Two Alternative Frameworks for Deploying Spoken Dialogue Systems to Mobile Platforms for Evaluation "In the Wild"

Helen Hastie, Marie-Aude Aufaure*, Panos Alexopoulos, Hugues Bouchard, Heriberto Cuayáhuitl, Nina Dethlefs, Milica Gašić, Almudena González Guimeráns, James Henderson, Oliver Lemon, Xingkun Liu, Peter Mika, Tim Potter, Verena Rieser, Pirros Tsiakoulis, Yves Vanrompay, Boris Villazon-Terrazas, Majid Yazdani, Steve Young and Yanchao Yu

email: h.hastie@hw.ac.uk. See `http://parlance-project.eu` for full list of affiliations

Abstract

We demonstrate two alternative frameworks for testing and evaluating spoken dialogue systems on mobile devices for use "in the wild". We firstly present a spoken dialogue system that uses third party ASR (Automatic Speech Recognition) and TTS (Text-To-Speech) components and then present an alternative using audio compression to allow for entire systems with home-grown ASR/TTS to be plugged in directly. Some advantages and drawbacks of both are discussed.

1 Introduction

This abstract describes the EC FP7 PARLANCE project whose goal is to perform interactive search through speech in multiple languages. With the advent of evaluations "in the wild", emphasis is being put on converting research prototypes into mobile applications that can be used for evaluation and data collection by real users downloading the app from the market place. This is the motivation behind the work demonstrated here. We present a modular framework whereby research components from the PARLANCE project (Hastie et al., 2013) can be plugged in, tested and evaluated in a mobile environment. The domain is interactive search for restaurants in San Francisco, USA. All required restaurant information is obtained through a Yahoo search API which returns entities based on their longitude and latitude within San Francisco for 5 main areas, 3 price categories and 52 cuisine types containing approximately 1,600 individual restaurants.

2 Two System Architectures

The first framework adopts a client-server approach as illustrated in Figure 1 for the PARLANCE system in Mandarin (Hastie et al., 2014). This system uses third party Google ASR and TTS, where the recognised utterance is sent to the Stanford Segmenter[1] server and the segmented utterance is then sent to the Spoken Language Understanding (SLU), Interaction Manager (IM) (Thomson and Young, 2010), Natural Language Generation (NLG) (Dethlefs et al., 2013) and TTS components in sequence. For details of all the PARLANCE components please see (Hastie et al., 2013) and the project website[2].

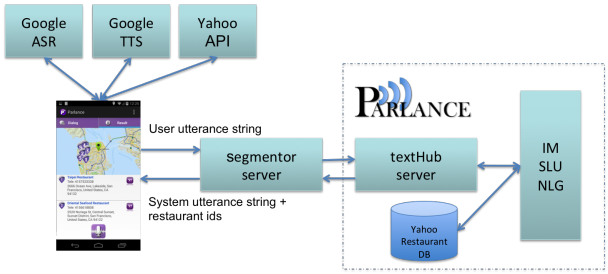

Figure 1: Architecture 1: the PARLANCE Mandarin mobile application system architecture using third party ASR/TTS.

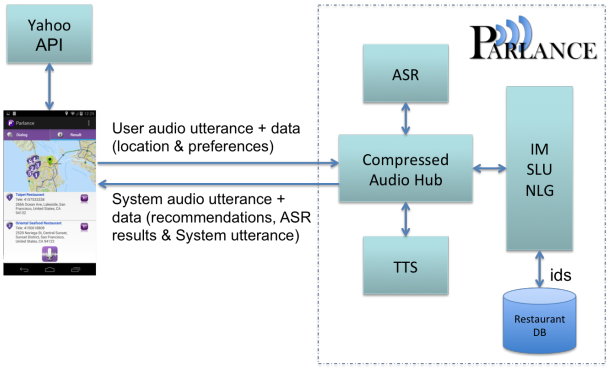

Figure 2: Architecture 2: the PARLANCE English mobile application system architecture using audio compression.

[1] http://nlp.stanford.edu/projects/chinese-nlp.shtml
[2] http://parlance-project.eu

*Authors are in alphabetical order

However, we were also interested in integrating and evaluating all the PARLANCE capabilities, such as user barge-in and incrementality (Hastie et al., 2013) and did not want to rely on third party software. Therefore, we developed an alternative architecture for the English version using a SIP client-server communication. However, this proved sensitive to bandwidth variations and some carriers and Internet service providers block it. The final version avoids this problem by transferring highly compressed audio and data using internet connectivity as illustrated in Figure 2.

Similar dialogue system frameworks also make use of audio compression for network-based ASR (Pieraccini et al., 2002) and TTS (Kruijff-Korbayová et al., 2012). They also transfer audio files (but without compression) for network-based ASR and use either a server TTS (Gruenstein et al., 2008) or a client TTS (Fuchs et al., 2012). Others train language understanding components from crowdsourcing based on speech input and output components running on a server (Liu et al., 2013).

2.1 Discussion of Architectures

Advantages of the first architecture include rapid development and easy portability to new domains. This is due to off-the-shelf components being used which save effort in development and testing. This is true for dialogue systems in multiple languages, where home-grown ASR/TTS do not exist. An advantage of the second architecture is that home-grown and domain-specific ASR and TTS components can often lead to better performance than off-the-shelf components (Dušek et al., 2014; Tsiakoulis et al., 2014). However, reasonable response times per turn should be taken into account (between 100 and 500 milliseconds) (Strömbergsson et al., 2013). Another advantage of the second architecture is that is allows incremental processing for input analysis and output planning. This has been shown to lead to more natural interactions that human users prefer over their non-incremental counterparts (Skantze and Schlangen, 2009).

2.2 Multimodal Functionality

In addition to spoken dialogue, the mobile app features substantial multi-modal interaction functionality. It displays the set of results during the conversation with the system and allows refinement and inspection of the results while talking. Hyper-local features include being able to sort results by distance from the user and also organised by neighbourhoods or nearby Points-of-Interest (POIs) (Bouchard and Mika, 2013). This last feature is particularly appealing in a tourism scenario where the user may not be aware of neighbourhoods in the city, but might remember the location of major sights. Screenshots of the English mobile app (Architecture 2) are shown in Figures 3 and 4.

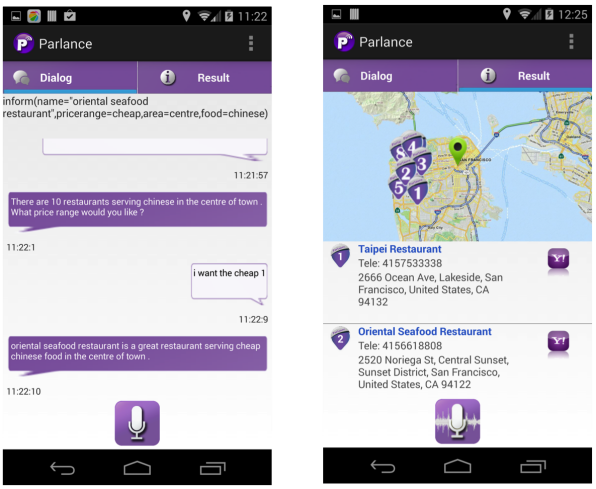

Figure 3: Screenshot of a dialogue and the list of recommended restaurants also shown on a map.

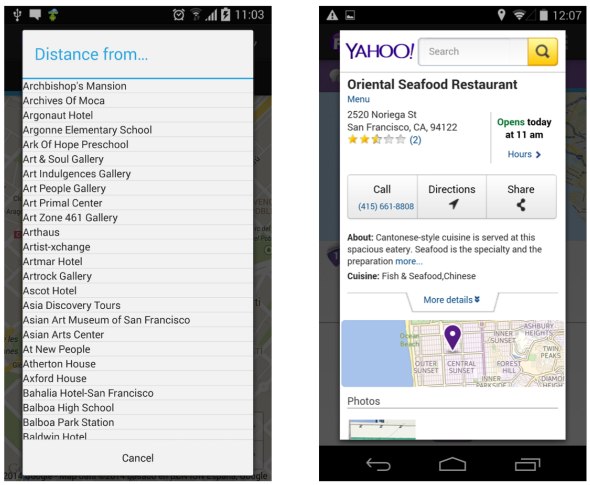

Figure 4: Screenshot of recommended restaurants and ordering by distance from points of interest.

3 Future Work

Future work involves developing a feedback mechanism for evaluation purposes that does not put undue effort on the user and put them off using the application. In addition, this framework could be extended to leverage social information of the user when displaying items of interest.

Acknowledgements

The research leading to this work was funded by the EC FP7 programme FP7/2011-14 under grant agreement no. 287615 (PARLANCE).

References

H. Bouchard and P. Mika. 2013. Interactive hyperlocal search API. Technical report, Yahoo Iberia, August.

N. Dethlefs, H. Hastie, H. Cuayáhuitl, and O. Lemon. 2013. Conditional Random Fields for Responsive Surface Realisation Using Global Features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

O. Dušek, O. Pltek, L. Žilka, and F. Jurčíček. 2014. Alex: Bootstrapping a spoken dialogue system for a new domain by real users. In *Proceedings of SIGDIAL*, Philadelphia, PA, U.S.A.

M. Fuchs, N. Tsourakis, and M. Rayner. 2012. A scalable architecture for web deployment of spoken dialogue systems. In *Proceedings of LREC*.

A. Gruenstein, I. McGraw, and I. Badr. 2008. The WAMI toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In *Proceedings of ICMI*.

H. Hastie, M.A. Aufaure, P. Alexopoulos, H. Cuayáhuitl, N. Dethlefs, M. Gasic, J. Henderson, O. Lemon, X. Liu, P. Mika, N. Ben Mustapha, V. Rieser, B. Thomson, P. Tsiakoulis, Y. Vanrompay, B. Villazon-Terrazas, and S. Young. 2013. Demonstration of the PARLANCE system: a data-driven incremental, spoken dialogue system for interactive search. In *Proceedings of SIGDIAL*, Metz, France, August.

H. Hastie, M.A. Aufaure, P. Alexopoulos, H. Bouchard, C. Breslin, H. Cuayáhuitl, N. Dethlefs, M. Gašić, J. Henderson, O. Lemon, X. Liu, P. Mika, N. Ben Mustapha, T. Potter, V. Rieser, B. Thomson, P. Tsiakoulis, Y. Vanrompay, B. Villazon-Terrazas, M. Yazdani, S. Young, and Y. Yu. 2014. The PARLANCE mobile application for interactive search in English and Mandarin. In *Proceedings of SIGDIAL*, Philadelphia, PA, U.S.A.

I. Kruijff-Korbayová, H. Cuayáhuitl, B. Kiefer, M. Schröder, P. Cosi, G. Paci, G. Sommavilla, F. Tesser, H. Sahli, G. Athanasopoulos, W. Wang, V. Enescu, and W. Verhelst. 2012. Spoken language processing in a conversational system for child-robot interaction. In *Proceedings of WOCCI*.

J. Liu, P. Pasupat, S. Cyphers, and J. Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *Proceedings of ICASSP*.

R. Pieraccini, B. Carpenter, E. Woudenberg, S. Caskey, S. Springer, J. Bloom, and M. Phillips. 2002. Multimodal spoken dialog with wireless devices. In *Proceedings of ISCA Tutorial and Research Workshop - Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, Germany.

G. Skantze and D. Schlangen. 2009. Incremental Dialogue Processing in a Micro-Domain. In *Proceedings of EACL*, Athens, Greece.

S. Strömbergsson, A. Hjalmarsson, J. Edlund, and D. House. 2013. Timing responses to questions in dialogue. In *Proceedings of INTERSPEECH*.

B. Thomson and S. Young. 2010. Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. *Computer Speech and Language*, 24(4):562–588.

P. Tsiakoulis, C. Breslin, M. Gasic, M. Henderson, D. Kim, M. Szummer, B. Thomson, and S. Young. 2014. Dialogue context sensitive hmm-based speech synthesis. In *Proceedings of ICASSP*.