Play Something by The Beatles

Aidan Kehoe
Logitech
Cork, Ireland
akehoe@logitech.com

Amer Chamseddine
EPFL
Lausanne, Switzerland
amer.chamseddine@epfl.ch

Asif Ahsan
Logitech
Newark, CA, USA
aahsan@logitech.com

ABSTRACT
In certain scenarios, voice access to a music library can be a desirable method of interaction. This position paper reports experiences from a number user studies conducted in the course of a project that explored enabling voice access to a music library in mobile usage scenarios. Many of the problems encountered in the studies are clearly attributable to core speech recognition engine performance. But there is also a broad range of additional challenges that must be addressed in the context of designing such a system in order to enable a positive user experience.

Author Keywords
Speech Interfaces; Music Library; Interact Design.

ACM Classification Keywords
H.5.2. Information interfaces and presentation.

General Terms
Human Factors; Design.

INTEREST IN WORKSHOP
The text on the workshop homepage "many real-life applications using speech technologies do not require 100% accuracy to be useful" caught our attention. We had recently worked on a project in which there were significant challenges with speech recognition accuracy. However, in the context of this project, even with much less than 100% accuracy, we believe it is possible to create a pleasant user experience. Coming from industry, the workshop would be a good way for us to make contacts with people (mostly academics?) actively working in this area.

INTRODUCTION
The ubiquitous availability of smartphones allows people to access their music library in a wide variety of scenarios. Their content may be stored on their smartphone, tablet, personal computer, or in the cloud. However, it typically requires a number of discrete steps to access the content, e.g., access the physical device, swipe to unlock, select an app, select some content and start to play.

Designing Speech and Language Interactions Workshop 2014

DSLI Workshop at CHI'14 , 2014, Toronto, Canada.

This multi-step interaction is problematic is many real world scenarios involving mobile audio. People sometimes actively listen and focus on music. However, people also like to listen to music while performing other tasks. They listen to music while studying, commuting, showering, cooking, exercising, etc. In many of these situations the user's hands, and possibly eyes, are also busy; and in these cases voice access to a music library can be very desirable.

With the availability of powerful smartphones, ubiquitous network access and a range of commercially licensable speech recognition engines; many of the components required to prototype and explore such functionality are now readily available to developers. This position paper reports some high-level findings from a number of user studies conducted in the course of development and evaluation of such a prototype system. The findings highlight that there are many additional very significant challenges in addition to speech recognition accuracy.

Of course, given the potential widespread use and perceived value of voice-enabled access to music, such functionality has been attempted in several commercial products and also explored in the context of academic research. It is an active area of research and development.

Commercial Products
In recent years Apple iOS, Android and Microsoft Windows-based products have been making very significant enhancements in speech interaction on mobile devices. For example, with default out-of-the-box settings, iPod/iPhone iOS7 devices have a limited command and control vocabulary that allows users to play/pause music, play a specific album, etc. Voice assistant software such as Apple Siri, Samsung S-Voice and Nuance Nina allow for more advanced functionality through use of server-based recognition.

As of this date (January 2014), with respect to voice interaction with music library content, the functionality of such systems remains very limited. Firstly, the interaction model is often such that the music playing app can be launched but there is limited (or no) capability for on-going voice interaction after the app has been launched. Secondly, the speech recognition engines have not been optimized for use with music content (typically the recognition engines support settings for "dictation" or for "web search" usages only). This can ultimately result in a low success rate for user requests to play specific music. Finally, the "always

listening" capabilities of mobile devices are only now beginning to become mainstream. As a result, voice interaction must usually be triggered by physical interaction with the device, e.g., button press.

Academic Research

In addition to existing commercial products, academic research on a broad range of topics including speech recognition engines, natural language processing, search and recommendation systems are all very relevant to developers in this area. Research associated with voice interaction for car entertainment systems often address many of these issues. There are also many books, and developer implementation guides, to assist designers of voice enabled systems.

PROJECT PHASES & USER STUDIES

The project kicked off with a number of structured interviews with smart phone users to understand their current methods of interacting with music, and their requirements and desires for voice interaction. The next phase of the project involved creating a number of functional prototypes which were used during a number of limited-scope formative evaluations [7]. When the prototype system reached a certain level of maturity a larger and more structured lab-based study with 12 participants was conducted to benchmark overall system performance.

The development platform selected was Android 4.x. This open platform allowed for easier experimentation with a variety of cloud-based recognition services (Google Speech SDK, Nuance Dragon SDK, SoundHound SDK) and a number of embedded recognition engines (Nuance VoCon Embedded SDK, CMU Sphinx). The ability to try a number of different recognition engines was important to understand any cost/performance tradeoffs. At the start of the project there were no publically available relative comparison benchmarks across the various different recognition engines that could be used as a selection guide.

Phase One: Interviews

A number of user studies were conducted. The first study involved interviews with 12 smart phone users (9 iOS, 3 Android; 5 female, 7 male; ages ranging from 16 to 40) that accessed music frequently as part of their daily routines. Participants provided details of their music listening behavior and current interaction methods.

They also described the type of voice interaction that they would envisage to access their music via their smart phone. This provided seed material for usage scenarios, and samples of user vocabulary that the participants would use for voice interaction. Following the initial interviews with users, the project moved to the prototype development phase.

Phase Two: User Studies with Functional Prototypes

During the course of the development a number of smaller scale formative evaluation studies were conducted with typically 4-6 participants. These studies proved important to enhance the grammar, gain a better understanding of failures and explore recovery methods, and tune certain system parameters (when and whether to give feedback or not, time outs, etc.). There is a lot of great high-level guidance to developers of voice interaction systems literature, but for any given project there is also a lot of application-specific tuning based on user studies required.

These studies also allowed for collection of recordings of user interactions. Such recordings of user utterances were important, since they could be later re-submitted to speech recognition engines offline as part of a benchmarking process.

DISCUSSION

This section of the position paper discusses some of the high-level findings from the user studies, from the designer/developer perspective. The engine recognition accuracy was a noticeable problem in many of the studies, as had been expected. However, there were many additional issues encountered that also needed to be addressed to improve the overall user experience.

WER & Successful Completion Rates

Word Error Rate (WER) is a widely used metric in the context of evaluation of speech recognition engines. Ideally, there would be readily available published WER data (from engine developers) to give developers some indication of what can be expected in their application, but this is not the case today. Also, the various engines used in the prototypes did not have a standardized batch submission system for online processing of recordings; such a capability would be very useful for developers for benchmarking.

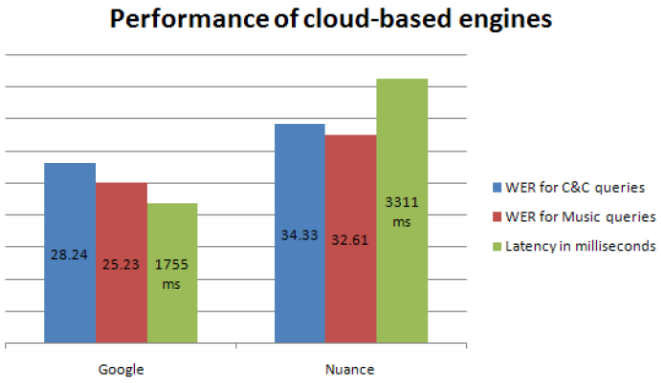

Figure 1. WER and Latency

Figure 1 shows the results for offline submission of the same batch of recordings to both Nuance Dragon SDK and to Google Voice SDK. The WER metric proved useful in understanding differences (and similarities) in system performance while using different recognition engines. In the case of this prototype system, the studies showed it was

also helpful to be able to look at the "command-and-control" recognition rate (for core commands such as play/play/stop/etc.) independently from those words that referred to specific contents of the music library (album/artist/playlist names).

While users were inclined to expect that basic commands always worked, they were be somewhat more forgiving of errors associated with music content requests, i.e., they were not always sure of an exact song title or album name.

However, WER alone is not a sufficient metric. Ultimately the goal is to achieve an acceptable success rate for interactions, and deliver a pleasant user experience. Even with a relatively high WER, many of the resulting problems could be addressed through use of supporting basic natural language processing, and not requiring 100% matches with music library content prior to playing music.

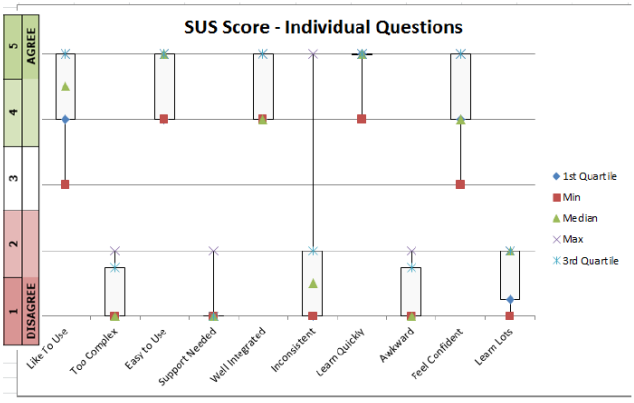

Figure 2. SUS Questionnaire Results.

In this project the System Usability Scale (SUS) [1] was found to be a good complement to the WER metric. Figure 2 shows the box results for some of the individual SUS questions for a 12-person user study. Even with a low WER, the SUS score can be impacted by a wide variety of other items such as latency, allowing freedom in user vocabulary, errors handling, etc.

Vocabulary & Keywords

When designing a vocabulary and grammar for voice interaction to a music library, there are a number of existing commercial products that can be a useful reference, .e.g., the Apple iOS (with Siri disabled, or enabled) and Google Now vocabularies. These vocabularies allow users to play specific content, and also allow a number of queries such as "What is playing?" and "Who is this song by?".

These vocabularies did cater for the majority of the interactions encountered in the user interviews and studies, particularly in relation to basic command-and-control and requests to play content. But there were also a number of other additional interactions that were encountered. These highlighted user interest in voice functionality beyond basic command-and-control and access to specific material in their library.

This included requests relating to:

- Recent activity, e.g., "play the album I listened to on iTunes last night"
- User context, e.g., "I'm tired, play something relaxing"

Another interesting component in the user vocabulary related to the user's understanding of the component they were interacting with. Some users were inclined to specify a particular device in their vocabulary (iPhone); others indicated a specific app (Spotify).

Challenges Associated with Music Library

There are also a variety of challenges associated with matching speech requests with contents of a music library. Some of these are related to user voice input. For example, according to Tashev et al. [9], more than 60% of songs are referred to by people using names that do not match their actual title field. This may be because the song is known under a different name, or because some metadata fields contain incorrect information; the end result is that a simple matching algorithm will not work well.

Most of the time, people include in their query information from multiple metadata fields (e.g., "Play Something by The Beatles") [8], and they can also mistake the song title with the artist or album, which renders the matching process a challenging task. Ju et al. [5] worked on improving the perceived accuracy of speech systems by accepting such ill-formed queries. But as with this Beatles example, the exact meaning of the user utterance can also be interpreted in two different ways.

There are many additional challenges in relation to the content of the music library itself. Music library metadata can contain a mixture of languages, and include many non-standard artist and media content spellings. Commercial products such as GraceNote SDK can help address these issues.

The user's interaction history and library metadata may provide opportunities to improve performance; but it also introduces additional challenges. For example, the average iTunes user only listens to 19% of his/her music library [3]. Most people listen to the same song over and over again [4]. 90% of iTunes music libraries have missing or incorrect metadata [6].

Another example is, according to TidySongs [2], the average iTunes user has 7,160 songs; with 490 songs without an artist name, 1,984 jams without track or year information, and 814 duplicate songs. This obviously can have a negative impact on the possibility to match user utterances with music library content.

Platform Constraints and Use "In the Wild"

The numerous additional challenges associated with

Bluetooth platform constraints, difficulties encountered with user studies outside the lab environment and multi-user scenarios; these are important real world challenges but not discussed in this position paper.

CONCLUSION

This position paper discusses some very high-level findings from a project the explored voice access to a music library. Many of the problems encountered in the studies are clearly attributable to core speech recognition engine performance. From a developer/designer perspective the ability to easily benchmark performance from a number of different speech engine vendors would be highly desirable.

But beyond recognition accuracy, there is also broad range of additional challenges that must be addressed in the context of designing such a system in order to enable a positive user experience.

While many of the software components required to attempt to build such a system are readily available, it still remains a very significant design and development challenge to create a system that works robustly outside of a lab environment with a broad range of users.

There are many large commercial companies devoting significant resources to development of voice-enabled personal assistants. Applications such as voice search, calendar management and texting have been the primary focus for voice assistant software and functionality to date. In the future, it seems reasonable to expect that these assistants would broaden their functionality to support a number of other activities, including improving access to music. As a result, this position paper is very much a partial snapshot of the situation as it exists today, and this may change rapidly.

REFERENCES

1. Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189, 194.
2. Ehrlich, B. Just how messy is the average user's itunes library? http://mashable.com/2011/01/04/itunes-library/, 2011.
3. Hiner, J. Average itunes user only listens to 19 http://www.techrepublic.com/blog/hiner/average-itunes-user-only-listens-to-19-of-music-library/, 2011.
4. Inquisitr. On demand music listening patterns over time, 2011. URL http://www.inquisitr.com/107151/graph-on-demand-music-listening-patterns-over-time/
5. Ju, Y. C., Seltzer, M. L., & Tashev, I. (2009, September). Improving perceived accuracy for in-car media search. In INTERSPEECH (pp. 979-982).
6. Kahney L, Average itunes library has 3k songs and is heavily mislabeled. http://www.cultofmac.com/103614/103614/, 2011.
7. Nielsen, J. (2000). Why you only need to test with 5 users.
8. Song, Y. I., Wang, Y. Y., Ju, Y. C., Seltzer, M., Tashev, I., & Acero, A. (2009, April). Voice search of structured media data. In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on (pp. 3941-3944). IEEE.
9. Tashev, I., Seltzer, M., Ju, Y. C., Wang, Y. Y., & Acero, A. (2009). Commute UX: Voice enabled in-car infotainment system. In Mobile HCI (Vol. 9).