

Talk it Out: Adding Speech Interaction To Support Informational and Transactional Applications on Public Touch-Screen Kiosks

Kheng Hui Yeo, Rafael E. M. Banchs

Institute for Infocomm Research, Singapore

{yeokh, rembanchs}@i2r.a-star.edu.sg

Abstract

In this paper we present a method that enables people to interact with large touch-screen displays using spoken language. Such panels are commonly used in public areas for information and advertisement. A personal mobile device is used to convert the speech signal into text, which is then sent to the panel via an Internet connection. This allows speaker adaptation and background noise issues to be mitigated. Information may also be sent to the user for later reference on their personal device.

Index Terms: human-computer interaction, dialogue systems, multi-modal interfaces, automatic speech recognition.

1. Introduction

Spoken dialogue systems have been deployed for different purposes, such as restaurant information, directory assistance, and conference information. A common characteristic amongst dialogue systems is the fact that they usually make use of a telephone connection. However, other interaction approaches are possible: for example, during the INTERSPEECH 2014 conference held in Singapore, a virtual agent with dialogue capability [1] was deployed as part of the official mobile application. This agent was used to provide event information.

Public places such as shopping malls typically have static directory signs providing information about a specific area of interest, be it a particular store or an amenity. In some places, touch-screen kiosks have been introduced instead. These kiosks serve not only information, but facilitate also personal registration, payment, and reservations, providing a more interactive experience. Adding speech would further enhance the user experience, allowing for intuitive interaction.

Therefore we present our idea for further enhancing such public kiosks with speech interaction, given a reliable internet connection. In this paper we will describe the system architecture, then the integration of its components, as well as provide a few use cases. Finally we will discuss possible applications areas for our system.

2. System Architecture

The hardware and software components, as well as their integration, can be described as follows:

2.1. Hardware

The kiosk is running Windows 7 Embedded Standard on a 1.99GHz quad-core Intel Celeron CPU and 4GB of RAM. It has a vertical 42" capacitive touch screen, WLAN connectivity, and audio output via an external speaker. This

can be substituted by a smaller device, for example, a laptop computer, iPad, or Android tablet.

Any mobile device with a wireless internet connection and a web browser can connect to this kiosk. In the current setup we use a HTC One X mobile phone running on Android 4.2.2.

2.2. Software

The main software component is the APOLLO dialogue engine [2], a platform developed in our department. It coordinates the communication between the web interfaces running on both the mobile phone and touch screen. XML scripts control the communication between the core engine and its respective plugins.

The MySQL database provides structured data for questions within a known domain. In this application, for example, it would be information relating to the shops in the mall. This information was collated from official sources and additionally verified with site visits.

An additional index implemented in Lucene was created to handle more general questions which may not fit the domain anticipated in the database. These could be generic food recommendations or questions about transportation options in the vicinity of the mall, for instance.

The kiosk GUI contains four main elements: a speech-enabled avatar, a text area where the conversation between the



Figure 1: The (a) kiosk and (b) mobile client UI

system and the user is displayed, a frame which offers additional information, and buttons for navigation. By default, the frame shows a listing of shops which the user may touch for more details, as seen in Figure 1.

In contrast, the mobile client connects to a much simpler web interface with only two elements, a text box and a button to end the interaction session. To start talking, users will

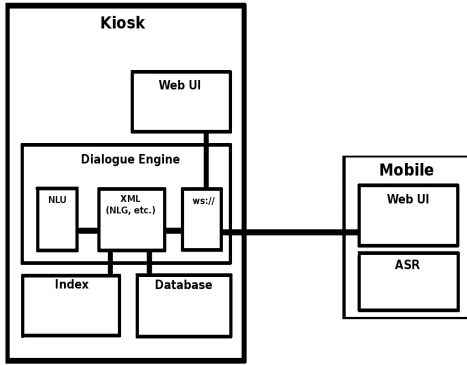


Figure 2: Components and how they interact

activate the speech recognition routines on their own devices after clicking in the text box. Pressing the Enter key will send the text to APOLLO.

Both web interfaces are hosted on the same web server, Apache 2.4, running on the kiosk.

2.3. Interaction of Components

Figure 2 shows the block diagram of the entire system, and the communication between components. Both the kiosk and mobile client are represented as discrete hardware components. They are connected via WebSockets (the component labelled ws:// in the diagram) to a running instance of APOLLO.

A spoken request is processed and translated into text by the user's mobile device. On an Android device, this would typically be Google's speech recognition engine. As it may vary, the link to an external service is not explicitly shown here. This text is then sent via the client web interface to the kiosk where APOLLO is running.

The text received by APOLLO is parsed into a list by a grammar-based NLU plugin. This list is then read by the XML script which calls other plugins to retrieve relevant data from both the database and the index. Here, a special heuristic is used to select the most suitable answer. Finally, the answer is displayed on the kiosk GUI.

After 20 seconds with no incoming interaction from the user, the kiosk will revert to its initial state to prepare for the next user. The kiosk may also be reset manually using the 'Home' button on the web interface.

3. Example Use Cases

Here is a sample dialogue in which a user is requesting directions to a particular outlet in the mall:

Speaker	Utterance	Data
System	Good morning, nice to meet you!	-
User	How do I get to black canyon	-
System	Please see the map above.	Map
User	Thank you.	Map

In this situation, the user requests the directions to a specific outlet explicitly. The system identifies this as a direction-type question, and prepares an answer which includes an animated map. This map can be downloaded to the user's device for later reference. The user then ends the session as they would a conversation with a human concierge. This also causes the UI to revert to its initial state.

In this next example, the user asks for a recommendation for a specific type of cuisine, rather than an outlet. In the system's database, there is only one option, so this is presented along with the official website in the kiosk UI.

Speaker	Utterance	Data
System	Good morning, nice to meet you!	-
User	Can you recommend some Thai food here?	-
System	There is one option: Black Canyon Coffee.	Web-site
User	How do I get there?	Web-site
System	Please see this map above.	Map
User	Thank you.	Map

In this example it is assumed that the user would follow the system's recommendation and would like to go there. As the most recent conversation topic was "Black Canyon Coffee", it is automatically identified by the system and the relevant map is produced.

4. Future Work

Further evaluation needs to be done to compare the performance of this system with its alternatives, such as recording the speech input from an embedded microphone on the kiosk rather than via the mobile device. Metrics such as recognition accuracy used in the evaluation of similar systems [3] may also be applied during user testing.

As far as the application is concerned, work is being done to fully exploit the data transfer capabilities offered by the connection between kiosk and mobile device. For instance, directions may be downloaded as an animation to the user's device so that navigation is easier when the user goes away from the kiosk. Retail tenants at the mall employing this kiosk solution may choose to send their latest offer as a downloadable coupon to the user's mobile device.

5. Conclusion

Our proposed technique allows public kiosks to incorporate speech while interacting with visitors. This method of using a personal mobile device as a speech input source rather than an in-built microphone on the kiosk allows us to leverage the customised speech recognition systems trained on a particular user's traits. In addition, the distance between the microphone and the user is reduced, minimising the effect of background noise on recognition performance.

6. References

- [1] D'Haro, L. F., Kim, S. K., Yeo, K. H., Jiang, R., Niculescu, A. I., Banchs, R. E., Li, H. CLARA: a multifunctional virtual agent for conference support and touristic information. Proceedings for International Workshop for Spoken Dialogue Systems 2015. Busan, South Korea.
- [2] Jiang, R., Tan, Y. K., Limbu, D. K., Tran, A. T., Li, H. A Configurable Dialogue Platform for ASORO Robots. Proceedings of Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2011. Xi'an, China
- [3] Georgila, K., Sgarbas, K., Tsopanoglou, A., Fakotakis, N., Kokkinakis, G. A Speech-Based Human-Computer Interaction System for Automating Directory Assistance Services. International Journal of Speech Technology 6, pp. 145-159, 2003.