

Comparing Isolated Word and Continuous Speech for Voice Search

Submission #307

ABSTRACT

Continuous, natural speech input to voice user interfaces enables a conversational interaction, but requires sophisticated speech recognition. Recognizing isolated keywords requires significantly less training data and computational resources. This paper assesses the usability impact of restricting input to isolated words in a voice-only search interface. To do so, we designed, implemented, and evaluated continuous speech and isolated word versions of *Farmers' Exchange*, a voice interface for searching for answers to agricultural questions. Task completion rates were comparable between the two conditions, but continuous speech was significantly faster. The difference was correlated to longer, more specific initial queries given more often with continuous speech input. However, more input wasn't always better; there was an "optimal" query length for Farmers' Exchange. Though a slower interface overall, isolated word was significantly faster for participants with extensive prior experience with search interfaces, suggesting that a longitudinal comparison of the two modes would be a worthwhile future direction.

ACM Classification Keywords

H.5.2 User Interfaces: Voice I/O User Interfaces; H.5.2 User Interfaces: Evaluation

INTRODUCTION

When the computing platform is a mobile phone, voice is a natural choice as an interaction modality [1]. However, "the bane of speech driven interfaces is the very tool which makes them possible: speech recognition." [2].

This paper compares the usability of isolated word to continuous speech input in a voice-only search interface. We are motivated to evaluate a restricted input style that leads to higher accuracy speech recognition, particularly for languages with limited training resources. We hypothesized that, like text-based search, speech queries will naturally tend toward a small number of keywords, making a re-

stricted input system comparable to an unrestricted system in performance and satisfaction. On the other hand, it is possible that the constraints of isolated word input (or, conversely, the freedom of continuous speech) would result in a difference in performance and/or satisfaction.

We evaluated the two styles using *Farmers' Exchange*, an English language voice interface for searching a repository of agricultural questions and answers. In our experiment, participants had comparable task completion rates in the two conditions, but session duration was significantly shorter with continuous speech. The difference was correlated to longer, more specific initial queries given more often with continuous speech. However, more input wasn't always better; empty result sets from over-specification were also more frequent. Isolated word was significantly faster for participants with extensive prior experience with search interfaces compared to those with less experience. This motivates a longer-term comparison of the two modes as a next step for this work.

BACKGROUND AND RELATED WORK

Commercial speech systems achieve recognition rates of 95% or more [3-4]. Successful speech recognition for large vocabulary, continuous speech requires three linguistic resources: first, a large corpus of speech hand-labeled at the phoneme or word level; second, a dictionary of words in the language with pronunciations for each word; and third, a language model trained on a large text corpus, or grammar rules created by a linguist. These linguistic resources are readily available in English, French, and Japanese, for example. By contrast nearly all of the world's languages currently have limited or no linguistic resources. The time, money, and expertise required to develop these resources represents a practical barrier for enabling large vocabulary, continuous speech recognition across many languages [5].

Isolated-word recognition requires significantly less transcribed data than continuous speech, and no language model. Comparing estimates of the training data required for large vocabulary, continuous speech recognizers [3, 6] to approaches that simplify the recognition task to hundreds of isolated words [7] (assuming high accuracy for both), the difference can be an order of magnitude.

Prior research comparing speech input modes has produced varying results depending on the task being studied. In a study comparing continuous, isolated word, and touchtone input for entering digit sequences, user preference favored

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2009, April 4-9, 2009, Boston, Massachusetts, USA.
Copyright 2009 ACM 978-1-60558-246-7/09/04...\$5.00.

continuous input and touchtone over isolated word [8]. In a comparison of touchtone to flexible dialog input on a voicemail management application, there were no significant differences in completion time, task completion, or user preference [9].

In the domain of text-based Web search, research has examined query formulation and specification (see [10] for a list, or [11] for a survey). Three observations come up repeatedly: most people use few search terms; infrequently view pages and iterate queries; and rarely use advanced search features. Studies of web search conducted between 1997 and 2005 report an average query length ranging from 2.4 to 2.8. While there is a slight upward trend, 75% of queries remain three words or less [12]. This note investigates search behavior in a voice interface.

FARMERS' EXCHANGE

The platform for this study, *Farmers Exchange*, is a voice search interface for agricultural information. Our fieldwork suggests that effective, telephone-based information access holds great appeal for farmers. A voice-based interface offers access in outdoor work environments, using technology many farmers use already. If translated into Spanish, Hmong, Chinese, and other languages, Farmers Exchange could also address the communication gap between local support institutions and non-English speaking farmers.

We developed Farmers Exchange in partnership with the National Sustainable Agriculture Information Service (ATTRA) [13]. The Farmers' Exchange database has over 300 English questions and answers excerpted from a larger repository maintained by ATTRA. We integrated Nuance's Dragon Naturally Speaking [4] speech recognizer to interpret speech queries. The system returns matches based on a full text search of the question and answer corpus.

During the parsing process, the dialog engine removes all ancillary words (such as "I", "wonder", or "about") from the user's input, leaving only content keywords. It then speaks these back to the user to verify that they were recognized correctly. This interactive feedback helps maintain common ground.

Farmers Exchange searches its database with a logical conjunction of the query's keywords. The system responds in one of three ways. If the search produced three or fewer matches, those are presented to the user for listening. If there are more than three results, Farmers' Exchange prompts the user to provide further input to narrow them to three or fewer. Three results was chosen as a threshold based on previous research on optimal branching factors for phone trees [14]. If the user does not respond within a specified time window, the system proactively suggests additional keywords. When the result set has been sufficiently narrowed, the caller may listen to the questions and answers. If the search produced no results, the system asks the user to go back a step in the refinement process (if possible), start

FE: Welcome to Farmer's Exchange! Please say your question, or hold for suggestions.

User: I'm looking for information on goat dewormers.

FE: You asked about goat; and dewormers.... is that correct?

User: Yes.

FE: I found 4 results that match your query. Please provide more detail to narrow your results, or hold for some suggestions.

User: <no input>

FE: Okay, here are some suggestions: drugs; pasture; or, poultry.

User: Drugs.

FE: You asked about drugs... is that correct?

User: Yes.

FE: Okay, I found 3 results that match your query. The first result is 'What information can you provide on organic dewormers for meat goats?'. Would you like to play the answer, go to the next question, or go back?

User: Play.

FE: Ok, here's the answer...

Figure 1 A sample interaction with Farmers' Exchange. The isolated word version of the system only differs in the initial prompt, which asks for keywords separated by pauses.

their search over, or submit their question to agricultural specialists (like ATTRA) by recording it directly. Figure 1 shows a sample dialog.

PARTICIPANTS AND METHODOLOGY

Fourteen farmers participated in this experiment; all spoke English. Farming experience varied from 2 years to 54 years, with a median of 22 years. The study was conducted in participants' homes, typically in a quiet room. Sessions were audio and video recorded. System logs recorded recognized input, keywords identified, hierarchy options selected, and navigation within results.

The experiment had a within-subjects design. Before interacting with each treatment, users were given a tutorial of the system via a hypothetical dialog between the system and a user acted out by two researchers. Treatment and task order were randomly assigned to each participant. For each treatment, the participant was given 3 tasks to complete, one at a time. The experimenter presented each task orally, reading from a script. Task descriptions consisted of some general information to set up the scenario, followed by a specific information retrieval request.

"You are a strawberry producer in California and are having weed problems in your rows. You are interested in becoming organic, so you don't want to use chemicals. Come up with two weed control options for organic strawberry production."

Participants were told that the system may not contain the answers for completing all tasks, though in actuality the specific information needed for all 6 tasks was present. Users 'finished' a task when they were satisfied they had found

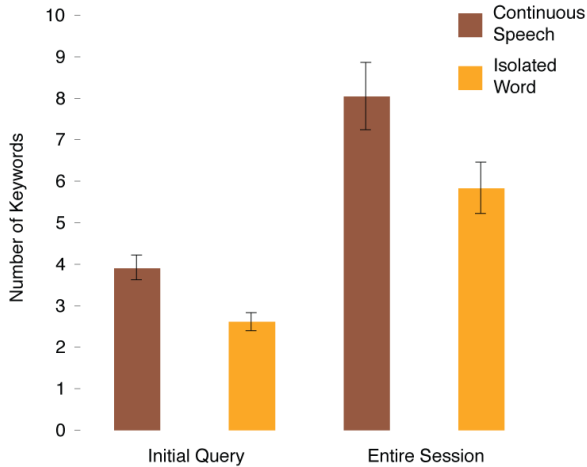


Figure 2: Continuous speech elicited more keywords than isolated word input.

the correct information, had submitted their question, or had given up.

RESULTS

Our 14 participants completed a total of 84 required tasks, as well as 9 additional tasks asking their own questions. We discarded two of the required sessions due to glitches in our prototype. In the following sections, we include data from the 91 remaining sessions, except where noted.

There was no significant difference in task completion rate between the two treatments. 70% of the required isolated word sessions were successful, compared to 67% for continuous speech.

Despite comparable completion rates, the time it took to complete those tasks varied significantly between the two conditions. For continuous speech, average session duration was 102.2 seconds per session, while with isolated word it was 131.6 seconds ($p < 0.03$). The number of *exchanges* (user input followed by a system output) was 8.3 for continuous speech versus 10.9 for isolated word ($p < 0.05$).

The difference in duration was correlated to the number of keywords users provided in their initial queries. On average, continuous speech sessions yielded 49% more keywords in the initial query than with isolated word (see Figure 2). The additional keywords led to a 20% reduction in the number of exchanges required to obtain an answer. In 48% of continuous sessions, participants were able to reach a correct answer after just one input (*i.e.*, no further refinements were required). In contrast, only 33% of isolated word sessions succeeded after a single input ($p < 0.04$).

Users who provided 4 keywords in their initial query reached an answer the fastest (see Figure 3). Overall, the total number of keywords provided by the *end* of a successful session, including refinement steps, was 4.9. This suggests

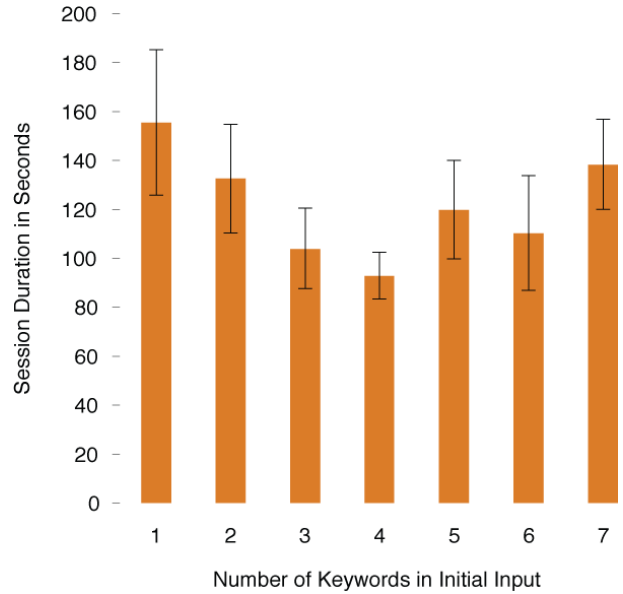


Figure 3: Users who initially provided four keywords reached an answer the fastest.

that the optimal number of keywords needed for obtaining an answer in Farmers' Exchange was about 5.

Isolated word users usually provided too few keywords; on average, 2.7 keywords in the initial query. As a consequence, they had to *iterate* on their search (on average, twice) by providing more keywords. In contrast, continuous speech users said significantly more keywords in their initial queries, with an average of 4 ($p < 0.001$). However, verbosity also worked against them; continuous speech users often received no results because their queries were too specific. 61% of continuous speech sessions resulted in at least one empty result set from the system, compared to only 30% of isolated word sessions ($p < 0.03$). Consequently, continuous speech users restarted more often than isolated word users: 24% of continuous speech sessions included at least one restart, versus 10% for isolated word sessions ($p < 0.03$).

Despite differences in the *number* of keywords uttered in the two conditions, *which* keywords were uttered was not affected. Of the 172 unique keywords uttered overall, 48 were common to both treatments. 82 of the remaining keywords were unique to the continuous speech treatment (63% of all continuous speech words), and 42 unique to isolated word (47% of isolated word words). This difference appeared to be largely the result of population variability; a test of variation within each treatment showed similar percentages when sample size was held constant.

In a series of Likert-scale questions, participants showed a significant overall preference for continuous speech over isolated word input. Users perceived continuous speech as being more pleasant, fun, effective, and easy to use than isolated word ($p < 0.03$). Notably, some users preferred conti-

nuous speech even though they completed more tasks successfully using isolated word.

While continuous speech sessions had a faster completion time on average, the very fastest completion times occurred with isolated word input. Of the ten fastest sessions, half were with isolated word, including the top three. These sessions were with users with the most experience using the Internet (and presumably search engines). As a group, they performed tasks significantly faster than less experienced users, and they also completed significantly more tasks with both interfaces. These results indicate that experience can make up for efficiency lost through restrictive isolated word.

DISCUSSION

Our main motivation for comparing continuous speech to isolated word input was to determine the usability and performance tradeoffs of simplifying the speech recognition task for voice search interfaces. The results showed that constraining the input to isolated word did not affect the ability to successfully retrieve search results, but it did have a negative impact on the speed of completing tasks, which could have a detrimental effect on user satisfaction. There was a strong positive correlation between a user's preferred input condition and the condition they completed more quickly ($r=0.4$).

On the other hand, results showed that in the hands of experienced web searchers, isolated word input can yield faster task completion times, owing to the brevity of isolated word input and the ability for experienced users to provide more, higher quality keywords than novices. The impact of expertise on search behavior in the text-based search engine has been well-reported [10-11, 15]. Our study in the voice search domain confirms the prior result that search experience determines query quality, which ultimately determines their success.

Given the potential of isolated word input as a superior input mode for experienced searchers, an interesting next step to this experiment would be to study search behavior with users over an extended period of time. If an isolated word search interface does not frustrate novice users early on, the long-term performance could be comparable to an unconstrained system.

CONCLUSION

In this study we compared continuous speech to isolated word input for a voice search interface. We found that there was no difference in task completion between the two systems overall. However, continuous speech input led to users specifying more search terms in their initial query, which correlated to significantly faster searches. Isolated word input yielded significantly shorter queries and slightly fewer unique keywords. However, isolated word session duration was significantly shorter for participants with extensive prior experience with web search interfaces. This motivates

a follow-up study of search behavior over a longer time period.

REFERENCES

- [1] Lai, J. Conversational interfaces. *Commun. ACM*, 43, 9 (2000), 24-27.
- [2] Nicole Yankelovich, G.-A. L., Matt Marx. Designing SpeechActs: Issues in Speech User Interfaces. In *Proceedings of the CHI '95 Proceedings, Conference on Human Factors in Computing Systems* (Denver, CO, May 7-11, 1995, 1995), [insert City of Publication],[insert 1995 of Publication].
- [3] Dan Jurafsky, J. H. M. *Speech and Language Processing*, 2009.
- [4] Nuance *Dragon Naturally Speaking product page*. City, 2009.
- [5] Madelaine Plauché, U. N. Speech Interfaces for Equitable Access to Information Technology. *Information Technologies and International Development*, 4(2007).
- [6] Frederick Weber, K. B., Roni Rosenfeld, Kentaro Toyama. Unexplored Directions in Spoken Language Technology for Development. In *Proceedings of the Spoken Language Technology Workshop* (2008), [insert City of Publication],[insert 2008 of Publication].
- [7] Madelaine Plauché, Ö. Ç., Udhaykumar Nallasamy. How to Build a Spoken Dialog System with Limited (or no) Language Resources. In *Proceedings of the IJCAI Workshop on AI in ICT for Development* (2006), [insert City of Publication],[insert 2006 of Publication].
- [8] J. Foster, F. M., M. Jack, S. Love, R. Dutton, I. Nairn. An experimental evaluation of preference for data entry method in automated telephone services. *Behavior and Information Technology* (1998).
- [9] M. Goldstein, I. B., E. L. Sallnas, H. Bjork. Navigational abilities in voice-controlled dialogue structures. *Behavior and Information Technology* (1999).
- [10] Wendy Lucas, H. T. Form and Function: The Impact of Query Term and Operator Usage on Web Search Results. *Journal of the American Society for Information Science and Technology* (2002).
- [11] Hsieh-Yee, I. Research on Web search behavior. *Library & Information Science Research* (2001).
- [12] Hearst, M. *Search User Interfaces*. Cambridge University Press, 2009.
- [13] Service, N. S. A. I. *ATTRA website*. City, 2009.
- [14] Halstead-Nussloch, R. The Design of Phone-based Interfaces for Consumers. In *Proceedings of the CHI* (1989), [insert City of Publication],[insert 1989 of Publication].
- [15] C. Hoelscher, G. S. Web search behavior of Internet experts and newbies. In *Proceedings of the Proceedings of the Ninth International World Wide Web Conference* (2000), [insert City of Publication],[insert 2000 of Publication].