Speech for content creation

Joseph Polifroni
Nokia Research Center
4 Cambridge Center
Cambridge, MA 02142 USA
joseph.polifroni@nokia.com

Imre Kiss
Nokia Research Center
4 Cambridge Center
Cambridge, MA 02142 USA
imre.1.kiss@nokia.com

Stephanie Seneff
MIT CSAIL
32 Vassar Street
Cambridge, MA 02139 USA
seneff@csail.mit.edu

ABSTRACT

In this position paper, we propose a different paradigm for using speech to interact with computers: speech for content creation. We survey the literature in automatic speech recognition (ASR), natural language processing (NLP), sentiment detection, and opinion mining to argue that the time has come to use mobile devices to create content on-the-fly. We examine recent work in user modelling and recommender systems to support our claim that using speech in this way can result in a useful interface to uniquely personalizable data. We describe a data collection effort we've recently undertaken to help us build a prototype system for spoken restaurant reviews. This vision critically depends on mobile technology, for enabling the creation of the content and for providing ancillary data to make its processing more relevant to individual users. We feel this type of system can be of use even where only limited speech processing is possible.

General Terms

speech applications, content creation, sentiment detection, user modelling

1. INTRODUCTION

A couple are visiting Toronto and have just finished a meal at a small Chinese restaurant. The wife makes a habit of scouting out Chinese food in any city she visits and this restaurant was particularly good. As she walks out of the restaurant, she pulls out her mobile phone, clicks a button on the side, and speaks her thoughts about the meal she's just eaten. She then puts her phone away, having recorded her speech, her location, and the time of day, hails a cab, and goes off to the theater. Figure 1 shows what a user might say in this context.

The scenario we describe above is the first-stage interaction with an overall system that uses speech for content creation, social media, and recommender systems. In subsequent sections, we will enlarge upon this scenario, with further glimpses into the user interaction and the underlying technology required for each step. We argue that these technologies are sufficiently advanced to enable the convenience of recording thoughts and impressions on the go, indexing the results, and extracting enough information to make it useful for others.

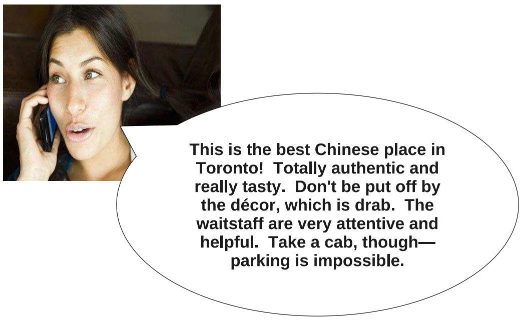

This is the best Chinese place in Toronto! Totally authentic and really tasty. Don't be put off by the décor, which is drab. The waitstaff are very attentive and helpful. Take a cab, though—parking is impossible.

Figure 1: A representation of how a user might create content via speech.

One of the most important aspects of this scenario, and the ones that follow, is that the user is in charge of the interaction the entire time. Users do not have to worry about getting involved in an interaction when they're busy, in a noisy environment, or otherwise unable to devote time to the interface. Users can describe an experience while it is fresh in their memory through an interface that is always available to them. When they have the time and the inclination to make further use of the information, they can examine, review, and, ultimately, share it. The spoken input takes the form of a "note to self," where the user does not have to plan carefully what to say.

In this initial scenario, the user's interaction with the system stops after the review is spoken. Either immediately, or when connectivity is re-established, speech is uploaded to a cloud-based system. With a combination of automatic speech recognition (ASR) and natural language processing (NLP) technologies, the system goes to work on indexing and deriving meaning from the dictated review. In the best case scenario, information about individual features, such as *food quality* or *service*, are extracted and assigned a scalar value based on user input. These values are used to populate a form, combined with other online sources of information,

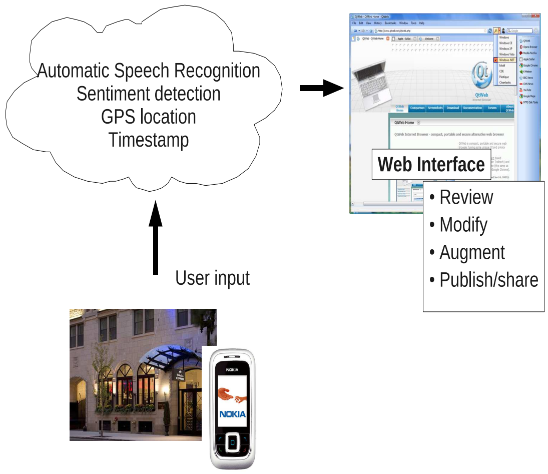


Figure 2: A schematic representation of data capture and processing in the restaurant review scenario.

and made available to the user to review, modify, and share (see Section 2). Various other fallback levels of analysis are always available, so that the information is never completely lost or ineffectual. For example, the system may be able to only assign a single overall polarity to the entire review, or just extract some keywords for indexing. In the worst case, a simple audio file is saved and associated with a timestamp and GPS location. The user remains unaware of this processing, which need not be real-time. Further input will come later, at the discretion of the user. Figure 2 shows how this process might unfold.

Speech for content creation has several characteristics that make it attractive from a technological perspective:

- It does not have to be real-time. As our scenarios illustrate, the user simply speaks to a mobile device to enter content. Any further interaction takes place at the convenience of the user.
- It does not involve a detailed word-by-word analysis of the input. As we will show, further processing of the text can be done using just keywords/phrases in the user's input.
- It can be designed with multiple fallback mechanisms, such that any step of the process can be perceived as useful and beneficial to the user.

The key components that make this vision possible are large-scale ASR; NLP for keyword/named entity extraction, as well as opinion mining and sentiment detection; and content creation and information presentation informed by user modelling and research in recommender systems.

In Section 2, we describe the current state of the art in the technologies to be used for acquiring content via speech. Section 3 describes recent research in recommender systems and user modelling and shows how content collected via a mobile device fits into the emerging paradigms in these fields.

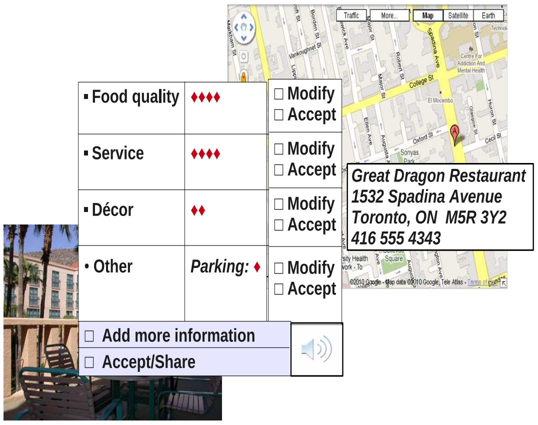


Figure 3: A representation of the user interface for reviewing, accepting, and sharing content.

Section 4 describes a data collection effort that is currently underway at Nokia Research to help us build a system for spoken restaurant reviews. In Section 5, we discuss how this technology could be used in developing markets.

2. ACQUIRING CONTENT VIA SPEECH

It is the following morning. The visitor to Toronto from Section 1 has had her morning coffee and is pleasantly thinking back on the previous evening. She goes to a website and sees a map of Toronto with an icon at the location of the Chinese restaurant she went to the day before. The system has inferred the name of the restaurant from positioning data and has added further information from online restaurant review sites, including address, phone number, and summaries of what other people have said about the place. The user clicks on the icon and sees a display of what the system has done with the input (see Figure 3). She looks it over, makes a small change, and decides to add a recommendation (via speech or typing) for a dish she particularly liked. After spending a few minutes on the review, she clicks **Share** *and makes it available to her friends.*

2.1 Extracting meaning from speech

The technologies described in this section demonstrate that extracting meaning from spontaneous speech is possible, and does not necessarily involve a complete analysis of the input utterance(s). As we describe below, there is value in simply extracting keywords and phrases from speech data. More sophisticated analysis for sentiment detection, although currently only applied to text data, also involves processing only parts of the text.

Keyword/phrase spotting has long been used to perform at least *partial understanding* of spontaneous speech in spoken dialogue systems [54, 63, 32]. when a complete understanding of spoken input is impossible. In the context of a dialogue system, however, partial understanding must be accompanied by some mechanism to remember context, incorporate new information correctly, and draw inferences among varied and possibly competing input data. Although the systems that use it are typically automatically trainable, using this technology to sustain an interaction requires some heuristic component.

We feel that partial understanding is especially valuable for "one-shot" type applications, where there is no need for a more detailed analysis that will drive an ongoing dialogue. If an immediate and detailed interaction is not required, or even desired, the system has the luxury of performing computationally expensive processing while, at the same time, not having to completely understand everything the user spoke or involve them in a tedious confirmation dialogue. Many of the underlying technologies are currently available in frameworks that allow for easy exploration and experimentation [10].

The extraction of *named entities* from speech has been used with large vocabulary ASR, most notably Broadcast News, associated with the DARPA HUB-4 task [4, 39, 27], as well as with similar corpora in Chinese [64] or French [20]. Although the speech in these corpora is not, for the most part, spontaneous, the extraction of proper names, locations, and organizations represents a significant advancement in the processing of this type of data.

Bechet *et al.* extracted named entities from *spontaneous speech* within the *HMIHY* corpus, concentrating on the extraction of phone numbers from utterances spoken to a customer care application [8]. Huang *et al.* [29] and Jansche and Abney [31] perform named entity extraction on two separate speech corpora of a similar nature, i.e., voicemail transcripts. In both cases, they were looking for "caller phrases," phrases within the voicemail, typically near the beginning of the message, where the caller identifies him-/herself. In addition, caller phone numbers were extracted.

Keyword/phrase-spotting has been used for partial understanding of spontaneous speech in systems where the interaction component is restricted to a single turn. The *How May I Help You* (*HMIHY*) system at AT&T [25] was one of the first to show the viability of extracting salient phrases from unconstrained speech input to perform call routing. Text categorization technology, applied to the same data, showed its applicability to the problem [53].

The output of large vocabulary ASR has also been shown to be accurate enough to be used to *segment* and *index* audio recordings. In developing the SpeechBot system, van Thong *et al.* found that information retrieval performance using text derived via ASR was actually higher than would be expected given the error rate of the ASR engine [59]. Suzuki *et al.* found they were able to perform keyword extraction on radio news using ASR and use that information to identify domains of individual segments [58].

Lecture browsing is another domain in which the output of ASR engines has been shown to be sufficiently accurate to provide real value [22, 41]. To provide entree into hours of otherwise unsegmented spontaneous speech, topics must be discovered and delimited automatically, using keywords and phrases. One of the models used to partition these data, the minimum-cut segmentation model, was originally developed on text data and has been subsequently found to be robust to recognition errors. In comparisons of performance using this model, only a moderate degradation in classification accuracy was observed for speech transcription vs. text [36].

More recent work at AT&T, focussing on voice search, has also sought to extract locations from spoken input, along with query search terms [21]. This level of natural language understanding helps in both making local search more precise and also in inferring a user's intent. In recent work on extracting named entities from utterances derived from a spoken corpus, the identification of named entities was found to significantly improve a later stage of parsing in which a more complete analysis is done [51].

Dowman *et al.* describe a framework for annotating, segmenting and searching audio from television and radio news sources [19]. An interesting component of this system is the use of key phrases extracted from ASR output to find related text documents on the Web, providing a more complete view of a particular news story from both text and audio sources. By unobtrusively offering additions of this sort to searches, the system enhances the information seeking activity of the user. Augmentations such as these need not be limited to keyword searches; GPS coordinates can provide place names that can also be used in such an information mash-up.

All of these studies show that a completely accurate transcription of speech, and a complete analysis of that transcript, is not necessary to build systems that benefit users. Given that we are not engaging the user in a lengthy interaction, we anticipate that the level of detail available from ASR transcripts will be sufficiently robust to make the service we propose a convenient and useful way to annotate one's life.

2.2 Opinion mining/Sentiment detection

Opinion mining and *sentiment detection* represent an extension of the paradigm of performing useful NLP in the absence of a complete parse of the input. These technologies work from tokenized collections of individual words. Many of the techniques employed for opinion mining, sentiment detection, and feature extraction make use of words and phrases found within the text rather than on a complete analysis of each word as it functions within a sentence. Their power comes from the amount of data used to draw inferences, data from consumer-generated media on the Web.

Unlike text written by professional journalists, consumer-generated media cannot be relied on to be grammatical, free of typos, or coherent within the context of an utterance. We anticipate that spoken reviews will be equally informal, but, as long as users speak, in general, the same words and phrases they use in informal written reviews, the technology should be portable, as was shown with text categorization data [53].

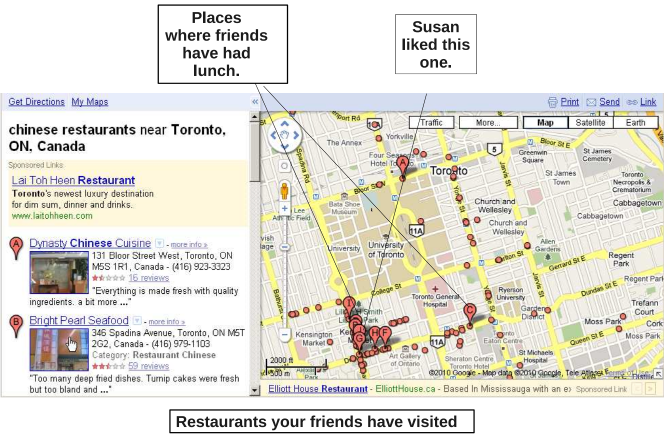
The initial work in sentiment detection focussed on extracting the polarity of a given text, applied to written reviews for a variety of consumer products, as well as entities such as movies and restaurants [49, 42]. As the technology matured, it became possible to determine a more fine-grained rating, indicating a scale of sentiment [17, 23].

For completely unstructured text, such as that found in user-generated content on the Web, it is also useful to automatically extract information about individual features mentioned in reviews. This process of automatic knowledge acquisition can be complemented with sentiment detection to enable a more nuanced understanding of users' opinions [11, 14, 28]. In a further refinement, automatically extracted features are combined with gradient rating of attributes to enable even deeper insight into consumer-generated media [57, 60, 34], It becomes possible to approach the insights of guides such as Zagat's or Consumer Reports, derived from a broad spectrum of opinions, with correspondingly little effort or time.

Some techniques make use of individual utterances and, therefore, utterance boundaries, for either computing the overall polarity of a given text [48] or to subset a larger text into just segments of interest for a market department, for example [30]. The concept of a sentence is evident in dictated speech, as well, and we anticipate it will be part of the spoken reviews we are collecting (see Section 4). Although not all users will speak flawlessly complete sentences, we expect an underlying prosodic and language model to be present nonetheless. Automatic addition of periods and other punctuation has already been shown to be possible in speech and beneficial to performance of automatic speech recognizers [35, 33]. It has been further shown to help in identifying names in speech [26]. We anticipate that aspects of this technology will have to be applied to speech for content creation, and our data collection effort has been devised with this in mind, as well. In two separate styles of interaction, users will provide spoken reviews in ways that encourage both multiple utterance input and individual utterances (see Section 4).

3. USING CONTENT

It is several months later. A friend of the original reviewer is now in Toronto, looking for a place to have lunch. He logs onto the same website, navigates to the Toronto page, and sees a list of reviewed sites. He has shared information himself with his friends and the system has learned that he and our original reviewer have similar tastes. The Chinese restaurant fits perfectly into the profile for both users, and is highlighted (see Figure 4) in the results.



chinese restaurants near **Toronto, ON, Canada**

Sponsored Links

Lai Toh Heen **Restaurant**
Toronto's newest luxury destination for dim sum, dinner and drinks.
www.laitohheen.com

A Dynasty **Chinese** Cuisine - more info »
131 Bloor Street West, Toronto, ON M5S 1R1, Canada - (416) 923-3323
16 reviews
"Everything is made fresh with quality ingredients. a bit more ..."

B Bright Pearl Seafood - more info »
346 Spadina Avenue, Toronto, ON M5T 2G2, Canada - (416) 979-1103
Category: **Restaurant Chinese**
59 reviews
"Too many deep fried dishes. Turnip cakes were fresh but too bland and ..."

Elliott House **Restaurant** - ElliottHouse.ca - Based In Mississauga with an e› Sponsored Link

Figure 4: A representation of the user interface for displaying the results of content creation plus recommendation.

Confronted with a wealth of choices– and even with knowledge summarized from numerous user-generated reviews on consumer websites–users can be as easily overwhelmed by knowledge as empowered by it. Modelling user preferences and behavior helps filter that information into relevant, personalized segments. The acquisition and building of these models depend on the availability and transparency of the data that can be collected about user behavior. We argue that content created by voice on mobile devices enables enriched models of user behavior/preference based on *trust* and *context*.

For building models of both trust and context, the content creation paradigm we envision would enable seamless and nonintrusive collection of the necessary data. Users' interaction can be time-stamped, at either the moment the speech is collected or in a later, post-processing stage. Because the information can be easily associated with specific users–in fact, many users may see that as the entire point of the interaction–sharing will be determined by who users feel would either benefit from or want to know that information. Existing social media and mobile phone contact lists can be leveraged to provide initial sharing networks.

In the remainder of this section, we examine how user models have been used effectively in dialogue systems, and then look more closely at enhancements to these systems via mobile technology.

3.1 User modelling in dialogue systems

User models have shown their utility in spoken dialogue systems for tailoring domain information to fit user preferences. Carenini and Moore applied a mathematical formalism to score, select and organize content for presentation to users in generated recommendations [12, 13]. This work showed that tailoring an evaluative argument to a user's preferences does increase its effectiveness. Carenini and Moore's work has been extended and applied to content selection for recommendations in a spoken dialogue system in a restaurant domain [62].

In deciding what to tell a user about options in a dialogue system, Demberg and Moore show the importance of pointing out trade-offs [18]. They found that, for example, a user who prefers both flying on KLM and taking direct flights is more confident of a system that offers best-possible matching flights but also mentions a sub-optimal (e.g., connecting) flight if it is also on KLM. A follow-up study showed that these refinements to the options space, when volunteered by a system, help reduce dialogue turns [47]. Carenini and Rizoli specifically examine the presentation of opinion-based data in a multimedia setting and argue for the inclusion of dissimilar information and data on the degree of "controversiality" of opinions (i.e., how split the opinions were between positive and negative) [15].

Even outside a traditional dialogue, it is important for systems to process content to help users meet their information-seeking goals. Belkin *et al.* argue that browsing is a natural human activity and guidance is a necessary part of any information interface [9]. In a study in the restaurant domain, it was shown that domain knowledge could be automatically summarized using a combination of machine learning and user modelling [52]. The same automated technology for content selection was applied to a news corpus and shown to help journalists find background data for breaking news stories [6]. By enabling a richer set of meta-data to associate with entities in a system, we anticipate more informative summaries (e.g., “Most people liked this restaurant but Victor didn’t.”).

3.2 Recommender systems

Regardless of the methodology or formalism, modelling preferences helps users make their way through large amounts of data. Both collaborative and content filtering are demonstrably useful and have become an expected aspect for any online shopping experience. Collaborative filtering, however, can be plagued by problems of sparsity, i.e., when few people have recommended only a handful of products, it’s difficult to create and infer from communities of users. Content filtering, which can be more robust in the face of a limited population of users, still requires a stage of either online enrollment or a period of monitored usage in order to learn preferences. Hybrid systems have been proposed to address these issues [5, 24], usually combining the two approaches for recommendations perceived as better overall by users.

Building on computational models of trust [1], systems make use of this notion as an adjunct to collaborative filtering, i.e., as a way of improving on recommendations by using this additional information [44, 38]. If a friend has dined with you in the past and knows your tastes, a particular restaurant you reviewed and liked will be of interest, possibly defining interest. Metrics for trust can be gathered and refined via knowledge of the people users share reviews with, and the degree to which they agree or act upon those reviews.

Adomavicius *et al.* propose that the next generation of recommender systems make use of contextual information, both to address the sparsity problem, as well as to fill a specific and missing need in current systems [3, 2]. They cite a hypothetical example of making movie recommendations, where a system might offer a different choice for a Sunday afternoon matinee (a time when a given user might typically be seeing a movie with her children) than for a Saturday evening (when that same user may have established a pattern of seeing a more adult-oriented film with her partner). In an empirical study, it was shown that time and place had a significant effect on user ratings, in this case of movies [2]. Time-stamped reviews gathered from families with children could provide a prototype for another family’s visit to the same city, including critical information implicitly, e.g., what were good morning activities.

4. DATA COLLECTION

We have already begun the first step in making the restaurant review scenario a reality. Our initial effort at collecting speech for content creation is currently underway. We are collecting these data in a laboratory setting, with subjects recruited from among a self-reported population of people who frequently eat out at restaurants and who are familiar with on-line restaurant review sites. Each of these subjects speaks to a Nokia handset instrumented for the purpose of data collection. Users see the questions shown in Table 1 on the handset and respond by clicking and holding to talk. All utterances are transcribed after recording.

Subjects are randomly assigned to answer one of two questionnaires, both in the restaurant domain. In both questionnaires, the users are asked to rate the *food quality*, *service*, and *atmosphere* of each individual restaurant on a scale of 1-5. In one set of questions, shown at the top of Table 1, users are asked to simply assign a scalar value to the attributes and then rate the restaurant and their experience as a whole in a single, albeit lengthy turn. In the second set of questions, shown at the bottom of Table 1, users are asked to assign a scalar value *and* verbally describe each individual attribute. These users are also asked to provide an overall spoken review, in which they can repeat information previously spoken.

1. What is the name of the restaurant?
2. Where is this restaurant located?
3. What type of cuisine does it serve?
4. What is its phone number?
5. Rate this restaurant on a scale of 1-5, where 1 is poor and 5 is excellent.
6. Rate the food quality on a scale of 1-5, where 1 is poor and 5 is excellent.
7. Rate the quality of service on a scale of 1-5, where 1 is poor and 5 is excellent.
8. Rate the atmosphere on a scale of 1-5, where 1 is poor and 5 is excellent.
9. Please review the restaurant and your experience there in your own words.

1. What is the name of the restaurant?
2. Where is this restaurant located?
3. What type of cuisine does it serve?
4. What is its phone number?
5. Rate this restaurant on a scale of 1-5, where 1 is poor and 5 is excellent.
6. Rate the food quality on a scale of 1-5, where 1 is poor and 5 is excellent.
7. In words, please summarize the food quality.
8. Rate the quality of service on a scale of 1-5, where 1 is poor and 5 is excellent.
9. In words, please summarize the service.
10. Rate the atmosphere on a scale of 1-5, where 1 is poor and 5 is excellent.
11. In words, please summarize the atmosphere.
11. Please review the restaurant and your experience. Repeating information is okay.

Table 1: The two questionnaires used for the ongoing data collection effort in spoken restaurant reviews.

This data collection effort was designed to give us flexibility in designing an initial application and also to provide insight into review data elicited under slightly different protocols. Both sets of questionnaires are designed primarily to collect data that can be used to associate users’ spoken reviews with an automatically derived value representing their sentiment about the restaurant and its attributes. The first set of questions represents an ideal situation, i.e., where a user simply speaks in a free-form manner and we determine both features and polarity ratings. The latter set of data has been designed to capture specific information that might be useful to bootstrap training algorithms for automatically detecting specific features and assigning a graded sentiment representation for each. The latter set may also help us to train models for automatic assignment of utterance boundaries in free-form speech. Both sets of data will be used for language model training.

The data collected here will be valuable in porting technology developed for text to a speech corpus, in addition to giving us insight into the specific issues involved in using speech for creating review data of this sort. To make the vision real, more data will be required. In the initial stages, we may rely more heavily on a user feedback stage (i.e., the scenario at the beginning of Section 2) until algorithms have sufficiently matured to accurately extract all the information we want. However, speech data, along with ancillary information such as position and time, can be acquired via a relatively thin client. National and international consumer review sites, along with existing social media applications, have already accustomed people to the idea of expressing their opinions and sharing them with groups of friends. A convenient platform for expressing these opinions is the next logical step.

5. IN A BROADER CONTEXT

A user in a small village in the developing world critically depends on local, long-distance bus service to buy and sell goods. The bus comes irregularly and the bus stop is far away. However, the bus does follow a prescribed route. When it departs each stop, a user calls a central number and reports the bus's departure via speech. Users can call another number to quickly check on the position of the bus. Each user has to pay a small fee for the service, but users who call and report reliable departure information (e.g., determined by follow-up calls from farther along the route), receive credits for subsequent calls.

This last scenario shows our concept in a broader context. Speech for content creation should not be considered solely an idea for smartphone markets. We hope to make use of this idea to address a current need in developing countries: developing content and providing access to it [37]. Recent advances in the development of ASR for resource-poor languages [43, 7, 16] indicate that "good enough" versions of ASR engines can be obtained for a relatively small cost. These ASR engines can be used in basic spoken dialogue systems. If the system for creating content is also reduced to a more basic functionality, e.g., posting an alert that a bus is on its way, it should be feasible to use speech here, as well.

Sherwani *et al.* showed that information access of a more restricted kind can be successful for low-literate users interacting with a speech interface [56]. In their study, both high-literate and low-literate users achieved higher success rates using a speech interface than one that used touch-tone. This study highlighted the importance of training new users, along with the importance of local facilitators to help introduce new technology. If such an infrastructure is in place, speech interfaces can be profitably used. In general, it seems that an "orality-grounded" HCI is possible in the developing world; specific implementations must be designed carefully, with the exigencies of the developing world in mind [55].

The information needs of farmers in Nigeria were the subject of a study from the Consultative Group on International Agricultural Research, based at the World Bank [46]. Among the specific needs mentioned in this study were current prices, timing of crop planting, and information on group marketing. Infrastructure for information delivery is rudimentary for these farmers, who have limited access to television or even radios. A simple spoken interface could provide critical information in a relatively straightforward and easy-to-learn way.

In the simplest scenario, information could simply be recorded and a key sent via SMS to interested farmers. By calling a number and entering the key, users could hear product planning information. There would be no need for speech processing whatsoever. A simple small-vocabulary system, such as that developed in [50], could provide access to the information via a menu-driven dialogue system instead of keypad input. Farmers who have information could follow the same sort of menu-driven dialogue to input information, e.g., speaking the name of the crop and the town where it was sold, followed by the price.

Economic interests are not the only driver for such systems. Another similar possibility would be a system that allows patients who take certain drugs to share their experiences with those drugs via a "living database." The system could provide access to both typed and spoken testimonials from contributors who share a common medical problem such as side effects from a particular drug. To access the database, people would speak a query such as "is there an association between Lipitor and shoulder pain?" and get back a display listing succinct summaries of all matching hits. Clicking on any one of them would launch an audio playback or open a window showing the complete text entry. The user could then enter their own new contribution to the database as well, if they so desired.

6. CONCLUSION

Spoken language systems should make people's lives easier. An ideal system would be viewed as a convenience, as something users turn to when they are trying to simplify their lives. However, many studies have shown that speech actually increases user's cognitive load. Experiments in using speech in mobile environments have shown that multi-tasking and time pressure have measurable effects on users' speech patterns [40]. Oviatt hypothesizes that less constrained speech systems increase cognitive load by imposing a demand for planning on the part of users [45]. It would seem that a system deployed on a mobile device, eliciting unconstrained speech, could actually be an annoyance.

The key difference in the systems we propose, however, is that the interaction is managed the entire time by the user. Users can choose to devote attention to the system when it is convenient for them. In our most ambitious scenario, the system does not provide any immediate feedback at all. Users are, therefore, not held hostage to a system that cannot proceed until it has understood some part of the input. Whatever processing needs to be done on user input can be delayed indefinitely. When the user does choose to review the results, she can do so later, when she has the time to devote to the task and/or is in an environment where text-based corrections are supported.

Mobile devices make all these scenarios possible, for different reasons. Tourists carry them when they are sightseeing. People have them when they dine, see movies, or go to museums. In the developing world, they are more common than landlines or internet connections [61]. Many of these devices can run thin clients that enable time-stamped speech capture. Higher end devices can associate geo-positioning data with speech. Used effectively, we feel this information can enrich the user experience with mobile devices.

7. ACKNOWLEDGMENTS

The authors would like to thank Chao Wang, Janet Slifka, and Kalina Bontcheva for their valuable help and input on this paper.

8. REFERENCES

[1] A. Abdul-Rahman and S. Hailes. A distributed trust model. In *NSPW '97: Proceedings of the 1997 workshop on New security paradigms*, pages 48–60, New York, NY, USA, 1997. ACM.

[2] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems*, 23:103–145, 2005.

[3] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[4] D. Appelt and D. Martin. Named entity extraction from speech: Approach and results using the textpro system. In *In Proceedings of the DARPA Broadcast News Workshop*, pages 51–54, 1999.

[5] M. Balabanovič and Y. Shoham. Fab: content-based, collaborative recommendation. *Commun. ACM*, 40(3):66–72, 1997.

[6] E. Barker, J. Polifroni, M. Walker, and R. Gaizauskas. Angle-seeking as a scenario for task-based evaluation of information access technology. In *Proc, IUI*, 2009.

[7] E. Barnard, M. Davel, and C. van Heerden. Asr corpus design for resource-scarce languages. In *Proc, Interspeech*, 2009.

[8] F. Béchet, A. L. Gorin, J. H. Wright, and D. Hakkani-Tür. Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How may i help you? *Speech Communication*, 42(2):207–225, 2004.

[9] N. Belkin, C. Cool, A. Stein, and U. Thiel. Cases, scripts, and information seeking strategies: On the design of interactive information retrieval systems. *Expert Systems and Applications*, 9(3):379–395, 1995.

[10] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, 10(3/4):349—373, 2004.

[11] S. R. K. Branavan, H. Chen, J. Eisenstein, and R. Barzilay. Learning document-level semantic properties from free-text annotations. *J. Artif. Int. Res.*, 34(1):569–603, 2009.

[12] G. Carenini and J. Moore. A strategy for evaluating generative arguments. In *Proc. First Int'l Conference on Natural Language Generation*, pages 1307–1314, 2001.

[13] G. Carenini and J. D. Moore. Generating and evaluating evaluative arguments. *Artif. Intell.*, 170(11):925–952, 2006.

[14] G. Carenini, R. T. Ng, and E. Zwart. Extracting knowledge from evaluative text. In *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture*, pages 11–18, New York, NY, USA, 2005. ACM.

[15] G. Carenini and L. Rizoli. A multimedia interface for facilitating comparisons of opinions. In *IUI '09: Proceedings of the 13th international conference on Intelligent user interfaces*, pages 325–334, New York, NY, USA, 2009. ACM.

[16] M. D. Charl van Heerden, Etienne Barnard. Basic speech recognition for spoken dialogues. In *Proc, Interspeech*, 2009.

[17] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 519–528, New York, NY, USA, 2003. ACM.

[18] V. Demberg and J. Moore. Information presentation in spoken dialogue systems. In *Proc., EACL*, 2006.

[19] M. Dowman, V. Tablan, H. Cunningham, C. Ursu, and B. Popov. Semantically Enhanced Television News through Web and Video Integration. In *Second European Semantic Web Conference (ESWC'2005)*, 2005.

[20] B. Favre, F. Béchet, and P. Nocéra. Robust named entity extraction from large spoken archives. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 491–498, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[21] J. Feng, S. Bangalore, and M. Gilbert. Role of natural language understanding in voice local search. In *Proceedings Interspeech, 2009*, 2009.

[22] J. Glass, T. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. Recent progress in the mit spoken lecture processing project. In *Proc., Interspeech*, 2007.

[23] A. B. Goldberg and X. Zhu. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *TextGraphs '06: Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing on the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[24] N. Good, J. B. Schafer, J. A. Konstan, A. Borchers, B. Sarwar, J. Herlocker, and J. Riedl. Combining collaborative filtering with personal agents for better recommendations. In *In Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 439–446, 1999.

[25] A. L. Gorin, B. A. Parker, R. M. Sachs, and J. G. Wilpon. How may i help you? *Speech Communication*, 23:113–127, 1997.

[26] D. Hillard, Z. Huang, H. Ji, R. Grishman, D. Hakkani-tur, M. Harper, M. Ostendorf, and W. Wang. Impact of automatic comma prediction on pos/name tagging of speech. In *Proc. of the IEEE/ACL 2006 Workshop on Spoken Language Technology*, pages 58–61, 2006.

[27] J. Horlock and S. King. Discriminative methods for improving named entity extraction on speech data. In *In Proc., Eurospeech*, 2003.

[28] M. Hu and B. Liu. Mining opinion features in customer reviews. In *AAAI'04: Proceedings of the 19th national conference on Artifical intelligence*, pages 755–760. AAAI Press / The MIT Press, 2004.

[29] J. Huang, G. Zweig, and M. Padmanabhan. Information extraction from voicemail. In *In Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 290–297, 2001.

[30] M. Hurst and K. Nigam. Retrieving topical sentiments from online document collections. In *In Document Recognition and Retrieval XI*, pages 27–34, 2004.

[31] M. Jansche and S. P. Abney. Information extraction from voicemail transcripts. In *In Proc. Conference on Empirical Methods in NLP*, 2002.

[32] T. Kawahara, C.-H. Lee, and B.-H. Juang. Combining key-phrase detection and subword-based verification for flexible speech understanding. In *Proc., ICASSP*, 1997.

[33] Kolář, S. Jáchym, and J. Psutka. Automatic punctuation annotation in czech broadcast news speech. In *Proc., SPECOM*, 2004.

[34] J. Liu and S. Seneff. Review sentiment scoring via a parse-and-paraphrase paradigm. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 161–169, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[35] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(5):1526–1540, 2006.

[36] I. Malioutov and R. Barzilay. Minimum cut model for spoken lecture segmentation. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 25–32, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[37] G. Marsden. Using hci to leverage communication technology. *interactions*, 10(2):48–55, 2003.

[38] P. Massa and P. Avesani. Trust-aware collaborative filtering for recommender systems. In *In Proc. of Federated Int. Conference On The Move to Meaningful Internet: CoopIS, DOA, ODBASE*, pages 492–508, 2004.

[39] D. Miller, R. Schwartz, R. Weischedel, and R. Stone. Named entity extraction from broadcast news. In *In Proceedings of the DARPA Broadcast News Workshop*, pages 37–40, 1999.

[40] C. Müller, B. Grossmann-Hutter, A. Jameson, R. Rummer, and F. Wittig. Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In *UM '01: Proceedings of the 8th International Conference on User Modeling 2001*, pages 24–33, London, UK, 2001. Springer-Verlag.

[41] C. Munteanu, G. Penn, and X. Zhu. Improving automatic speech recognition for lectures through transformation-based rules learned from minimal data. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 764–772, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[42] K. Nigam and M. Hurst. Towards a robust metric of opinion. In *Proc., AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 2004.

[43] S. Novotney and C. Callison-Burch. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Proc, NAACL*, 2010.

[44] J. O'Donovan and B. Smyth. Trust in recommender systems. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, pages 167–174, New York, NY, USA, 2005. ACM.

[45] S. Oviatt. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 871–880, New York, NY, USA, 2006. ACM.

[46] V. N. Ozowa. Information needs of small scale farmers in africa: The nigerian example. *Consultative Group on International Agricultural Research News*, 4(3), 1997.

[47] T. Paksima, K. Georgila, and J. D. Moore. Evaluating the effectiveness of information presentation in a full end-to-end dialogue system. In *SIGDIAL '09: Proceedings of the SIGDIAL 2009 Conference*, pages 1–10, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[48] B. Pang and L. Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[49] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[50] M. Plauche, U. Nallasamy, J. Pal, C. Wooters, and D. Ramachandran. Speech recognition for illiterate access to informaiton and technology. In *Proc., International Conference on Information and Communications Technologies and Development*, 2006.

[51] J. Polifroni and S. Seneff. Combining word-based features, statistical language models, and parsing for named entity recognition. In *Interspeech, 2010, submitted*, 2010.

[52] J. Polifroni and M. Walker. Intensional summaries as cooperative responses in dialogue: Automation and evaluation. In *ACL '08*, 2008.

[53] R. E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. In *Machine Learning*, pages 135–168, 2000.

[54] S. Seneff. TINA: A natural language system for spoken language applications. *Computational Linguistics*, 18(1):61–86, 1992.

[55] J. Sherwani, N. Ali, C. P. Rosé, and R. Rosenfeld. Orality-grounded hcid: Understanding the oral user. *Information Technologies & International Development*, 5(4), 2009.

[56] J. Sherwani, S. Palijo, S. Mirza, T. Ahmed, N. Ali, and R. Rosenfeld. Speech vs. touch-tone: Telephony interfaces for informatio access by low literate users. In *Proc. ICTD, Information and Communications Technologies and Development*, 2009.

[57] B. Snyder and R. Barzilay. Multiple aspect ranking using the good grief algorithm. In *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL*, pages 300–307, 2007.

[58] Y. Suzuki, F. Fukumoto, and Y. Sekiguchi. Keyword extraction using term-domain interdependence for dictation of radio news. In *Proceedings of the 17th international conference on Computational linguistics*, pages 1272–1276, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[59] J.-M. V. Thong, D. Goddeau, A. Litvinova, B. Logan, P. Moreno, and M. Swain. Speechbot: a speech recognition based audio indexing system for the web. In *Proc. of the 6th RIAO Conference*, pages 106–115, 2000.

[60] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 111–120, New York, NY, USA, 2008. ACM.

[61] United Nations International Telecommunications Union. ICT statistics. http://www.itu.int/ITU-D/ict/statistics/at_glance/KeyTelecom99.html, 2009.

[62] M. Walker, S. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, and G. Vasireddy. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840, 2004.

[63] W. Ward. Understanding spontaneous speech. In *HLT '89: Proceedings of the workshop on Speech and Natural Language*, pages 137–141, Morristown, NJ, USA, 1989. Association for Computational Linguistics.

[64] L. Zhai, P. Fung, R. Schwartz, M. Carpuat, and D. Wu. Using n-best lists for named entity recognition from chinese speech. In *HLT-NAACL '04: Proceedings of HLT-NAACL 2004: Short Papers on XX*, pages 37–40, Morristown, NJ, USA, 2004. Association for Computational Linguistics.