users' negative feelings.

Although there are clear divisions between the two types of errors, the participants did not specifically report whether either one is more serious or troublesome than the other.

4.2.2 Topic familiarity and complexity

Topic familiarity, or users' topic knowledge, can facilitate users when selecting information when searching (Spyridakis & Wenger, 1991). Similar to conventional search systems (using keyboard for query input), voice search users also found that topic familiarity and complexity are factors that affect search difficulty. Four participants stated that topic familiarity was a major obstacle they faced during the experiment: "*I didn't know enough about those topics to re-word the speech properly*" (P01). P07 also reported about topic complexity regarding the topic "marine vegetation": "*... I mean, finding marine vegetation was easy but how it ... but I couldn't find anything on how it was used in relation to food and drug and it kept ...*"

4.2.3 Query formulation

After finishing each topic, we also asked the user to rate the topic on whether it is difficult to formulate queries (using a 6-point Likert scale). We found that users' ratings do correlate with the seriousness of the errors and their actual search performance on the topics.

We characterize the influence of voice input errors by the average proportion of words spoken by the users that were missed in the system's transcription (% missing words), the Jaccard similarity between the results of the voice queries and the transcribed queries, and the drop of nDCG@10 in the transcribed queries compared to the voice queries' actual content. As shown in Table 3, the easier a topic is perceived by the user, the less severe the voice input errors are and the less likely users' search performances are affected by the errors (although the results of two adjacent rating values are sometimes inconsistent). This indicates that users can correctly perceive the difficulties of query formulation.

There are many reasons why it may be difficult to formulate queries for a topic. Aside from topic familiarity and complexity (as reported in last section), difficulties arise when the topic has theme words that are necessary and cannot be replaced and the system has specific difficulty recognizing these theme words. For example, P03 reported that the reason the topic "Culpeper national cemetery" was the most difficult: "*I could not pronounce. I couldn't get the name. I could not even find anything on it*" (P03). In the next section, we will identify the typical difficult words in detail.

Table 3. Users' Perceived Easiness of Topics and the Influence of Voice Input Errors on Users' Search Performance

Perceived difficulty	% Missing words	Jaccard similarity	Drop of nDCG@10
6 (the least difficult)	0.3304	0.4900	0.1023
5	0.2805	0.5140	0.1045
4	0.3274	0.3725	0.1411
3	0.3336	0.4147	0.1187
2	0.3825	0.3261	0.1464
1 (the most difficult)	0.4658	0.1365	0.1831

4.3 Perceived difficult words and reformulation strategies

We asked participants whether they noticed any types of words or phrases that were specifically difficult to be recognized, as well as their reformulation strategies. For each of the following sub-sections, we describe one type of the perceived difficult words and the corresponding strategies.

4.3.1 Acronyms and single-worded queries: Create more clues

When acronyms or single-worded queries are not recognized correctly, participants tended to create more clues (such as using a full name, adding extra words for disambiguation, or changing the part of speech of the word). As shown in Table 4, we categorized acronyms and single-word queries in the same group because users identified their common characteristic: the lack of context. Several participants (n=5) mentioned that acronyms, abbreviations, or very short words could lead to serious recognition issues. For example, P02 reported, "*... short words, like art, was really hard for it to pick up*" (P02). In the search log, we also found that the queries "ER" (the TV show; acronym for "Emergency Room") and "AVP" (acronym for "American Volleyball Professionals") had a 100% error rate.

To cope with this type of difficulty, participants reported that they tried to use the full name of the acronyms, or to add additional clues. For example, "*If I know the word, like ER for example. I kind of like use a keyword that makes it obvious what I'm referring to. ER George Clooney*" (P17).

4.3.2 Frequently misrecognized words with observable phonetic features: Repeat

Participants summarized certain phonetic features of the frequently misrecognized words. For example, P17 reported that some words with syllables that "slide together" were hard for the system to recognize, such as "horse hooves" or "rap and crime." As we examined the search log, we found that the word *rap* in "rap and crime" was misrecognized 13 times out of 36 uses. P17 and P18 both reported that a diphthong word (i.e., two adjacent vowels) caused confusing results (e.g., "hooves" was misrecognized as "who" or "whose"). Participants P04 and P07 also reported that they observed errors on voiced and unvoiced consonants, respectively: "*consonant, P, T, K, those are... it doesn't hear them as well and so for example saying Irish Peace Talks*" (P04); "*Violent. I guess where it ... words that don't have kind of like sharp consonants in them ... to them, it has trouble finding those words, I would guess*" (P07).

In response to this group of errors, some participants (n=3) reported that they would repeat or overstate the error words (e.g., speak slower, clearer, louder) (Type II in Table 4). For example, P07 was asked about how she dealt with the errors of the word "violence": "*I would speak clearly and enunciate. I would definitely speak in a manner that I wouldn't speak to control.*"

4.3.3 Words with pronunciation uncertainty

Words with questionable pronunciation were also perceived as difficult words by the participants. For example, non-English words such as "El Niño" resulted in a high error rate (31 out of 46 being misrecognized). One user tried to pronounce it as the "ninjoʊ" sound: "*my voice's trying to mimic the sounds of the Spanish*

Table 4. Difficult-to-Recognize Words and Corresponding Strategies

Types of words	Example (error rate*)	Users' strategies for given words
A. Acronyms	ER (100%), AVP (100%)	I. ♦ Use full name (e.g., AVP, Association of Volleyball Professionals) ♦ Add extra key word (e.g., ER George Clooney; kiwi fruit) ♦ Change the part of speech (e.g., tax and taxing, P03; use to using, P04)
B. Single-worded queries without context	sun (58.5%), theft (100%), art (45.3%)	
C. Two syllables can slide together easily	rap in "rap and crime" (36.1%)	II. ♦ Repeat the same query with the same tone ♦ Repeat the same query but speak differently in terms of: ○ Making pauses between words (e.g., horse [pause] hooves) ○ Slowing down ○ Putting an emphasis
D. Diphthong	fraud (85.7%), horse (27.8%)	
E. Unvoiced/voiced consonants may fail	**v**iolence (70.4%) "**t**alks" in "Irish Peace talk" (60%), ethnics (47.6%)	
F. Non-English words	El Niño (67.4%)	III. ♦ Try different pronunciations ○ Switch the pronunciations in different languages (e.g., /ninjoʊ/ and /nino/, P05, P07, P11) ○ Trial and error—work around different pronunciations and see which the system will pick up better (e.g., "Falkland", P13) ♦ Spelling letter by letter (e.g., Niño and n-i-n-o, P09) ♦ Avoid perceived difficult words in terms of: ○ Picking a synonym (e.g., theft to "espionage", P09; achievements to "accomplishments", P07); woman to "female", P06) ○ Describing associated things, but nothing directly related (e.g., polygyny to "one man two wives", P19; tornado to "hurricane", P07)
G. Named entities	Ralph (61.1%), Owen (96.2%), Culpeper (66.7%)	
H. Other words that "I don't think I pronounce properly"	polygyny (100%)	

Note. *: occurrence of used times / occurrence of errors

language, didn't come across as well, as the English words [sic]" (P17). Table 5 demonstrates P09's search log and her reformulation strategy. The voice query indicates what a participant's query sounds like, whereas the transcribed query is the one that Google actually picks up. The participant, P09, tried to switch pronunciations back and forth in Spanish and English when "El Niño" kept being misrecognized. Later, she added "flood" and "natural disaster" to create more clues.

Users also reported that they were unfamiliar with the proper pronunciation of some relatively

Table 5. An Example Search Log and Reformulation Strategy of Topic "El Niño" (Participant 09)

#	Voice query	Transcribed query	Reformulation strategy
1	El Niño support	el minya support	- -
2	El Niño support	I mean your support	Try different pronunciations
3	El Niño	aluminium	Emphasis
4	El Niño	Antonio	Repeat
5	flood and el Niño	Antonio	Try different pronunciations and adding context
6	El Niño	and I am now	Try different pronunciations
7	natural disaster el nino	natural disaster antonio	Repeat "el nino" and adding context
8	natural disaster el nino	natural disaster el nino	Repeat

rare words, such as "Culpeper" (18/27) and "polygyny" (8/8). We found that participants used different strategies when they encountered unfamiliar or non-English words. According to the experiment log, P09 spelled out "n-i-n-o" letter by letter when she performed her sixth attempt on the topic. Table 6 shows a participant's (P19) search log when she tried to input the query "polygyny" (with "gyny" pronounced "dʒənɪ"):

Firstly, the user used "gəni," but the system did not return the result that she expected. After repeating the same sound (gəni) with overstating, in the fourth attempt, she pronounced it differently as "gaɪni." However, the "gaɪni" sound seemed to have a critical error as well. Finally, she abandoned the word and used "one man two wives" instead. Therefore, we anticipate that if a user continues to fail after many attempts of saying the same word, it is very possible that the user will employ Type III strategies in Table 4. P18 stated that sometimes the Repeat strategy might not work very well because *"I feel that if you were to say it again there's not going to be a big difference* [sic]*"*. At this point, Type III strategies seem to be "a shelter of last resort" across any type of difficult word, because it can at least generate "some differences."

4.4 Emotions and usage occasions

18 out of 20 participants stated how they felt when encountering voice input errors. Not surprisingly, the majority of participants expressed negative emotions while facing voice input difficulties. More specifically, nine participants stated that they felt frustrated while encountering the voice input errors, two felt annoyed, and one felt angry. On the other hand, three participants expressed that they found the error results hilarious, and the other three participants were not bothered by the input errors.

When asked for the situations where speech searching should be avoided, many participants mentioned public spaces, quiet public spaces (e.g., a library [P19]), or places with a noisy background: *"I guess if you were in a public area and people would be wondering who are you talking to, or if you need to search something quickly and it can't recognize it"* (P18). Interestingly, although some participants feel they

Table 6. An Example Search Log and Reformulation Strategy of Topic "polygyny" (Participant 19)

#	Voice query	Transcribed query	Reformulation strategy
1	polygyny	poligamy	- -
2	polygyny	paul inca ny	Emphasis
3	polygyny	polly guinea	Emphasis
4	polygyny	call gary	Try different pronunciations
5	polygyny	polygamy	Emphasis
6	one man two wives	1 man to live	Describe associated things

should avoid using voice search in a public space, other participants are not very worried about it: *"On the bus all the time people are yapping on their cell phones"* (P12), *"nowadays people have like Bluetooth, so it wouldn't be awkward if you kind of yelled out something"* (P20).

One participant also noted the impossibility of using speech search academically, but this might also be related to avoiding it in a quiet space. *"I don't think it would be useful academically because you wouldn't really want to use a voice search in like a library"* (P19).

When we asked participants to list some situations where they would prefer to use a speech search, many of them stated that voice input is especially helpful when hands are unavailable to use, e.g., while walking (P10), if hands are hurt (P15), or while driving (P19). The participant P14 even provided an interesting scenario that speech voice input is preferred when dealing with very hard-to-spell words in a laboratory: *"sometimes we'll have to use words like* immunohistochemistry *or words like the normal person wouldn't have to use…so maybe if you say it it'll recognize it better. That might be helpful with that there"* (P14).

5. Discussion

By analyzing participants' responses in the interviews, we found that user-recognized difficulties in using voice search systems are both from those related to voice query input and unrelated (e.g., topic familiarity and complexity) inputs. The users perceived query formulation difficulties associated with the seriousness of the error and their performance, clearly indicating that users are aware of this issue. As shown in previous sections, users reported strong dissatisfaction about voice input errors and their tendency to use alternative input methods when encountering errors (*"I'd rather type* (P16)*"*), which suggests the necessity of equipping multi-modal query inputs in current voice search systems. As we restricted users to only utilizing voice inputs in our experiments, it is interesting and necessary to further explore user interactions in systems with multi-modal query inputs. This also indicates that voice search and similar applications (such as intelligent voice assistants) should be equipped with assistance techniques for voice query formulation and reformulation. Unfortunately, such techniques are rare. For example, query

auto-completion is included in almost all current conventional search engines as an important assistance feature, but similar techniques for voice search, to the best of our knowledge, do not exist.

We categorized three types of difficult words reported by the users, which helps us explain why certain categorization of words have high error rates, a problem that we left unanswered in our previous study (Jiang et al., 2013). Based on Table 4, most of the categories in our previous work also existed in those reported by the users (i.e., acronyms, named entities, and non-English words). Participants also provided possible explanations for some of the uncategorized words with high error rates, e.g., "sun", "talks." This may provide people studying automatic speech recognition with firsthand examples of errors.

Although "partial emphasis" (overstating a part of the voice) and query term substitution are the two reformulation patterns mostly highly associated with previous error words, our participants did not specify the cases of using the latter. We suspect this is probably due to the fact that participants were only aware of and thus summarized the acoustic features of the words with high error rates.

We also found that a participant's first reaction to voice input errors is to repeat or improve their pronunciation (Strategy II in Table 4), rather than switch to alternative words (Strategy III). However, Strategy II worked less successfully than query term substitution. Therefore, voice search systems may benefit from providing better guidance or suggesting the adoption of reformulation strategies, e.g., reminding users it is probably more effective to try other words when speech recognition errors occur. Additionally, as two of the participants adopted spelling as a query input strategy, voice search systems should probably support such user actions.

We further found that the majority of participants in this study expressed negative emotions (i.e., mostly frustrated) while encountering voice input difficulties during their voice search. Moreover, our participants revealed both environmental and social concerns about using voice search. For example, some expressed concerns about the awkwardness of using voice search in public. Such issues do not exist in most scenarios of using conventional search engines, which are assumed to be private and protected. Implications drawn from users' preferred situations suggest that voice search is best suited for users whose hands are unavailable. This indicates a possible need for applying voice search in mobile applications.

6. Conclusion

We found that users identified voice input errors and topic familiarity and complexity as the major obstacles of voice search. The users also reported the characteristic types of words that were difficult to be recognized and the corresponding reformulation strategies to solve the issues, as well as their feelings and preferred usage situations. For most of the misrecognized words, the most natural reformulation technique is to repeat it again. When the queries have a lack of context, e.g., acronyms and single-worded query, users tend to create more clues by using a full name or adding extra words. Spelling letter by letter is the fail-safe option, which is usually a user's final step. These findings help us better understand the current issues and user interaction

within voice search. Future work is needed in order to design a better mechanism for allowing users to easily reformulate their queries.

References

Anick, P. (2003). Using terminological feedback for web search refinement: A log-based study. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 88-95. doi: 10.1145/860435.860453

Bohus, D., & Rudnicky, A. I. (2005, September). *Sorry, I didn't catch that!-An investigation of non-understanding errors and recovery strategies*. Paper presented at the Proceedings of 6th SIGdial Workshop on Discourse and Dialogue, Lisbon, Portugal.

Bruza, P., & Dennis, S. (1997). Query reformulation on the internet: Empirical data and the hyperindex search engine. *Proceedings of the Conference Adaptivity, Personalization and Fusion of Heterogeneous Information, 97*, 488-499.

Crestani, F., & Du, H. (2006). Written versus spoken queries: A qualitative and quantitative comparative analysis. *Journal of the American Society for Information Science and Technology, 57*(7), 881-890. doi: 10.1002/asi.20350

Huang, J., & Efthimiadis, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 77-86. doi: 10.1145/1645953.1645966

Jansen, B. J., Booth, D. L., & Spink, A. (2009). Patterns of query reformulation during web searching. *Journal of the American Society for Information Science and Technology, 60*(7), 1358-1371. doi: 10.1002/asi.21071

Jiang, J., Awadallah, A. H., Jones, R., Ozertem, U., Zitouni, I., Kulkarni, R. G., & Khan, O. Z. (2015). Automatic online evaluation of intelligent assistants. *Proceedings of the 24th International Conference on World Wide Web*, 506-516. doi: 10.1145/2736277.2741669

Jiang, J., Jeng, W., & He, D. (2013). How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 143-152. doi: 10.1145/2484028.2484092

Kim, T. H. J., Gupta, P., Han, J., Owusu, E., Hong, J., Perrig, A., & Gao, D. (2012). OTO: Online trust oracle for user-centric trust establishment. *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, 391-403. doi: 10.1145/2382196.2382239

Narang, A., & Bedathur, S. (2013). Mind your language: Effects of spoken query formulation on retrieval effectiveness. arXiv:1312.4036 [cs.IR]

Raux, A., Langner, B., Bohus, D., Black, A. W., & Eskenazi, M. (2005). Let's go public! Taking a spoken dialog system to the real world. *Proceedings of Interspeech 2005*, 885-888.

Rieh, S. Y., & Xie, H. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval

context. *Information Processing & Management, 42*(3), 751-768. doi: 10.1016/j.ipm.2005.05.005

Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., ... Strope, B. (2010). "Your word is my command": Google Search by voice: A case study. In A. Neustein (Ed.), *Advances in speech recognition* (pp. 61-90). New York, NY: Springer. doi: 10.1007/978-1-4419-5951-5_4

Shin, J., Narayanan, S., Gerber, L., Kazemzadeh, A., & Byrd, D. (2002). Analysis of user behavior under error conditions in spoken dialogs. *Proceedings of Interspeech 2002*, 2069-2072.

Shokouhi, M., Jones, R., Ozertem, U., Raghunathan, K., & Diaz, F. (2014). Mobile query reformulations. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1011-1014. doi: 10.1145/2600428.2609497

Spyridakis, J. H., & Wenger, M. J. (1991). An empirical method of assessing topic familiarity in reading comprehension research. *British Educational Research Journal, 17*(4), 353-360. doi: 10.1080/0141192910170405

Swerts, M., Litman, D., & Hirschberg, J. (2000). Corrections in spoken dialogue systems. *Proceedings of the International Conference on Spoken Language Processing, 2*, 615-618.

(Received: 2015/10/1; Accepted: 2016/3/10)

圖書資訊學刊　第14卷第1期　（2016.6）　頁25-39
doi:10.6182/jlis.2016.14(1).025

Google Voice語音搜尋中使用者對困難的感受與檢索詞重構策略

Users' Perceived Difficulties and Corresponding Reformulation Strategies in Google Voice Search

鄭　瑋[1]　姜捷璞[2,†]　何大慶[3]

Wei Jeng[1], Jiepu Jiang[2,†], Daqing He[3]

摘　要

語音搜尋（Voice search）為近年來行動裝置應用上的趨勢，透過使用者實驗與後測訪談，本研究旨在探討一般使用者利用Google Voice進行語音搜尋並遭遇困難時，所因應的檢索詞重構策略。本研究的結果揭示了：(1)受測者進行語音搜尋時，常遭遇的困難有語音辨識錯誤（speech recognition errors）以及主題複雜度（topic complexity）；(2)受測者面對容易辨識錯誤的字詞時（如縮寫字、單音節詞、外來語等），自然而然地發展出若干不同的因應策略；(3)受測者在面對辨識錯誤結果的情緒反應（emotional reactions）皆有不同，以及表達了其偏好的語音搜尋適用情境（usage occasions）。

關鍵字：語音搜尋、語音輸入、語音辨識錯誤、檢索詞重構、Google Voice

[1,3]美國匹茲堡大學資訊科學學院
School of Information Sciences, University of Pittsburgh, USA
[2] 美國麻薩諸塞大學電腦科學學院智慧資訊檢索中心
Center for Intelligent Information Retrieval, School of Computer Science, University of Massachusetts, USA
† 本研究為姜捷璞於美國匹茲堡大學資訊科學學院時完成。
* 通訊作者Corresponding Author: 鄭瑋Wei Jeng, E-mail: wej9@pitt.edu

註：本中文摘要由作者提供。
以APA格式引用本文：Jeng, W., Jiang, J.-P., & He, D.-Q. (2016). Users' perceived difficulties and corresponding reformulation strategies in Google Voice Search. *Journal of Library and Information Studies, 14*(1), 25-39. doi: 10.6182/jlis.2016.14(1).025
以Chicago格式引用本文：Wei Jeng, Jiepu Jiang and Daqing He. "Users' perceived difficulties and corresponding reformulation strategies in Google Voice Search." *Journal of Library and Information Studies* 14, no. 1 (2016): 25-39. doi: 10.6182/jlis.2016.14(1).025