Voice Access to Music: Evolution from DSLI 2014 to DSLI 2016

Aidan Kehoe
Logitech
Cork, Ireland
akehoe@logitech.com

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

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

Denis O'Keeffe
Logitech
Cork, Ireland
dokeeffe@logitech.com

Abstract

In certain scenarios, voice access to a music library can be a desirable method of interaction. This position paper is a follow on to one accepted for the DSLI 2014 CHI workshop. In the two years since that workshop, speech recognition engine performance has measurably improved, and there are a number of widely available systems that support voice access to music functionality. However, there remains a broad range of additional challenges that must be addressed to continue to improve the user experience of such interactions.

Author Keywords

Speech Interfaces; Music Library; Interaction Design.

ACM Classification Keywords

H.5.2. Information interfaces and presentation.

Interest in Workshop

This is a follow up to the 2014 DSLI CHI workshop paper by the same authors. Several of our product concept explorations involve speech and multimodal interaction. Coming from industry, the workshop would be a good way for us to make contacts with other people actively working in the speech and multi-modal interaction area.

Introduction

The ubiquitous availability of smartphones, and increasing presence of voice enabled wearables and IoT devices, allows people to access their music library in a wide variety of scenarios. Their music content may be either stored locally on their device, or increasingly commonly, may be streamed from the cloud. Voice offers the possibility of natural, direct and hands-free access this is especially useful in multi-tasking scenarios.

Many of the components required to prototype and explore such functionality have continued to improve their capabilities since the 2014 position paper; smartphones are even more powerful, network access more ubiquitous, and a range of licensable speech recognition engines are improving speech recognition rates. In addition, there are a number of commercially available solutions that offer natural language processing, AI and cloud computing components that are necessary parts creating a complete functioning system.

Commercial Products

Apple iOS, Android, Microsoft and Amazon products have been making continuous enhancements for speech interaction with music content. Out-of-the-box settings for the smartphone-based products, without network connectivity, typically 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, Microsoft Cortana and Amazon Alexa leverage more sophisticated cloud processing and offer more advanced functionality, with specializations for interaction with music content.

As of this date (January 2016), with respect to voice interaction with music library content, the capabilities of such systems have notably improved since 2014. Voice commands for specific artists, albums, playlist and genres are supported both in smartphones, and in systems optimized for in-car use. Since the launch of Apple Music, Siri has been enhanced with capabilities to “like” songs, play “top songs”, etc. Amazon Echo offers music search capability and integrates with a number of different music services including TuneIn and Pandora.

The “always listening” capabilities of mobile devices are only now beginning to become mainstream. As a result, today smartphone voice interaction is typically triggered by physical interaction with the device, e.g., button press. Amazon Echo has “always listening” capability with interaction triggered by a specific spoken key phrase.

Benchmarking Improvements

For a developer, the ability to prototype with a variety of different recognition engines is important to understand potential cost/performance tradeoffs. There is still a lack of transparent and publically available relative performance comparison benchmarks across the various different recognition engines that could be very helpful to developers in their engine selection process [4].

Vocabulary

The DSLI 2014 position paper describes the development and use of the vocabulary that was utilized in that study. For a relative comparison with performance in August 2013, that same vocabulary was used again to gather performance data in December 2015. It consisted of 100 music-related requests, 66 of which related to requests to play specific content (the other 34 phrases were related to command-and-control and settings). Some example music query phrases from the 2013 vocabulary that were re-used in 2015:

- Play Achtung Baby by U2
- Listen to some Adele songs
- Play the first song in Enrique's album Insomniac

WER Comparison for Aug 2013 and Dec 2015
Figure 1 shows the average WER (Word Error Rate) for the user study phrases for August 2013 and December 2015 when submitted to Google Speech Recognition API.

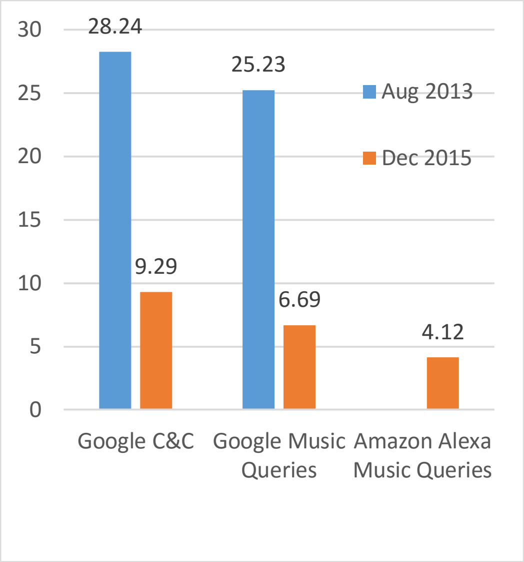

Figure 1: Word Error Rate in Aug 2013 and Dec 2015.

The WER is shown for both command-and-control queries, and for music-related queries. There is a very noticeable reduction in average WER versus 2013.

For Amazon Echo only the 2015 data for music queries is shown (there were some technical problems in collecting the command-and-control data). The Amazon Echo product was not available in 2013, so it is not possible to compare relative WER.

It is important to note that the recordings of the study participants took place in a quiet office environment as they read the queries from the study script; so it was in close to ideal settings. Regardless, these improvements since 2013 are indicative of the performance improvements being reported by the developers of speech recognition engines [2, 8].

Challenges for Developers

This section of the position paper outlines some of the challenges from the perspective a developer trying to use speech recognition engines, together with other tools, to enable voice access to a music library.

Method for Offline Recording Submission
WER, even if it has notable limitations [5], 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 still not the case today. Doubtless such capabilities are available internally in the companies that develop the engines, they are just not publically available.

In the absence of this publically available data it would be very helpful for engines to support a standardized batch submission system for offline processing of user recordings; such a capability would be very useful for developers in iterative benchmarking at scale.

Shared Music Search Query Corpus

In the past, open publically available corpora of speech recordings have proved valuable for developers in benchmarking and tuning performance, e.g., TIDIGITS [9]. Something similar for music-related queries has been suggested in the past by Bainbridge et al [1]. This type of resource would be very valuable given the diversity in music and artist names, spelling, pronunciation, etc. The public availability of such resources, together with a standardized way to process offline recordings, would be helpful for developers in benchmarking. Such a resource would also be helpful for engine developers too.

Language Model Flexibility

Some of the publically available speech recognition APIs allow an application to specify details of the desired language model. For example, the Android speech recognizer can specify settings for WEB_SEARCH (for web search queries) or FREE_FORM (for dictation). Some engines also allow specification of an associated grammar file, against which the speech input can be processed and constrained.

Would performance be improved if applications (and engine training) specifically supported a MUSIC_SEARCH category? Apps that are specifically targeting music access would have the ability to give that additional information to the speech engine.

Challenges Associated with Music Library

The challenges associated with matching speech requests with contents of a music library, as outlined in the DSLI 2014 position paper, still remain. Some of these are related to user voice input, even with zero WER, due to incomplete or incorrect utterances [10]. For example, according to Tashev et al. [11], more than 60% of songs are referred to by people using names that do not match their actual title field. As a result, basic matching algorithms do not work well in many cases. However, we observed that the Spotify search API can sometimes give meaningful matches in the absence of incomplete or inconsistent data.

Back in 2011, people were primarily using locally storage and had a limited music library (often constrained by on-device storage space). For example, according to TidySongs [3], the average iTunes user had 7,160 songs. Today most streaming services offer access to a broadly similar catalogues of 30+ million songs. This growth in catalogue size could increase the probability of an incorrect or duplicate match. In our user studies we've encountered "errors" where the same song title was sung by many different artists; and also where artists had multiple versions of the same song in the catalogue and the user wanted to listen to a specific one. To mitigate these difficulties it is important to have a user model that considers the interaction history, and incorporates the probabilities that people re-listen to the same music frequently [6, 7].

Conclusion

Our DSLI paper in 2014 ended with this paragraph:
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.

And it did change rapidly! In the 2+ years since the data was collected for the 2014 DSLI workshop paper, speech recognition engine performance has measurably improved. Today there are a number of widely available and frequently used assistants that have enabled voice access to music.

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. In our studies we encounter many failures in noisy environments (outdoors, when music playing), when multiple people are speaking, when people are speaking in a less formal manner trying to remember a song or playlist title, etc.

References

1. David Bainbridge, John R. McPherson and Sally Jo Cunningham. Forming a Corpus of Voice Queries for Music Information Retrieval. *ISMIR*. 2002.
2. Li Deng et al. "Recent advances in deep learning for speech research at Microsoft." *Acoustics, Speech and Signal Processing (ICASSP)*, IEEE 2013.
3. B Ehrlich. 2011. Just how messy is the average user's itunes library? Retrieved Jan 2, 2016 from http://mashable.com/2011/01/04/itunes-library
4. Huang, Xuedong, James Baker, and Raj Reddy. "A historical perspective of speech recognition." *Communications of the ACM 57.1* (2014): 94-103.
5. Xiaodong He, Li Deng and Alex Acero. Why word error rate is not a good metric for speech recognizer training for the speech translation task *Acoustics, Speech and Signal Processing (ICASSP)*, 2011 IEEE International Conference.
6. Hiner, J. 2011. Average iTunes user only listens to 19% Retrieved Jan 2, 2016 from http://www.techrepublic.com/blog/hiner/average-itunes-user-only-listens-to-19-of-music-library
7. Lapcaprica 2011. On demand music listening patterns over time. Retrieved Jan 2, 2016 from http://www.inquisitr.com/107151/graph-on-demand-music-listening-patterns-over-time/
8. Lei, Xin, et al. "Accurate and compact large vocabulary speech recognition on mobile devices." INTERSPEECH. 2013.
9. Leonard, R. Gary, and George Doddington. "Tidigits speech corpus." Texas Instruments, Inc. 1993.
10. Song, Y. I., Wang, Y. Y., Ju, Y. C., Seltzer, M., Tashev, I., & Acero, A. 2009. Voice search of structured media data. *In Acoustics, Speech and Signal Processing, IEEE ICASSP* April 2009.
11. 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)*.