How Do Users Respond to Voice Input Errors? Lexical and Phonetic Query Reformulation in Voice Search

Jiepu Jiang
School of Information Sciences,
University of Pittsburgh
jiepu.jiang@gmail.com

Wei Jeng
School of Information Sciences,
University of Pittsburgh
wej9@pitt.edu

Daqing He
School of Information Sciences,
University of Pittsburgh
dah44@pitt.edu

ABSTRACT

Voice search offers users with a new search experience: instead of typing, users can vocalize their search queries. However, due to voice input errors (such as speech recognition errors and improper system interruptions), users need to frequently reformulate queries to handle the incorrectly recognized queries. We conducted user experiments with native English speakers on their query reformulation behaviors in voice search and found that users often reformulate queries with both lexical and phonetic changes to previous queries. In this paper, we first characterize and analyze typical voice input errors in voice search and users' corresponding reformulation strategies. Then, we evaluate the impacts of typical voice input errors on users' search progress and the effectiveness of different reformulation strategies on handling these errors. This study provides a clearer picture on how to further improve current voice search systems.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation, relevance feedback.*

General Terms

Measurement, Experimentation, Human Factors.

Keywords

Query reformulation; voice search; voice input errors.

1. INTRODUCTION

Supporting query reformulation has long been recognized as an important strategy to help users further their search progress [3]. Users may need to reformulate queries several times until their information needs are fully satisfied. The need for reformulation may be attached to the users themselves. As users may have limited understanding of their information needs, the retrieval system and the collection, it is difficult for them to develop one single query to complete the search. At the same time, the need for reformulation may come from search problems being explorative where relevant documents may be scattered among different subtopics, so that it is impossible to retrieve all relevant documents with a single query. Therefore, many studies [7, 22] concentrated on supporting reformulation of textual queries.

Along with the rapidly increasing usage of mobile devices and the improvement of speech processing, voice search becomes an alternative search mode. During voice search, users can vocalize their queries and the retrieval system utilizes the voice recognition results for retrieval [6, 19]. Though previous studies found that query reformulation plays an important role in conventional textual search systems, to the best of our knowledge, there are very limited studies on voice search, especially concerning users' query reformulation in voice search.

In this paper, we therefore focus on explaining query reformulation behaviors in the context of *voice search*. The term *voice query*[1] refers to the query in voice search. It contains not only the lexical contents, but also the phonetic characteristics such as the speaker's stress, speed, and intonation. In comparison, we refer to those searches in which users need to type queries on a keyboard as *conventional searches*.

We mainly concentrate on three research objectives in this study. First, voice search relies on users' vocalization of queries and systems' automatic speech recognition to transcribe voice queries, which may result in various *voice input errors*. Voice input errors include not only the errors from automatic speech recognition but also the system's interruptions during users' vocalization of queries. Therefore, our first objective is to characterize the types of voice input errors in voice search and evaluate their impacts on voice search.

Second, upon recognition of voice input errors, users will take actions in their subsequent query reformulation to overcome the errors. As voice queries involve both lexical and phonetic characteristics, users' reformulation choices and preferences would also be different from those in conventional searches. Therefore, our second objective is to identify and characterize users' query reformulation patterns in voice search.

Third, as the ultimate goal of this study is to shed light on how to support query reformulation in voice search, it is important to analyze users' preferences of using different reformulation patterns and examine the effectiveness of the reformulation patterns in handling voice input errors. In this study, we evaluate the effectiveness of the reformulation patterns by how they overcome the voice input errors and improve the retrieval performance.

To meet our research objectives, we conducted a series of voice search experiments involving native English speakers working on TREC search topics using the Google voice search app on the iPad. The participants were only permitted to speak voice queries to initiate searches and reformulate queries. Within a certain time limit, the participants could freely issue multiple voice queries, read or click on returned search results, and use Google's query suggestions. Users' voice queries, the system's transcription

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
Copyright © 2013 ACM 978-1-4503-2034-4/13/07…$15.00.

[1] In this paper, we use voice queries to refer to *spoken queries* and *speech queries*, which were used in previous studies [5]. Our rationale is to keep a consistency with Google Voice Search, the platform used in our experiment.

results to the voice queries, and the clicked documents were all recorded for analysis.

The rest of the paper is organized as follows: Section 2 reviews related studies in query reformulation and voice-based search; Section 3 introduces our methods for experimentation and analysis; in Section 4, we characterize voice query input errors and voice query reformulation patterns; Section 5 evaluates the impacts of voice input errors on voice search; Section 6 evaluates the effectiveness of each type of voice query reformulation; finally, we discuss suggestions for future development of voice-based search systems and outline our conclusions.

2. RELATED WORKS

2.1 Voice Search

Voice search [8, 23] or voice-enabled search [2, 20] refer to the search systems that allow users to input search queries in a spoken language and then retrieve the relevant entries based on system-generated transcriptions of the voice queries. Currently, voice search is commonly applied via mobile devices. Researchers examined the scenario of using voice search compared with traditional desktop search. For example, Schalkwyk et al. [19] analyzed Google's search logs and found that users utilized Google Voice Search more frequently when they tried to find information such as food and local geographical information (e.g. city names and local restaurants). However, it remains unclear whether the location-related information needs are intrinsically related to voice search, or are due to the fact that the current devices supporting voice search are mostly mobile devices.

Existing studies on voice search are very limited, especially those related to users' voice queries and query reformulations. Schalkwyk et al. [19] reported statistics of queries from Google Voice's search logs which found that voice queries are statistically shorter than desktop search queries. Crestani et al. [6] conducted a user experiment based on collections of users' voice queries. However, the experiment environment did not involve a real search system. Participants were asked to formulate voice queries without knowing whether their voice queries could be recognized, or if they would retrieve meaningful results. In comparison, in our experiments, participants freely interacted with the voice search systems, so that the participants' interactions, particularly their responses to voice input errors, could be collected and studied.

2.2 Query Reformulation

The lexical query reformulation patterns we adopted in this paper come from the summarization of many previous studies, including [9–11, 18, 21]. As we did not aim to create a systematic taxonomy of the reformulation patterns in voice search, we simplified the patterns to only four types: addition, substitution, removal, and re-ordering. However, our substitution pattern involved many other patterns defined in previous works, such as stemming [9, 21] and acronyms [9]. Also, many textual reformulation patterns that do not exist in voice search were removed, including: punctuation [21], URL stripping [9], substring [9], spelling correction [1], and capitalization [21].

3. METHODS

3.1 Settings and Experiment System

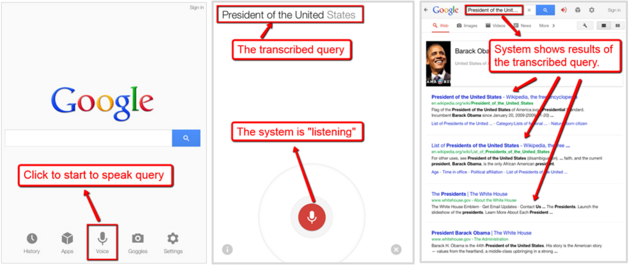
As stated, we are interested in users' query reformulation behaviors in voice search, especially how they utilize query reformulations to cope with voice input errors. Admittedly, as currently voice search is mostly used on mobile devices, an ideal experiment setting for our study should simulate mobile search environment, including many issues previously found to have an impact on automatic speech recognition (ASR) and voice search, such as the background noise [19]. However, after consideration, we decided to conduct our study in a controlled laboratory experiment setting for the following reason: our focus is on how users change their queries when voice input errors happen in voice search. Therefore, automatic speech recognition (ASR) errors and the often concerned noise and vocabulary issues in ASR [19], though important in voice search, are just part of the problem and have secondary importance in our study.

Among the state-of-the-art web search engines that support voice search, we adopted the Google search app on the iPad for our experiment because of the popularity of Google in conventional web search. We believed that users with Google search experience could more easily understand its voice search function. In addition, using the iPad for experiment also replicated some form of mobile search environment.

As our study focus on query reformulation behaviors in voice search, we simply adopted Google voice search as an out-of-box system, despite it is unclear how the voice search system and its ASR were implemented. Although the voice search system and the effectiveness of ASR can influence experiment results, we believe that Google voice search system is probably the best choice for this experiment and the experiment results would be still representative of users in other voice search systems.

Figure 1 contains screenshots of the system[2]. As shown in Fig.1 (a), a user can touch the voice search icon to issue voice queries. If the user stops speaking or pauses for a while, the system concludes that the user has completed the voice query. Then it starts the recognition of the voice query and uses the transcribed query to search (see Fig.1 (b)).

Google voice search system provides different audio cues to indicate its various statuses, which includes: starting or stopping "listening" a voice query; displaying the transcribed query; and failing to generate the transcribed query. These audio cues are very useful in our transcriptions of the experiment recordings.


Sign in
Google
Click to start to speak query
History
Apps
Voice
Goggles
Settings
President of the United States
The transcribed query
The system is "listening"
Google
President of the Unit...
Sign in
Web
Images
Videos
News
More
Barack Oba
System shows results of the transcribed query.
President of the United States - Wikipedia, the free encyclopedia
List of Presidents of the United States - Wikipedia, the free ...
The Presidents | The White House
President Barack Obama | The White House

(a) (b) (c)

Figure 1. Screenshots of the Google search app on iPad.

3.2 Search Tasks and Topics

Our experiment setting is similar to the one adopted by the TREC session track [17], in which users can issue multiple queries to work on one search topic.

Ideally, search topics should be representative of users' information needs in the mobile search environment. However, as discussed in Section 3.1, our experiment setting was not a real mobile environment, therefore we selected conventional TREC ad hoc search topics in our experiments. On the one hand, we could not find many mobile search topics due to limited resources. On the other hand, we also wanted to study the connections between query reformulation in voice search and those in conventional

[2] The screenshots were made in January 2013, 6 months after the experiments. However, the main system features did not change.

textual queries as part of our future study. Therefore, we selected 50 TREC topics for our study, of which 30 are from the TREC robust track in 2004, and 20 are from the TREC web track in 2009 and 2010. The selected topics were representative of informational search problems [4]. Table 1 shows the selected TREC topic numbers.

Although the literature shows that many searches on mobile devices involve location-related information needs [19], we did not want to restrict our findings by not including other types of information needs. The first reason is that there is no absolute demarcation line between mobile devices and portable computers. The second is that many voice search systems such as Google can be used on laptop and desktop computers.

Table 1. Selected 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, 672, 683, 698
Web Track 2009, 2010	51, 52, 54, 56, 68, 70, 72, 73, 74, 91, 94, 100, 104, 107, 108, 110, 112, 113, 122, 141

3.3 Experiment Procedure

We recruited 20 participants (14 females and 6 males). The majority of them were college students (13) and graduate students (3). All 20 participants were native English speakers, and their average age is 23.7 with standard deviation being 4.72.

Each participant was compensated financially for their involvement in the experiment, which lasted for about 1.5 hours. At the beginning of the experiment, each participant was trained to work on one TREC topic (other than the 50 topics in table 1) to make sure that they knew how to use the voice search system, and were clear about what operations they were allowed to do during the experiments.

They then each worked on 25 of the 50 topics listed in Table 1. We alternated the topic assignments to reduce learning and fatigue bias. For each topic, the participant first read the topic description on a computer screen and then worked on each topic using the Google voice search app on iPad for 2 minutes. The participant could only vocalize queries, browse and click the search results, and use Google's query suggestions. The participant was not allowed to type queries on the iPad touch screen. After each topic, the participants were asked to answer a short questionnaire regarding their perceptions of topic difficulty and familiarity.

The experiment paused for a 5-minute break after the participant finished 15 topics. When all 25 topics were completed, each participant was interviewed for about 10 minutes on his/her perceptions of the voice search and query reformulation. The whole experimental process was recorded for later transcription and analysis of users' voice queries and interviews.

3.4 Data

Two coders manually transcribed the voice queries and agreed on 100% of the transcribed texts except for the use of plurals and prepositions (which are difficult to identify and usually do not affect search results after stemming and stopword removal).

Google's search history automatically records the system's transcribed queries and the users' click-through pages. For each participant, we created a Google user account and recorded the user's search history during the experiments.

Each participant went through a semi-structured interview at the end of the experiments on their opinions of using voice search systems and especially on how they constructed and reformulated voice queries. Some of the interview questions were based on our own experience of using voice search and a pilot study. As the study was highly exploratory, we also developed new interview questions for the experiments. We hired a professional transcription company to transcribe the interview texts.

The experiment was conducted in July 2012. In total, we collected 1,650 voice queries and 32 cases of using query suggestions. On average, each subject issued 3.30 voice queries per topic (SD=2.50). Among the 1,650 voice queries, 742 were correctly recognized. Voice input error happened in 908 voice queries, of which 810 were caused by speech recognition error and 98 by system interruption. We also found 42 voice queries for which the system could not provide any transcription results. For these queries, we simply counted their transcribed queries as empty strings and their search results as empty lists. These voice queries and query suggestions provided us with 1,182 query reformulation pairs. The average number of results clicked by a user throughout the session of a search topic were 1.41 (SD=1.14). On average, for each topic, 9.76 unique clicked results were aggregated as qrels (SD=3.11).

3.5 Search Effectiveness Evaluation

For each topic, we assume that a set of topic-level relevant results can be collected to evaluate each query in the search sessions dealing with that topic. Such evaluation method was widely adopted in multi-query search session evaluation, e.g. [12–14, 16]. Due to the time limitation of the experiments, we did not ask the participants to make relevance judgments, but relied on the clicked results as relevant documents for evaluation. Similar methods were widely adopted in web search [15].

Due to the voice input errors, sometimes a participant will not be able to find any meaningful results within the 2-minute session. Thus, for each topic, we aggregated the results clicked by any of the participants when they were working on that topic. Each clicked result was assigned a relevance score of 1 for that topic. Other results were considered non-relevant (relevance score is 0). On such a basis, we can calculate standard evaluation metrics such as nDCG of the queries.

Note that this method will be biased toward the transcribed queries in evaluation, because only those results retrieved by the transcribed queries have the chance to being clicked upon (i.e. some of the voice queries' results were not clicked upon because they were never shown to the participants). Thus, the evaluated effectiveness of the voice queries may be underestimated. However, this problem does not affect the validity of our study. As will be shown in Section 6, even if they are underestimated, voice queries still outperform their corresponding transcribed queries in nearly all the cases.

Google search history only records clicked results of queries. Thus, we crawled the first page of Google results for each of the voice queries and system transcribed queries. These results were accessed 5 months after our experiments. Although these results may be somewhat different from those at the time we conducted the experiments, we assume they do not influence the comparison between queries.

4. VOICE INPUT ERRORS AND REFORMULATION PATTERNS

4.1 Voice Input Error Identification

In voice search, a user speaks a *voice query* (q_v), and the search system generates the transcription of q_v for search, which is referred to as the *transcribed query* (q_{tr}). We say a *voice input error* occurs when the actual content of a voice query q_v differs from its transcribed query q_{tr}. Let {$q_v^{(1)}$, … , $q_v^{(n)}$} be *n* voice queries, and {$q_{tr}^{(1)}$, … , $q_{tr}^{(n)}$} be the corresponding *n* transcribed queries. The transition from $q_v^{(i)}$ to $q_v^{(i+1)}$ is referred to as a *voice query reformulation*.

Through comparison of manually transcribed voice queries with system transcribed voice queries, we can obtain recognition errors, which include:

Missing words: words in q_v that do not appear in q_{tr}.

Incorrect words: words in q_{tr} that do not appear in q_v.

When identifying recognition errors in this experiment, we did not consider the word differences caused by letter case (e.g. "United States" and "united states" are considered as equivalent) and plurals (e.g. "neil young tickets" is considered as equivalent to "neil young ticket"). The reason for this is that these types of errors do not have a significant impact on search results.

In addition to the system's speech recognition errors, voice input errors can also be caused by the system's interruption of the participants' voice inputs. While vocalizing a query, if the participant pauses for a certain amount of time, the system will "think" that the participant has completed the query. So the system will stop listening to the participant's voice input and directly transcribe the unfinished voice query for search. This type of error can be reliably identified by listening to the recording. The participant would pause and then start to talk again but the system had already issued the audio cue for stopping listening. Therefore, we manually annotated each voice query with one of the following four categories, two of which indicate the voice input error type:

Speech Recognition Error: the participant completed a query without any interruption, but the voice query was not recognized correctly. This error can be characterized by missing words or/and incorrect words as mentioned earlier.

System Interruption: the participant was improperly interrupted by the system and failed to speak all of the query words.

No Error: no voice input error.

Query Suggestion: the participant used a Google's query suggestion. If the search history recorded that the participant searched for a query while we did not hear it in the recording, we consider that to be a case of using Google's query suggestion.

During the annotation of voice input errors, the two coders agreed on 100% of the voice queries' category types. Because the participants usually stopped speaking when system interruption happened, we cannot determine the unspoken contents of the queries (i.e. for queries with system interruption, we can only have information on q_{tr} but not q_v). Thus, in much of the later analysis that requires the information on q_v, we mainly focus on queries without voice input errors and those with speech recognition errors.

4.2 Voice Query Reformulation Patterns

As voice queries have both lexical and phonetic characteristics, voice query reformulation can incorporate not only textual changes to the query but also phonetic changes. Thus, voice query reformulation can have *lexical query reformulation*, *phonetic query reformulation* or both. In the remainder of this section, we will discuss the patterns of voice query reformulation, which were summarized from previous works [9] and our observations on the experiment's results.

4.2.1 Lexical Query Reformulation

Expanded from previous studies [9], we characterized lexical query reformulation into addition, substitution, removal, and re-ordering of words, or the combination of these patterns. Although these patterns also exist in conventional search, users may utilize them for different reasons in voice search.

Addition (ADD): adding new words to the query. We refer to the newly-added words as *ADD words*. For example:

	Voice Query	Transcribed Query	ADD words
q_1	the sun	the son	
q_2	the sun solar system	the sun solar system	solar system

Substitution (SUB): replacing words with semantically-related words. In voice search, we noticed that users may substitute the words that were incorrectly recognized with other words of similar meanings. We refer to the words being replaced and the new words as *SUB words*. For example:

	Voice Query	Transcribed Query	SUB words
q_1	art theft	test	
q_2	art embezzlement	are in Dublin	theft→embezzlement
q_3	stolen artwork	stolen artwork	embezzlement→stolen art→artwork

Different from the substitution pattern in [9], we also count "acronym", "abbreviation", and "word stemming" in [9] as word substitution patterns, for example:

avp → association of volleyball professionals
united states → us
ireland peace talk interruption → ireland peace talk interrupted

Removal (RMV): removing words from the query. In voice search, we noticed that the participant may remove a part of a voice query, if the part was not correctly recognized and was not essential to the search topic. The words being removed are referred to as *RMV words*. For example:

	Voice Query	Transcribed Query	RMV words
q_1	advantages of same sex schools	andy just open it goes	
q_2	same sex schools	same sex schools	advantages of

Re-ordering (ORD): changing the order of the words in a query. The words being re-ordered are referred to as *ORD words*. In voice search, we noticed that the words being re-ordered are usually those wrongly recognized. For example:

	Voice Query	Transcribed Query
q_1	interruptions to ireland peace talk	is directions to ireland peace talks
q_2	ireland peace talk interruptions	ireland peace talks interruptions

4.2.2 Phonetic Query Reformulation

Phonetic query reformulation is unique in voice search. During our transcription of experiment recordings, we found the following human recognizable phonetic query reformulation patterns:

Partial Emphasis (PE). Partial emphasis refers to the behavior of phonetically emphasizing a part of the current query that also appeared in the previous query. Typically, the users can put stress (STR) on certain words, or slow down (SLW) at these words, or use both. Sometimes the users may only emphasize a vowel or consonant in the word. We also noticed other ways of emphasizing words when speaking voice queries. For example, some users spell out each letter in the word (SPL), or try different pronunciations (DIF) for some non-English words (e.g. Puerto Rico). Overall, STR and SLW are the two primary patterns of partial emphasis, whereas SPL and DIF occurred rarely in our experiments. The recurring words being emphasized during speaking are referred to as *PE words*. We use the following methods to represent the PE methods:

PE	Example	Explanation
STR	***rap*** and crime	put stress on "rap"
SLW	rap and c-r-i-m-e	slow down at "crime"
SPL	P·u·e·r·t·o Rico	spell out each letter in "Puerto"
DIF	Puerto Rico	pronounce "Puerto" differently

In voice search, we notice that the part of the query being emphasized is usually that part being incorrectly recognized in previous searches. For example:

	Voice Query	Transcribed Query	PE words
q_1	rap and crime	rap and crying	
q_2	rap and c-r-i-m-e	rob and crime	crime
q_3	***rap*** music influence	rap music influence	rap

Whole Emphasis (WE). Whole emphasis is to place emphasis on every part of the query, usually by putting stress or slow down on each of the words. It usually happens when the majority of the previous query was wrongly recognized. For example:

	Voice Query	Transcribed Query
q_1	art embezzlement	are in dublin
q_2	a-r-t e-m-b-e-z-z-l-e-m-e-n-t	art embezzlement

We did not find other meaningful phonetic reformulation patterns other than PE and WE in our transcription.

4.2.3 Recognition of Query Reformulation Types

We recognize lexical query reformulation types by automatic and manual methods. Let $q_1 \rightarrow q_2$ be a lexical query reformulation, then the procedures of recognizing the patterns are:

Step 1: automatically check whether all words in q_1 also appear in q_2. If yes, any extra words in q_2 are recognized as ADD words, and q_2 is an ADD of q_1. Similarly, if all q_2's words are in q_1, any extra words in q_1 are recognized as RMV words, and q_2 is an RMV of q_1.

Step 2: For the rest of the query pairs, check manually whether q_2 contains SUB words of q_1. The two coders agreed on 93.9% of the cases at the beginning, and finally came to agreements on the remaining 6.1% after further discussion.

Step 3: Compared with q_1, if some newly appeared words in q_2 are not recognized as SUB words, we mark them as ADD words and q_2 as an ADD of q_1. Similarly, if q_2 removed some words in q_1 and the removed words are not substituted by other words, we mark them as RMV words and q_2 as an RMV of q_1.

Step 4: Finally, if two words appeared in both q_1 and q_2, and their sequence was changed, we mark q_2 as an ORD of q_1.

Note that ADD, RMV, SUB, and ORD are not exclusive of each other. For example:

Reformulation	q_1: information retrieval system q_2: search system development
Reformulation Type & Words	ADD: development SUB: retrieval → search RMV: information

The phonetic query reformulation types and the PE words were manually recognized. In transcribing the recordings, we found that STR and SLW almost always happened together. Thus, we mark STR and SLW as one type "STR/SLW". Finally, we come to four exclusive phonetic reformulation patterns: STR/SLW, SPL, DIF, and WE. The two coders agreed on 87.6% of the cases at the beginning, and finally came to agreement on the remaining 12.4% after further discussion.

5. INFLUENCE OF VOICE INPUT ERROR

5.1 Voice Input Errors in Individual Queries

RQ1*: How do speech recognition errors affect voice queries?*

Speech recognition error is the major type of voice input error. It occurred in 810 voice queries (89.2% of all 908 queries with voice input errors in our study). We found that speech recognition error can greatly change the content and results of voice query, most likely hurting the performance of voice search.

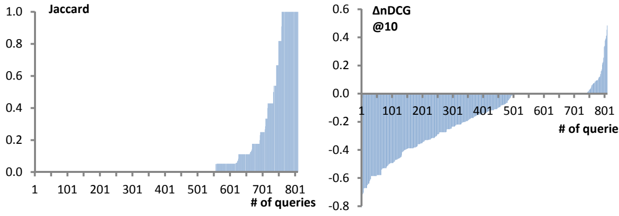
At the word level, we calculated the average percentage of missing words in voice queries and the average percentage of incorrect words in transcribed queries. As shown in Table 2, when speech recognition error occurred, about half of the words (49.7%) in voice queries were missing in the transcribed queries. Similarly, about half of the words (49.3%) in transcribed queries were incorrect transcriptions. On average there were 1.77 missing words and 1.84 incorrect words per query.

Such high proportions of missing words and incorrect words greatly affected the results of voice search. For each of the 810 voice queries with speech recognition errors, we calculated the Jaccard similarity of Google's first pages of results between voice query and transcribed query (i.e. Jaccard(q_v, q_{tr}) in Table 2). As shown in Table 2, the average Jaccard similarity was only 0.118, indicating very low overlap between those retrieved by the transcribed queries and those that should have been retrieved by the voice queries' true content. Figure 2(a) further illustrated the low overlap by showing the distribution of Jaccard similarity, which indicated that, for 69% (556 out of 810) of voice queries with speech recognition errors, the search results will be totally different from users' expectations (i.e. Jaccard similarity is 0).

Table 2. Comparison of voice queries that contained no errors, speech recognition errors, or system interruptions.

	No Errors 742 Queries		**Speech Recognition Errors 810 Queries**		**System Interruptions 98 Queries**	
	mean	**SD**	**mean**	**SD**	**mean**	**SD**
nDCG@10 of q_v	0.275	0.20	0.264	0.22	-	-
nDCG@10 of q_{tr}	0.275	0.20	**0.083**$^{*\downarrow}$	0.16	**0.061**‡	0.14
Length of q_v	3.82	1.68	**4.14***	1.99	-	-
Length of q_{tr}	3.82	1.68	**4.21***	2.31	**2.34**‡	1.41
# missing words	-	-	1.77	1.09	-	-
# incorrect words	-	-	1.84	1.44	-	-
% missing words	-	-	49.7%	29%	-	-
% incorrect words	-	-	49.3%	31%	-	-
Jaccard(q_v, q_{tr})	-	-	0.118	0.27	-	-
ΔnDCG@10	-	-	-0.182	0.23	-	-

*: the difference between queries with no errors and recognition errors is significant at 0.01 level according to Welch t-test; ‡: the difference between queries with no errors and system interruptions is significant at 0.01 level according to Welch t-test; $^{\downarrow}$: the difference between q_v and q_{tr} under the same error conditions is significant at 0.01 level according to paired t-test.



(a) (b)

Figure 2. Jaccard similarity and ΔnDCG@10 of the top 10 results of q_v and q_{tr} for 810 queries with recognition errors.

In addition, speech recognition errors hurt the performance of voice search significantly. As shown in Table 2, the average nDCG@10 of the 810 voice queries with speech recognition errors was 0.084. However, if all the speech recognition errors were corrected, the average nDCG@10 could be significantly improved to as high as 0.264, comparable to the average nDCG@10 of voice queries with no voice input errors (0.275).

Figure 2(b) further shows the distribution of ΔnDCG@10 for the 810 queries (i.e. the difference of nDCG@10 between the transcribed query and the voice query). For 500 queries (62% of the 810), nDCG@10 declined. The remaining 310 queries, whose search performance was not hurt, were intrinsically inefficient queries. Even inputted correctly, these queries could only have an average nDCG@10 value of 0.117, which is significantly less than other queries. Therefore, these queries' performance was not hurt probably because there was not much room to degrade their search performance.

***RQ2**: How do system interruptions affect voice queries?*

System interruptions occurred in 98 queries (10.8% of all 908 queries with voice input errors), which also greatly altered the content of queries and hurt the performance of voice search. When system interruption occurred, it was impossible to determine the real content of the voice queries. Therefore, we calculated statistics only for the transcribed queries.

Compared with the 742 correctly recognized voice queries, the 98 queries with system interruptions performed significantly worse (0.061 vs. 0.275 in average nDCG@10). When system interruption occurred, the transcribed queries were also significantly shorter than those of the correctly recognized queries (2.34 vs. 3.82 words), probably because the users were interrupted improperly and were not able to vocalize the entire query words.

***RQ3**: When do speech recognition errors happen?*

We found that query length may be one factor related to speech recognition errors. As shown in Table 2, queries with speech recognition errors were significantly longer than those correctly recognized queries (4.14 vs. 3.82 words). On the one hand, this is not surprising: as recognition error may happen in any word of a voice query, the more words spoken, the more likely an error happens. On the other hand, the longer the query, the richer the context it provides, which may help the speech recognition. Therefore, further study is needed on whether or not query length can affect speech recognition errors.

We also explored the relationship between speech recognition errors and certain types of words. We calculated recognition error rates for the words used by the participants, which is defined as the number of times a word was not recognized correctly divided by the total number of times the word was used in voice queries. We only calculated error rates for words being used at least 10 times. Table 3 shows the categorization of the 20 words with the highest recognition error rate.

The first recognizable category of words with high recognition error rates are acronyms, such as "ER" (emergency room, a TV show), "AVP" (the Association of Volleyball Professionals), US and USA. One can hardly expect the system to recognize certain obscure acronyms, such as "ER" and "AVP".

Our interviews showed that more than half of the participants (N=14) reported their concerns about the use of acronyms. When the acronyms were not recognized, they tended to reformulate queries using the full words. For example, participant S14 said that "I was a little concerned … Like how I said AVP, and it pops up APP, which would be a totally different topic. I was a little worried about that … Once I realized what AVP was, I tried to use association, the full name. [sic]". Participant S20 said that "When I did the NRA, instead of giving me a single letter, N-R-A, it spelled out 'in' like that. Then I just switched over to actually saying the National Rifle Association because that was quicker."

Acronyms, named entities and non-English words comprise half of the top 20 words with the highest error rates. Examples of the uncategorized words are also listed in Table 3 as "other words".

5.2 Voice Input Errors in Search Sessions

***RQ4**: How do voice input errors influence search sessions?*

We collected 500 search sessions (20 participants with each working on 25 topics). We divided the 500 sessions into two groups by whether or not voice input errors occurred in the session. As shown in table 4, voice input errors occurred in 187 sessions.

Table 3. Categorization of 20 words with the highest recognition error rates.

Type	Examples (# NOT recognized correctly / # used)
Acronym	ER(29/29), AVP(11/11), US(57/61), USA(6/11)
Named Entity	Owen(25/26), Culpeper(18/27), Ralph(22/36), Gulf(13/24), Falkland(14/27)
Non-English	Nino(31/46)
Other words	theft(14/14), achievement(10/10), taxing(18/21), fraud(12/14), violence(19/27), talk(9/15), sun(24/41), aspirin(23/43), embezzlement(9/18), maglev(8/16)

Table 4. Comparison of session-level statistics between sessions with and without voice input errors.

	187 Sessions w/o Voice Input Errors		**313 Sessions w/ Voice Input Errors**	
	mean	**SD**	**mean**	**SD**
# voice queries	1.44	0.82	**4.41***	2.51
# unique voice queries	1.44	0.82	**3.30***	1.87
# queries w/o voice input errors	1.44	0.82	1.51	1.36
# queries w/ recognition errors	0	0	**2.59***	2.14
# queries w/ system interruptions	0	0	**0.31***	0.65
# unique results by q_v	13.38	6.66	**26.69***	13.90
# unique results by q_{tr}	13.38	6.66	**37.95**$^{*\uparrow}$	21.00
# unique relevant results by q_v	2.90	1.56	3.04	1.59
# unique relevant results by q_{tr}	2.90	1.56	**2.78**$^{\downarrow}$	1.71
# clicked results in the session	1.39	1.01	1.34	1.23
% sessions user clicked results	84.49%	-	69.97%	-
% sessions q_{tr} found relevant results	95.72%	-	92.01%	-

*: the difference between sessions w/ and w/o voice input errors is significant at 0.01 level according to Welch t-test; $^{\uparrow}$ and $^{\downarrow}$: the difference between q_v and q_{tr} is significant at 0.01 level according to paired t-test.

We found that, within the same period of time (a 2-minute search session), the participants issued significantly more voice queries when voice input errors occurred in the search session. As shown in Table 4, the average number of voice queries in sessions with errors was 4.41 and 1.44 without errors (the difference is significant). When voice input errors occurred in the search session, on average 1.11 queries in the session were repeating previously used queries, whereas when no errors occurred, users seldom repeated used queries. After removing the repeated queries, the participants still issued significantly more unique voice queries when voice input error occurred (3.30 vs. 1.44).

One consequence of the increased number of voice queries in sessions with voice input errors was that the participants had to spend more efforts to browse and examine the extra returned results. As showed in Table 4, the unique number of results returned by the transcribed queries in sessions with voice input errors was significantly higher than that of those without voice input errors. Although some of the participants could immediately reformulate the voice query without looking at any results, the increased number of returned results at least would not reduce the participants' search efforts.

We further looked into retrieval effectiveness of search sessions. In sessions with voice input errors, although more results were returned within a session, on average less unique relevant results were actually found. In the 313 sessions with voice input errors, on average the transcribed queries returned only 2.78 unique relevant results within a session. Whereas, if no voice input errors occurred, those sessions' voice queries should result in on average 3.04 relevant results (the difference is significant). Compared with the 313 sessions with voice input errors, the transcribed queries also returned more relevant results in the 187 sessions without any voice input error (2.90 vs. 2.78) and triggered more clicks (1.39 vs. 1.34), but the differences are not statistically significant.

Voice input error also has a higher likelihood of causing a failed search session, in which no relevant result were found. On average, 95.72% of the sessions without voice input errors returned at least one relevant result and in 84.49% of the sessions the participants clicked at least one result. In comparison, when voice input error occurred, only 92.01% of the sessions returned at least one relevant result and in 69.97% of the sessions the participants clicked at least one result.

In addition, voice input errors can also affect the participants' affective feelings. In our interviews, 90% of our participants reported frustration with their search experience when voice input error occurred. For example, participant S15 reported: "It's frustrating! I know I'm saying the word right and I know what I'm looking for, but it's just not connecting, and that disconnection is like arrgh! … (hope I can) just type it. [sic]".

To summarize, our results demonstrated that voice input errors significantly affected the performance of voice queries, and consequently made the whole search process more difficult and less effective. In response, users utilized both lexical and phonetic reformulations to handle the errors, which will be analyzed in the next section.

Table 5. Change in nDCG@10 after query reformulation.

			$q_v^{(2)}$				
			No Error	**Recognition Error**	**System Interruption**	**Query Suggestion**	**All**
			ΔnDCG@10	**ΔnDCG@10**	**ΔnDCG@10**	**ΔnDCG@10**	**ΔnDCG@10**
$q_v^{(1)}$	**No Error**	q_v	0.266 → **0.218** $^{\downarrow}$	0.255 → 0.204	-	-	-
		q_{tr}	0.266 → **0.218** $^{\downarrow}$	0.255 → **0.095** $^{\downarrow}$	0.256 → **0.059** $^{\downarrow}$	0.290 → 0.244	0.262 → **0.164** $^{\downarrow}$
		# cases	209	143	27	15	394
	Recognition Error	q_v	0.248 → 0.248	0.261 → 0.267	-	-	-
		q_{tr}	0.053 → **0.248** $^{\uparrow}$	0.058 → 0.074	0.096 → 0.062	0.099 → 0.226	0.059 → **0.135** $^{\uparrow}$
		Frequency	231	392	44	14	681
	System Interruption	q_{tr}	0.071 → **0.237** $^{\uparrow}$	0.038 → 0.085	0.134 → 0.012	-	0.056 → **0.128** $^{\uparrow}$
		Frequency	30	56	7	0	93
	Query Suggestion	q_{tr}	0.299 → 0.100	0.189 → 0.020	0.235 → 0.000	0.233 → 0.110	0.233 → **0.061** $^{\downarrow}$
		Frequency	4	6	1	3	14
	All	q_{tr}	0.150 → **0.233** $^{\downarrow}$	0.104 → **0.079** $^{\downarrow}$	0.156 → 0.056	0.201 → 0.**223** $^{\uparrow}$	0.129 → **0.143** $^{\uparrow}$
		Frequency	474	597	79	32	1,182

$^{\uparrow}$ and $^{\downarrow}$: the difference between $q_v^{(1)}$ and $q_v^{(2)}$, or between $q_{tr}^{(1)}$ and $q_{tr}^{(2)}$, is significant at 0.01 level according to paired t-tests.

6. VOICE QUERY REFORMULATION

In this section, we focus on users' query reformulations. In the following discussion, we use $q_v^{(1)}$ and $q_{tr}^{(1)}$, $q_v^{(2)}$ and $q_{tr}^{(2)}$ for the voice query and transcribed query both before and after query reformulation, respectively.

6.1 Effectiveness

***RQ5**: Can users' query reformulations improve search performance of voice queries?*

We found that query reformulation in voice search led to overall improvements in performance, but the magnitude depends on whether voice input errors occurred before or after reformulation.

Table 5 shows the comparison of search performance before and after query reformulation when different types of voice input errors occurred in $q_v^{(1)}$ and $q_v^{(2)}$. If counting all 1,182 cases of reformulation, search performance (as measured by nDCG@10) improved significantly from 0.129 to 0.143 (+10.85%) because of query reformulation. However, the improvements mainly occurred in the cases where voice input error occurred in $q_v^{(1)}$ and $q_v^{(2)}$ was correctly recognized, e.g. "Recognition Error" → "No Error" and "System interruption" → "No Error". If no voice input error occurred in $q_v^{(1)}$ or voice input error occurred in $q_v^{(2)}$, query reformulation resulted in limited improvements and it sometimes even hindered search performance.

Since results in Section 5 demonstrated the great influence of voice input errors on search performance, it is not surprising that the effectiveness of query reformulations also largely relied on whether or not voice input errors occurred in $q_v^{(2)}$.

***RQ6**: Can users' query reformulation correct the speech recognition errors in previous queries?*

We found that when recognition error occurred in $q_v^{(1)}$, users' query reformulation corrected some of the missing words in $q_v^{(1)}$. However, at the same time, new voice input errors could also happen in $q_v^{(2)}$, which may counteract the corrected errors and finally lead to degradation in search performance.

Table 6 shows the missing and incorrect words before and after query reformulation for 681 query reformulation cases in which speech recognition errors occurred in $q_v^{(1)}$. We separately calculated the statistics by the different types of queries and voice input errors in $q_v^{(2)}$. As showed in Table 6, when no voice input error occurred in $q_v^{(2)}$ (231 out of 681 cases), it is not surprising that the number of missing and incorrect words both dropped to 0 after query reformulation. When speech recognition errors occurred in $q_v^{(2)}$ (392 out of 681), the number of missing words only dropped slightly from 1.89 to 1.74 (the difference is significant at 0.05 level of significance) and the number of incorrect words slightly increased (the difference is not significant).

Does this mean users' query reformulations can only correct voice input errors when the reformulated queries are correctly recognized? On the contrary, in further analysis, we found that even when speech recognition errors occurred again in $q_v^{(2)}$, users' query reformulation did correct parts of the errors in $q_v^{(1)}$. However, at the same time, new errors also appeared in $q_v^{(2)}$.

To better explain the case, we calculated: the number of missing words in $q_v^{(1)}$ that were correctly recognized in $q_{tr}^{(2)}$; the number of missing words in $q_v^{(1)}$ that were removed in $q_v^{(2)}$; and the number of new missing words in $q_v^{(2)}$ (those are missing words in $q_v^{(2)}$ but not in $q_v^{(1)}$). As shown in Table 6, when speech recognition error occurred in $q_v^{(2)}$, 27.5% (0.52 out of 1.89) of the missing words in $q_v^{(1)}$ were corrected after query reformulation and 18.0% (0.34 out of 1.89) were simply removed. However, on average, 0.72 new missing words were produced in $q_v^{(2)}$, which still impeded the performance.

When system interruption occurred in $q_v^{(2)}$, on average, only 0.23 missing words in $q_v^{(1)}$ were corrected, which is significantly less than the 0.52 missing words corrected in the cases in which speech recognition error occurred in $q_v^{(2)}$.

Table 6. Comparison of the missing and incorrect words before and after query reformulation for the 681 query pairs in which speech recognition error happened in $q_v^{(1)}$.

$q_v^{(2)}$	# missing words	# incorrect words	# missing words in $q_v^{(1)}$ corrected in $q_{tr}^{(2)}$	# missing words in $q_v^{(1)}$ removed in $q_v^{(2)}$	# new missing words in $q_v^{(2)}$
	$q_v^{(1)} \rightarrow q_v^{(2)}$	$q_{tr}^{(1)} \rightarrow q_{tr}^{(2)}$			
No Errors	1.75 → 0.00**	1.81 → 0.00**	1.13	0.61	0.00
Rec Errors	1.89 → 1.74*	1.72 → 1.78	0.52	0.34	0.72
Sys Interrupt	1.71	-	0.23	-	-
Suggestion	1.14	-	0.86	-	-

* and **: the difference of $q_v^{(1)}$ and $q_v^{(2)}$ is significant at 0.05 and 0.01 level.

Table 7. The frequencies of using reformulation patterns.

$q_v^{(1)}$	ADD	SUB	RMV	ORD	Lexical	Lexical & Phonetic
No Errors	90.50 %	15.04 %	66.75 %	33.51 %	99.74 %	0.26 %
Rec Errors	32.98 %	16.34 %	37.93 %	43.03 %	77.36 %	11.99 %
Overall	53.82 %	14.87 %	48.37 %	39.58 %	85.47 %	7.74 %
$q_v^{(1)}$	**STR/ SLW**	**SPL**	**DIF**	**WE**	**Phonetic**	**Repeat w/o PE or WE**
No Errors	0 %	0 %	0 %	0.26 %	0.26 %	0 %
Rec Errors	14.84 %	0.60 %	0.90 %	9.30 %	25.64 %	20.54 %
Overall	9.46 %	0.39 %	0.57 %	6.02 %	16.44 %	13.58 %

6.2 Use of Reformulation Patterns

***RQ7**: How do users utilize different query reformulate patterns in voice search? Do voice input errors influence the use of query reformulation patterns?*

Table 7 shows the frequency of using different reformulation patterns in voice search. Despite how the query input mechanism changes dramatically in voice search, lexical reformulations were still the primary forms of query reformulation. No matter if speech recognition errors occurred, lexical reformulations were consistently used much more frequently than phonetic reformulations.

However, speech recognition errors did significantly affect the use of specific lexical query reformulation patterns. When speech recognition errors occurred, the participants tended to reformulate queries using more substitution (SUB) and re-ordering (ORD) patterns but dramatically less addition (ADD) and removal (RMV) patterns. As further examined in RQ8, this is probably because substitution and re-ordering can effectively correct the missing words in previous queries, whereas addition and removal cannot.

The use of phonetic reformulation patterns is almost always associated with speech recognition errors. As shown in Table 7, when no voice input error occurred in $q_v^{(1)}$, only 0.26% of the query reformulations adopted phonetic reformulation patterns. In comparison, 25.64% of the query reformulations adopted phonetic reformulation patterns when speech recognition errors happened in $q_v^{(1)}$. In addition to the phonetic reformulation patterns, repeating is also closely connected with speech recognition errors. When speech recognition errors occurred in $q_v^{(1)}$, we found that 20.54% of the reformulations were simply repeating $q_v^{(1)}$ without any recognizable phonetic changes.

Among all of the phonetic reformulation patterns, partial emphasis (PE) was used more frequently than whole emphasis (WE). As we mentioned in Section 4, stressing (STR) and slowing down (SLW) were the most frequent patterns for partial emphasis, while spelling (SPL) and using different pronunciations (DIF) rarely happened. Repeating was used as frequently as phonetic reformulation patterns when recognition errors happened in $q_v^{(1)}$.

To conclude, our results indicate that in voice search, a user's adoption of both lexical and phonetic query reformulation patterns were greatly impacted by voice input errors. As further illustrated in RQ8, many of the reformulation patterns were used specifically to correct the missing words occurred in previous queries.

***RQ8**: How do users utilize different reformulation patterns to handle speech recognition errors? Are these patterns effective in correcting speech recognition errors?*

When speech recognition errors happen, it is very common for some of the words spoken by the users to be incorrectly recognized or missing from the system's transcribed queries. Solutions to speech recognition errors should be able to effectively correct these errors. Among the lexical and phonetic query reformulation patterns summarized in our paper, four patterns can be used specifically related to the missing words: substitution (SUB), removal (RMV), re-ordering (ORD), and partial emphasis (PE). Users can substitute other words for the missing words, or remove the missing words, or re-order the missing words and other words, or phonetically emphasize the missing words. In comparison, the other patterns affect equally the missing words and other words in the query.

We evaluate the reformulation patterns by their effectiveness of correcting the missing words in voice queries. Similarly, we can evaluate by their effectiveness of reducing the incorrect words in transcribed queries. However, due to space limitation, we only reported the following measures regarding the missing words:

(1) For each of the four patterns that can be used specifically for handling the missing words (i.e. SUB, RMV, ORD, and PE), we calculated the percentage that the pattern was used specifically related to the missing words (i.e. the missing words were substituted, removed, re-ordered, or emphasized) out of all the cases that the reformulation pattern was used.

(2) The success rate of each pattern in correcting the missing words. For re-ordering (ORD) and partial emphasis (PE) patterns, the success rate was calculated as the percentage of missing words being corrected out of all the cases that the missing words were re-ordered or specifically emphasized. For addition (ADD), whole emphasis (WE), and repeating patterns, the success rate was calculated as the percentage of missing words being corrected out of all the cases that ADD, WE, or repeating was used (since it is difficult to identify whether these patterns were used specifically on the missing words). For substitution, the success rate was calculated as the percentage of the replaced words being correctly recognized out of all the cases that the missing words were replaced.

(3) The improvement in nDCG@10 between $q_{tr}^{(1)}$ and $q_{tr}^{(2)}$ when each pattern was used.

As shown in Table 8, the percentage of the patterns used specifically related to the missing words indicates users' adoption of the pattern to solve speech recognition errors. Among all of the patterns, partial emphasis (PE) has most usage. When PE was used, it was nearly always (93.69%) the case that the words emphasized were the missing words from $q_v^{(1)}$. In comparison, substitution (SUB), removal (RMV), and re-ordering (ORD) patterns have fewer but still considerably high usage (84.30%, 62.82% and 75.23%). Results indicate that, when recognition errors happened, these lexical patterns were primarily used to correct speech recognition errors, which is different from the intention to use these patterns in conventional searches.

Table 8 also reveals the effectiveness of different reformulation patterns in correcting speech recognition errors. As indicated in the results, different reformulation patterns vary widely in their success rates in correcting missing words in previous queries. Among these patterns, substitution (SUB) and re-ordering (ORD) had the two highest success rates (73.5% and 69.1%). In comparison, partial emphasis (PE) was less effective (62.5%). It is indicated that when recognition errors happened, it was usually more effective to modify the missing words into others (SUB) or to change the contexts around the missing words (ORD), rather than emphasizing with phonetic changes (PE).

Table 8. Effectiveness of reformulation patterns in correcting speech recognition errors that occurred in previous queries.

	% used specifically related to the missing words in $q_v^{(1)}$	Success rate of correcting missing words	nDCG@10 $q_{tr}^{(1)} \rightarrow q_{tr}^{(2)}$
ADD	-	40.73 %	0.085 → 0.119
SUB	84.30 %	73.53 %	0.052 → **0.156** $^{\uparrow}$
RMV	62.82 %	-	0.077 → 0.111
ORD	75.23 %	69.14 %	0.062 → **0.147** $^{\uparrow}$
PE	93.69 %	62.50 %	0.022 → **0.150** $^{\uparrow}$
WE	-	60.94 %	0.028 → **0.110** $^{\uparrow}$
Repeat w/o PE and WE	-	59.73 %	0.051 → **0.142** $^{\uparrow}$
Overall	-	47.45 %	0.058 → **0.132** $^{\uparrow}$

$^{\uparrow}$: the difference of nDCG@10 is significant at 0.01 level according to paired t-tests.

We suspect that users' adoption of partial emphasis (PE) is directly related to their everyday life experience: when others miss your words, it is natural to repeat and emphasize the missing part. However, it seems that this method cannot work well for automatic speech recognition systems. The speech recognition algorithms are usually trained with samples of the normal way of speaking, but the phonetic query reformulations may make the queries quite different from the normal way of speaking.

According to the success rates, partial emphasis (PE), whole emphasis (WE), and repeating effectively helped to correct the missing words (compared to the overall success rate of only 47.45%). However, we suspect that the effectiveness of the phonetic reformulation patterns is over-estimated. Compared with repeating, the phonetic patterns emphasized either certain parts of the queries or the entire queries. Therefore, we can use repeating as a baseline to evaluate the effectiveness of phonetic emphasis. However, as partial emphasis (PE) and whole emphasis (WE) had only slightly higher success rates compared to repeating, it is arguable whether or not phonetic emphasis was truly useful.

Finally, we looked into the improvement of the transcribed queries' search performance (by nDCG@10) after each pattern had been used in reformulated queries. Except for addition (ADD) and removal (RMV), we observed significant improvements with other patterns. In addition, the magnitude of nDCG@10 improvements for other patterns was also greater than those of ADD and RMV patterns. This indicates that ADD and RMV are less effective solutions to speech recognition errors.

To conclude, we found that substitution, re-ordering, partial emphasis, whole emphasis, and repeating were five effective reformulation strategies in voice search to handle recognition errors. Among these patterns, substitution and re-ordering are lexical patterns, but they outperformed the other three phonetic patterns in solving speech recognition errors.

7. DISCUSSION AND FUTURE WORK

(1) Should we use and support long and natural language queries or short and keyword queries in voice search?

Our results show that query length is an important factor associated with speech recognition errors (see Table 2 and discussion in RQ3). Long queries are prone to speech recognition errors. This reminds us of the different findings in previous studies: Schalkwyk et al. found that voice search queries were tend to be shorter than in conventional searches [19], whereas Crestani et al. found that voice queries tend to longer and more similar to natural language [6].

Since we did not conduct conventional search experiments for comparison, we cannot come to an answer to this disputable issue. We suggest that further studies are needed to identify the characteristics of queries in voice search. We believe that users' adoption of short or long queries depends on various factors. On the one hand, as voice search may be closer to people's normal ways of speaking, voice queries are probably also closer to natural language queries. On the other hand, as long queries may have more speech recognition errors, users may also prefer shorter and simpler keyword queries in voice search.

(2) Query suggestion in voice search.

Although the participants were told explicitly that they could use Google's query suggestions in our experiment, we did not observe many cases of them doing so (see Table 5). We tried some cases in Google and found that currently, Google's query suggestion in voice search is simply suggesting queries based on the transcribed queries' texts. Therefore, it is not surprising that the suggestions are ineffective when the transcribed texts are likely to be incorrect (due to voice input errors). For example, we submitted an incorrect transcription "rap and crying" (the correct one is "rap and crime") to Google and obtained two suggestions that are irrelevant to "rap and crime" but probably relevant to "rap and crying": "rapper crying at bet awards" and "soulja boy crying". This shows that query suggestion is more challenging in voice search.

In addition, we believe that query suggestion is more important for users in voice search than in conventional search. As shown in our results, despite various query reformulation methods have been developed, users' voice query reformulations might not totally resolve the old recognition errors, and at the same time could introduce new errors. In comparison, it may be a better solution for users to accept a good query suggestion for query reformulation. This calls for studies on query suggestion algorithms specifically designed for voice search. Probably a promising solution is to develop effective query suggestion algorithms considering not only the transcribed texts, but also speech recognition results.

(3) Interface for supporting voice query inputs and voice query reformulation.

Considering the effort and risk of issuing a voice query, voice search systems should employ proper methods to reduce the efforts and risks of constructing and reformulating voice queries. Based on our observation, one suggestion is to design a voice query reformulation interface that frees users from having to speak the whole voice query again if they only intend to correct one or two error words. For example, the users should be given the ability to specify and repeat the part of the query that they want to modify and let the search system recompose a new voice query based on the updated information.

In addition, our experiments also shown that system interruptions greatly harmed the performance of voice search, even though they occurred less frequently (see Table 2 & 5). The participants could not finish their voice queries, and sometimes became really frustrated after several consecutive interruptions. Voice query generation may impose higher cognitive load on the users than typing textual queries. Therefore, voice search systems should better manage their interruptions. For example, systems can allow users to control whether or not they will be interrupted while speaking voice queries.

8. CONCLUSION

In this paper, we studied two significant and closely related issues in voice search. First, what is the influence of voice input errors on search effectiveness in voice search? Second, how do users utilize different query reformulation patterns, including both lexical and phonetic query reformulation patterns, to handle these voice input errors? We conducted a controlled laboratory experiment for voice search, which helped answer these questions.

Our study systematically evaluated the influence of voice input errors on voice search from the aspects of individual queries and overall search sessions. We found that voice input errors greatly changed the content and results of queries, resulting significant decline of search performance for individual queries. This in turn led to increased efforts and negative feelings of users, hindering overall performance of the search session. In addition, current query suggestion algorithms may fail to generate effective suggestions due to voice input errors in transcribed queries.

Then, we characterized users' query reformulation patterns in voice search and evaluated the effectiveness of those patterns in handling voice input errors and improving search effectiveness. We found that users utilized both lexical query reformulation patterns that exist in conventional search and phonetic query reformulation patterns newly found in voice search. Despite some of the patterns effectively corrected voice input errors, users' query reformulation resulted in limited overall improvements in search performance, because voice input errors occurred frequently in reformulated queries.

Our study suggested voice input errors as the essential issue to be resolved in voice search. A possible solution is to better support users' query reformulation, which includes designing better interface supporting voice query reformulation and developing query suggestion algorithms using both lexical and phonetic information. To a broader extent, our study explored the influence of query input devices on user behaviors and search systems. Our methods and results may shed light on user behaviors and search systems in similar situations, such as when handwriting is used for input.

Admittedly, our study has one limitation in that the experiment setting did not fully replicate mobile search environment and tasks. This may influence the occurrences of the different types of voice input errors and users' adoption of the voice query reformulation patterns. However, it is very likely that the impacts of voice input errors on voice search systems and the effectiveness of different voice query reformulation patterns are representative of the cases in other voice search systems.

9. REFERENCES

[1] Anick, P. 2003. Using terminological feedback for web search refinement: a log-based study. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR '03): 88-95.

[2] Ballinger, B. et al. 2010. On-Demand Language Model Interpolation for Mobile Speech Input. Interspeech (2010): 1812–1815.

[3] Bates, M.J. 1979. Information search tactics. Journal of the American Society for Information Science, 30(4): 205–214.

[4] Broder, A. 2002. A taxonomy of web search. SIGIR Forum 36(2): 3-10.

[5] Crestani, F. 2002. Spoken query processing for interactive information retrieval. Data Knowl. Eng. 41, 1 (April 2002): 105-124.

[6] Crestani, F. et al. 2006. Written versus spoken queries: A qualitative and quantitative comparative analysis. J. Am. Soc. Inf. Sci., 57: 881–890.

[7] Dang, V. and Croft, W.B. 2010. Query reformulation using anchor text. In Proceedings of the third ACM international conference on Web search and data mining (WSDM '10): 41-50.

[8] Feng, J. and Bangalore, S. 2009. Effects of word confusion networks on voice search. (Mar. 2009): 238–245.

[9] Huang, J. and Efthimiadis, E. N. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09): 77-86.

[10] Jansen, B.J. et al. 2005. A temporal comparison of AltaVista Web searching. J. Am. Soc. Inf. Sci., 56(6): 559–570.

[11] Jansen, B.J. et al. 2009. Patterns of query reformulation during Web searching. J. Am. Soc. Inf. Sci., 60(7): 1358–1371.

[12] Järvelin, K. et al. 2008. Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. LNCS 4956: Proceedings of the 30th European Conference on Information Retrieval (ECIR '08): 4–15.

[13] Jiang, J. et al. 2012. Contextual evaluation of query reformulations in a search session by user simulation. In Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12): 2635-2638.

[14] Jiang, J. et al. 2012. On Duplicate Results in a Search Session. Proceedings of the 21st Text REtrieval Conference, (TREC 2012).

[15] Joachims, T. 2002. Optimizing search engines using clickthrough data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02): 133-142

[16] Kanoulas, E. et al. 2011. Evaluating multi-query sessions. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11): 1053–1062.

[17] Kanoulas, E. et al. 2011. Session Track 2011 Overview. The 20th Text REtrieval Conference Notebook Proceedings (TREC 2011).

[18] Rieh, S.Y. et al. 2006. Analysis of multiple query reformulations on the web: The interactive information retrieval context. Information Processing & Management. 42(3): 751–768.

[19] Schalkwyk, J. et al. 2010. "Your Word is my Command": Google Search by Voice: A Case Study. Advances in Speech Recognition SE - 4. A. Neustein, ed. Springer US. 61–90.

[20] Song, Y.-I. et al. 2009. Voice search of structured media data. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (Apr. 2009): 3941–3944.

[21] Teevan, J. et al. 2007. Information re-retrieval: repeat queries in Yahoo's logs. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07): 151-158.

[22] Wang, X. et al. 2008. Mining term association patterns from search logs for effective query reformulation. In Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08): 479-488.

[23] Wang, Y.-Y. et al. 2008. An introduction to voice search. Signal Processing Magazine, IEEE.