Breaking the Monotony of Telephony Voice Interfaces

Sheetal K Agarwal
IBM Research - India
Bangalore
sheetaga@in.ibm.com

Arun Kumar
IBM Research - India
New Delhi
kkarun@in.ibm.com

ABSTRACT
Traditionally, telephony voice applications or interactive voice response systems (IVR) are associated with frustrating caller experience. Yet they are used in several successful customer centre solutions primarily due to lower cost of operations and the scalability advantage. In the context of technology innovations for developing regions they have gained a lot of traction in last few years. For such population consisting of low-literate and low resource people, telephony voice applications make several services immediately accessible serving unaddressed needs. Our experience from field deployments suggests that telephony voice applications can be made lot more effective by not following the established *system controlled navigation* interaction mode available in most voice application platforms of today. In this paper, we present a set of design elements each of which contributes to a shift in Voice User Interaction (VUI) paradigm to a more *user controlled navigation* model closer to the GUI environments of today.

Author Keywords
HCI4D, Information Sharing, ICTD, User-Centered Design, Interactive Voice Systems, Voice-based Telephony Information Systems, India.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. See: http://www.acm.org/about/class/1998/ for more information and the full list of ACM classifiers and descriptors.

INTRODUCTION
Telephony voice applications have been around for more than a decade and are typically used to automate customer services by large companies and various organizations across the globe. However, majority of the voice applications are plagued with usability issues and continue to use touch tone menu interactions which are tightly system controlled. Such menu based interaction can be particularly frustrating, given a) the sequential nature of the voice modality b) fixed paced menus and c) users must remember the application hierarchy mentally to navigate it without getting lost. With advances in Automatic Speech Recognition (ASR) engines, speech input has been replacing touch tone menus to provide conversational interfaces but with limited success. For instance, services such as TellMe[1], provide a seamless user interface that enables users to navigate the application through speech input and without resorting to touch tone menus. However, in noisy environments reliable interaction is difficult to sustain on speech input alone and these services are typically restricted to information services.

The recently emerged natural user interfaces such as Apple's Siri and Nuance's Nina aim to provide human-like conversational interfaces on mobile phones. However, the lack of robust online speech recognition in international languages makes them unreachable for a majority of world's population. Also, similar noisy environment issues remain. This situation demands attention of HCI designers for interim solutions to work till technology advances enable reliable speech based natural language interaction in multiple languages. This is especially important since several services are being deployed to deliver services such as agricultural advice, job searches, healthcare advisories, community portals etc. [14, 15] Majority of these are targeted towards users in developing regions who are semi-literate or illiterate. In this context, researches have proposed new design guidelines to adapt the voice application interaction given the new constraints of the target users [7].

In this paper, we present a few design elements for telephony voice applications derived from our experiences with semi-literate populations in developing countries. These design elements provide ways to enable *user controlled interaction* as opposed to system controlled interaction design prevalent in traditional voice based

[1] http://en.wikipedia.org/wiki/Tellme_Networks. Tellme established an information number, 1-800-555-TELL, which provided time-of-day announcements, weather forecasts, brief news and sports summaries, business searches, stock market quotations, driving directions, and similar amenities

systems. User controlled interaction aims to let the user have control on how he or she would like to use the system rather than being boxed in a fixed menu structure. It focuses more on the services offered by the voice application and less on its structure. Users can concentrate on what they want instead of remembering menu options and the application tree.

DESIGN ELEMENTS FOR USER-CONTROLLED VOICE INTERACTIONS

We present several design elements that attempt to break the monotony that is inherent in voice applications of today.

Controlling the pace of application interaction: *Pause* and *Speedup* knobs

Navigating a non-visual voice based application requires user's undivided attention due to the transient nature of the voice information. Voice applications typically provide an option to allow the user to repeat system information. While this is a useful feature it can often be cumbersome and slow since one ends up navigating the same option multiple times and listening to the same content repeatedly just to reach the desired content. Letting users pause the application or change the speed of interaction (not just for content but for navigation as well), has been found to be very useful [10]. With a Pause knob, a user can effectively pause an application at any time, attend to the distraction while holding the line, and then come back to resume and complete the interaction.

Save as Draft

A related design element to eliminate tediousness from telephony voice applications is to be able to 'pause' an application automatically across calls. Nothing is more frustrating than a dropped call when you have spent several minutes searching for and reaching the desired content or service. With increased reliance on mobile access, it is imperative that telephony voice applications provide the ability to recall a caller's interaction history. This is available in limited form in telephony voice platforms of today. For instance, a user can start from the same content he or she was listening to last but user contributed content and navigation history may not be available in subsequent calls. Much richer interactions can be made persistent across dropped calls. This model can also be applied to other scenarios where users may want to save a browsing session and share it with others – for collective learning or sharing purposes [17].

Search

As more and more users contribute data on voice applications, wading through this content sequentially is a cumbersome task. Searching through audio files is a well-known challenge given the state of the art in speech recognition. Faceted search [14] proposes using meta-data available about the application and the content in itself to perform limited search over user content. To provide search within voice-based systems, Ajmera et al. [2] introduced algorithms that allow for automatic tagging of audio documents in voice-based information systems. Similarly, Srivastava et al. [5] proposed SWAicons that assign auditory cues to improve navigation in voice-based information systems. With large amounts of information being made available through voice applications, ability to search becomes a necessary and obvious tool to make the application interaction faster and rewarding.

Voice Hyperlinks, Bookmarks and Sharing

While searching lets you look for some unknown content, hyperlinks enable you to reach your desired content faster when you know what you are looking for. One simply needs to invoke the link that points to it. The emergence and success of the World Wide Web was driven primarily by the ability of documents to link with each other and create a web of information and services. For voice applications, voice hyperlinks [11, 12] embedded in content and navigation provide the much needed boost to reach content faster across application domains.

Voice hyperlinks also enable bookmarking. Dhanesha et al. [10] demonstrated bookmarking content in voice-based applications for users and how to re-visit that content. Bookmarking entails assigning an identifier, which must be unique for a caller. The application session information required to reach that point in the application is saved in order to revisit the bookmark. Voice hyperlinks across applications are also possible and enabled by protocols such as HSTP [11]

The power of voice hyperlinks can be amplified by letting them out of the voice application and make them shareable through other mediums such as SMS, email etc. Remy et al. [12] studied generation, sharing and use of voice hyperlinks in developing countries.

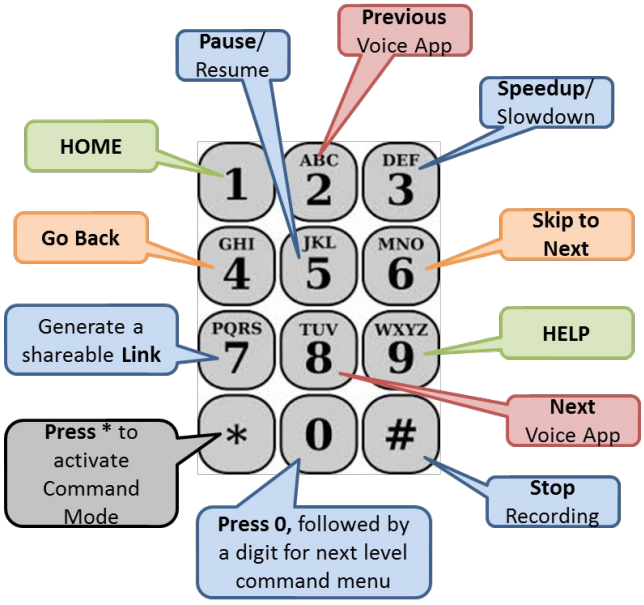
The access mechanism for invoking voice hyperlinks could either by voice based or touch tone based depending upon application needs. The interaction is designed such that the touch tone digit or the voice text is associated with the respective hyperlink. The voice grammar is constructed dynamically by the runtime engine to be able to recognize link invocation and interrupt the application to access the link target.

Interrupt Anytime and Take Control (IATC)

To overcome limitation of single channel communication inherent in voice modality, we propose IATC as a basic technique to allow switching of control from system space to user space. While voice markup standards such as VXML already have some support for it in terms of <bargein> element, it creates problems in noisy environments. Due to background noise, common to mobile users' environment, interruptible voice interfaces result into

false interruptions leading to a very poor user experience.

We found it extremely useful to be able to let the voice interaction be non-interruptible but offer an alternate interruption mechanism through touch tone signals sent by punching phone keypad. The figure below depicts the mapping of phone keys to concepts that we used to aid voice application navigation. '*' key is used to interrupt the application and switch to user controlled navigation. So, '1' represents Home or beginning of the application, '5' to pause or resume the application, '4'/ '5' to go back or forward in the application flow, '2' / '8' for previous or next item, '#' to stop a voice recording and '0' to extend the keypad menu to another series of commands (For eg. *01 activated a voice hyperlink).



Combining speech input with touch tone input brings the best of both worlds. It lets the users interact using speech in the supported language of their choice and the touch tone input provides quick interruptability without introducing any tedious delays.

The hated "Goto": Random Jump

Branched menus [8], present in a typical telephony voice application, allow users to navigate back up one level or to the root of the hierarchy at any point in the call interaction. Skip & Scan [9] menus present one menu option at a time to the user. The user must give a command to either select the current option or move to the previous or next menu option. With both these navigation styles user needs to remember menu options in previous levels which can be cumbersome and often leads to a frustrating user experience. We need a better mechanism to let users decide which portion of the application they would like to visit next. While a 'Goto' construct providing a random jump in the application logic is disliked by programming language experts, in voice interfaces it can do wonders. A 'Goto' enables the user to select a valid menu option from anywhere in the application irrespective of his current position in the application tree. The caller requests for visiting a particular section by interrupting the current flow and then speaking out the name of that section. The Spoken Web Application Framework [], includes a random jump module that computes the path to the target menu option from the current position and automatically traverses the hierarchy for the user.

It has been shown that for three-level menus, the branched menu structures are satisfactory[2]. Random jump can come in handy as the voice application content scales up. It removes the onus of remembering the application hierarchy from the user allowing him to interact with the application more naturally.

Multimodal Inputs

Another effective mechanism that alleviates the problem associated with the sequential and single dimensional nature of voice interfaces is to open up additional channels of communication. Researchers have proposed to augment voice inputs to a telephony voice application with gesture based and other haptic inputs such as tapping or scratching on the phone [13].

CONCLUSION

While natural language speech interfaces become a reality for telephony voice applications in international languages, HCI designers need to find innovative ways to overcome the cumbersome nature of current voice based systems. We proposed to bring a change in the interaction model of voice applications by switching from a system-controlled navigation to user-controlled navigation. We presented several design elements, derived from field-deployments and user feedback that help make such design possible.

While majority of phones in developing regions are still feature phones, the smartphone market is fast growing with phones available at same cost as feature phones. With users owning more sophisticated phones, augmenting the telephony voice interaction with a parallel graphical user interface is a promising direction to provide richer user experience. Essentially, an application on the smartphone run in synchrony with the server based telephony voice app and display pictures/icons/video on the smartphone screen in the context of content being played or accessed in the voice call. Such a telephony voice application driven mobile app can provide a very powerful interaction model. Remy et al [12] present a basic multimodal application to share voice hyperlinks in mobile app driven mode. However, such synchronized applications either need an active internet connection to synchronize the mobile app with server app, or use traditional channels such as SMS

[2]http://spotlight.ccir.ed.ac.uk/public_documents/technology_reports/No.6%20Menus.pdf

and USSD[3]. The application data such as images, text etc. could either be preloaded into the mobile application or sent to the device during the voice call.

REFERENCES

1. Agarwal, S., Kumar, A., Nanavati, A.A., and Rajput, N. Content creation and dissemination by-and-for users in rural areas. In Proc. International Conference on Information and Communication Technologies and Development (ICTD) 2009.
2. Ajmera, J., Joshi, A., Mukherjee, S., Rajput, N., Sahay, S., Shrivastava, M., and Shrivastava, K. Two Stream Indexing for Spoken Web Search. In Proc. International World Wide Web Conference (WWW) 2011.
3. Danesha, K.A., Rajput, N., and Srivastava, K. User Driven Audio Content Navigation for Spoken Web. In Extended Abstracts, Multimedia (MM) 2010.
4. Reddy, H., Annamalai, N., and Gupta, G. Listener-controlled dynamic navigation of voicexml documents, ICCHP, volume 3118 of Lecture Notes in Computer Science, page 347-354. Springer, (2004)
5. Hemambaradara Reddy, Narayan Annamalai, and Gopal Gupta. ICCHP, Volume 3118 of Lecture Notes in Computer Science, page 347-354. Springer, (2004)
6. Srivastava, S., Rajput, N., and Mahajan, G. SWAicons: Spoken Web Audio icons: Design, Implications and Evaluation. In Proc. International Conference on Computer Supported Cooperative Work 2012.
7. White, J., Duggirala, M., Srivastava, S., and Kummamuru, K. Designing a Voice-based Employment Exchange for Rural India. In *Proc. ICTD 2012*.
8. Sharma Grover, A, Stewart, O and Lubensky, D. 2009. Designing interactive voice response (IVR) interfaces: localisation for low literacy users. Proceedings of Computers and Advanced Technology in Education (CATE 2009), St Thomas, US Virgin Islands, 22-24 November 2009
9. Dailogues Spotlight Research Team, University of Edinburgh, Navigation in Structured and Unstructured Menus, Technical Report. http://spotlight.ccir.ed.ac.uk/public_documents/technology_reports/No.6%20Menus.pdf
10. Paul Resnick and Robert A. Virzi. 1992. Skip and scan: cleaning up telephone interface. In *Proceedings of the* SIGCHI Conference on Human Factors in Computing Systems (CHI '92), Penny Bauersfeld, John Bennett, and Gene Lynch (Eds.). ACM, New York, NY, USA, 419-426. DOI=10.1145/142750.142881
11. Dhanesha K, Rajput N , Srivastava K: User driven audio content navigation for spoken web. ACM Multimedia 2010
12. Agarwal S, Chakraborty D, Kumar A, Nanavati A, and Rajput N. 2007. HSTP: hyperspeech transfer protocol. In Proceedings of the eighteenth conference on Hypertext and hypermedia (HT '07)
13. C Remy, SK Agarwal, A Kumar, S Srivastava, Supporting Voice Content Sharing among Underprivileged People in Urban India Human-Computer Interaction–INTERACT 2013
14. Robinson S, Rajput N, Jones M, Jain A, Sahay S, and Nanavati A. 2011. TapBack: towards richer mobile interfaces in impoverished contexts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)
15. Diao M, Mukherjea S, Rajput N, and Srivastava K. 2010. Faceted search and browsing of audio content on spoken web. In Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)
16. Agarwal, S., Kumar, A., Nanavati, A.A., and Rajput, N. VoiKiosk: increasing reachability of kiosks in developing regions. In Poster Proc. International World Wide Web Conference (WWW) 2008
17. Patel, N., Chittamuru, D., Jain, A., Dave, P., and Parikh, T.S. Avaaj Otalo - A Field Study of an Interactive Voice Forum for Small Farmers in Rural India. In Proc. ACM Conference on Human Factors in Computing Systems (CHI) 2010
18. Farrell R, Das R, Rajput N: Social Navigation through the Spoken Web: Improving Audio Access through Collaborative Filtering in Gujarat, India**.** AAAI Spring Symposium: Artificial Intelligence for Development 2010
19. Kumar A, Agarwal S K, Manwani P., The spoken web application framework: user generated content and service creation through low-end mobiles, In Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)

[3] Support for USSD based applications in smartphone based OSes such as Android is limited.