Report on the 1st Workshop on Audio Collection Human Interaction (AudioCHI 2022) at CHIIR 2022

Gareth J. F. Jones
Dublin City University
Ireland
Gareth.Jones@dcu.ie

Maria Eskevich
CLARIN ERIC
The Netherlands
maria@clarin.eu

Ben Carterette
Spotify
United States
benjaminc@spotify.com

Joana Correia
Spotify
United States
joanac@spotify.com

Rosie Jones
Spotify
United States
rjones@spotify.com

Jussi Karlgren
Spotify
Sweden
jkarlgren@spotify.com

Ian Soboroff
National Institute of Standards and Technology
United States
ian.soboroff@nist.gov

Abstract

This is a report from the AudioCHI 2022 workshop which was organised in conjunction with the CHIIR conference in March 2022. The workshop was organised to bring together researchers in spoken content retrieval with expertise on human computer interaction in information access to examine opportunities and challenges for advancing technologies for search and interaction with spoken content. The workshop was originally planned to be a full day event, but due to pandemic-imposed travelling constraints, it was instead held as a virtual half day event to accommodate time zone differences for participants.

Date: 14 March, 2022.

Website: https://speechretrievalworkshop.github.io/.

1 Interaction is a key factor for access to spoken data

Recorded speech comes in many forms: collections of recordings, live audio chats, streamed and broadcast; in many styles and genres: factual or entertaining (or both!), timely or of historical interest, local or global, single speaker or conversations; and from various sources: media houses, entertainers, non-professional individual podcasters, art houses, stage production companies, news organisations, consumer brands, public agencies. Listeners approach these data with various expectations and various intentions.

The availability of recorded speech is growing and diverging from previous practice. New technologies and services for the distribution of audio content introduces new use cases, of which several but not all adhere to previously known media consumption patterns. The increasing number of audio creators brings with it a larger variety of material. New and emerging genres and new interaction models will allow new forms of interacting with spoken content.

The starting points of the workshop are that the study of human engagement with spoken material in search settings is crucial to understanding requirements for future system design and delineating future research. The organisers believe that there needs to be an ongoing conversation between researchers interested in understanding and studying usage, user needs and intentions, and interaction design; researchers interested in understanding and analysing content and form of spoken material; and researchers working on information retrieval systems of various types. We see that an convergence of knowledge and practice from these three strands of knowledge are necessary to move the field forward.

The AudioCHI workshop on Audio Collection Human Interaction was organised at CHIIR, where researchers with an interest in understanding interaction with information systems gather to discuss (i) how content analysis might establish verbal and non-verbal features for rich content representations, and (ii) how use cases and human factors in interaction with spoken audio content might inform the design of information access technology [Jones et al., 2022].

2 Challenge Questions

The workshop was organised around three challenge questions:

Use cases What is the parameter space of use cases for search, retrieval, exploration of speech audio?

Features What are some interesting audio features we do not pay enough attention to?

Shared tasks Can we suggest some new shared tasks that would span the above space interestingly?

3 Previous research

There is a considerable body of previous work studying spoken document retrieval or more generally spoken content retrieval, [Jones, 2019], but in line with most of information access technology it has primarily focussed on task-motivated topical access to fairly static collections. While this family of use cases is present for information access to speech, there are many more ways of using spoken material: listeners engage with spoken material for a variety of reasons, including entertainment, current affairs, education, and research [Jones et al., 2021b].

Search of spoken content collections has been the focus of a number of previous shared benchmark tasks such as the Cross-Language Speech Retrieval and the Cross-Language Spoken Document Retreival tracks at CLEF [Jones, 2019], the Spoken Document Retrieval track at TREC [Garofolo et al., 2000], the SpokenDoc and Spoken Query&Doc tasks at NTCIR [Akiba et al., 2016], and the Rich Speech Retrieval and Search & Hyperlinking tasks at MediaEval [Larson et al., 2011; Eskevich et al., 2014]. AudioCHI 2022 took as its most recent point of departure the first two years of the TREC Podcasts Track, which explored: (i) adhoc search for segments from podcast episodes, both with respect to topicality as well as other criteria of relevance, and (ii) summarisation of podcasts [Jones et al., 2021a; Karlgren et al., 2022].

4 Workshop Format and Deliberations

The workshop included two presentations:

- Moreno La Quatra: "Bi-modal Architectures for Deeper User Preference Understanding from Spoken Content" arguing for the joint processing of textual and audio data in speech retrieval, to provide a richer representation better to be matched against user preferences [La Quatra et al., 2022].
- Doug Oard: "Talking with the Planet" presenting a conversational approach to extracting information from recorded interviews and oral histories, pointing out that retrieval methods that are geared towards finding a succinct response to an expression of information need may not be well suited to finding an answer which is distributed over the course of a conversation and would need collation or aggregation and contextualisation to be useful to fulfill the information request [Oard, 2022].

Participation was about 25 people, mostly CHIIR participants, but also some with a background in phonetics and the analysis of archival recorded speech. The challenge questions were discussed, and most of the time was spent on use cases rather than features or shared tasks. On a shared board, the participants were asked to record some use cases they themselves engage in, including the latest session in which they listened to spoken material. These were then collectively organised on a surface with two axes:

Pickiness on a scale from "want this exact listening item" to "sure, I'll give this other item a shot" to reflect the differential between known-item search vs broad intent search.

Curation effort to reflect the difference between importance of the session to reflect the effort the listener is prepared to put into curating the session, with the one end representing a low threshold to abandon listening entirely and moving to another activity, and the other a considerable effort to put together a listening experience.

These axes are not independent of each other and the resulting board is given in Figure 1. The discussion also ranged over information needs that were less bound to a specific situation such as quote verification ("Did they really say that?"), rapid skimming over large collections, sampling and remixing, and tracking customer calls to a help desk. Some collections mentioned during the conversation were not amenable to transcription, due to their audio qualities or due to the variety of language in the recordings. An example of this type is dialectal field studies, where the objective itself is to collect speech which does not conform to standardised expectation with respect to intonation, pronunciation, or lexicon.

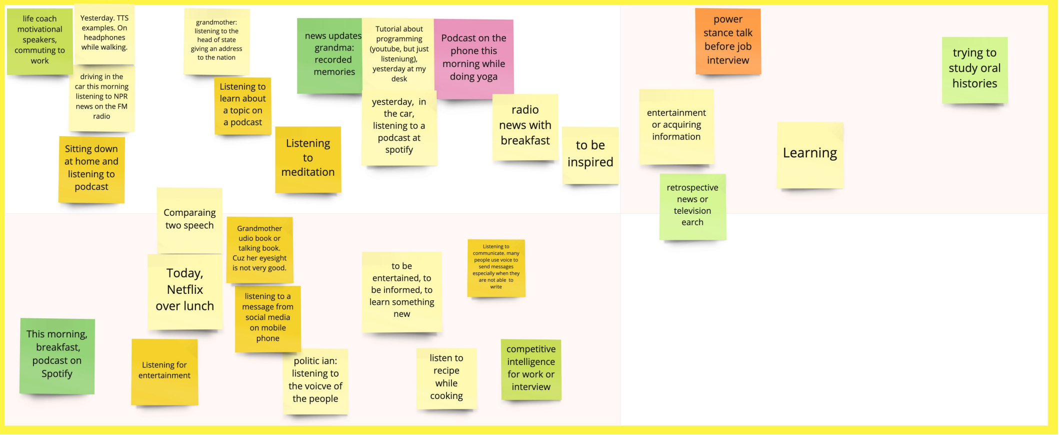
life coach motivational speakers, commuting to work
Yesterday. TTS examples. On headphones while walking.
driving in the car this morning listening to NPR news on the FM radio
Sitting down at home and listening to podcast
grandmother: listening to the head of state giving an address to the nation
Listening to learn about a topic on a podcast
news updates grandma: recorded memories
Listening to meditation
Tutorial about programming (youtube, but just listeniung), yesterday at my desk
yesterday, in the car, listening to a podcast at spotify
Podcast on the phone this morning while doing yoga
radio news with breakfast
to be inspired
power stance talk before job interview
entertainment or acquiring information
retrospective news or television earch
Learning
trying to study oral histories
Comparaing two speech
Grandmother udio book or talking book. Cuz her eyesight is not very good.
Today, Netflix over lunch
listening to a message from social media on mobile phone
to be entertained, to be informed, to learn something new
Listening to communicate. many people use voice to send messages especially when they are not able to write
This morning, breakfast, podcast on Spotify
Listening for entertainment
politic ian: listening to the voicve of the people
listen to recipe while cooking
competitive intelligence for work or interview

Figure 1: Shared discussion board from the workshop, placing use cases in relative positions according to pickiness (up to down) and curation effort (left to right)

5 Key Outcome

AudioCHI 2022 discussed information access to speech collections by considering how users engage with spoken material; how the characteristics of spoken content (temporal qualities, non-verbal and prosodic features, etc.) interact with those use cases; how novel usage scenarios add to the mix; and how the study of human factors can contribute to understanding the interplay of speech features, spoken content, and usage.

The workshop aimed to to provide a first landscape of the contact surfaces between audio and speech analysis on the one hand, and information access research in use cases and usage situations on the other. Since the workshop was shorter than originally intended many of the strands of thought — parameter spaces for features, the design of future shared tasks, most notably — were tabled for future events, where we look forward to continuing the conversation.

References

Tomoyosi Akiba, Hiromitsu Nishizaki, Hiroaki Nanjo, and Gareth J. F. Jones. Overview of the NTCIR-12 SpokenQuery&Doc-2 task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access*. National Institute of Informatics, 2016.

Maria Eskevich, Robin Aly, David Nicolas Racca, Roeland Ordelman, Shu Chen, and Gareth JF Jones. The search and hyperlinking task at mediaeval 2014. In *MediaEval*. Citeseer, 2014.

John S Garofolo, Cedric GP Auzanne, Ellen M Voorhees, et al. The TREC Spoken Document Retrieval Track: A Success Story. *NIST SPECIAL PUBLICATION SP*, 500(246):107–130, 2000.

Gareth J. F. Jones. About Sound and Vision: CLEF Beyond Text Retrieval Tasks. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World*. Springer, 2019.

Gareth JF Jones, Maria Eskevich, Ben Carterette, Joana Correia, Rosie Jones, Jussi Karlgren, and Ian Soboroff. CHIIR workshop on audio collection human interaction (audiochi 2022). In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 377–378, 2022. URL http://speechretrievalworkshop.github.io.

Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth JF Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. TREC 2020 Podcasts Track Overview. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC)*. NIST, 2021a.

Rosie Jones, Hamed Zamani, Markus Schedl, Ching-Wei Chen, Sravana Reddy, Ann Clifton, Jussi Karlgren, Helia Hashemi, Aasish Pappu, Zahra Nazari, LongQi Yang, Oguz Semerci, Hugues Bouchard, and Ben Carterette. Current Challenges and Future Directions in Podcast Information Access. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021b.

Jussi Karlgren, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth JF Jones, Rosie Jones, Sravana Reddy, and Edgar Tanaka. TREC 2021 Podcasts Track Overview. In Ian Soboroff and Angela Ellis, editors, *Proceedings of the Thirtieth Text REtrieval Conference (TREC)*. NIST, 2022.

Moreno La Quatra, Lorenzo Vaiani, Luca Cagliero, and Paolo Garza. Bi-modal architectures for deeper user preference understanding from spoken content. In Gareth J. F. Jones, Maria Eskevich, Ben Carterette, Joana Correia, Rosie Jones, Jussi Karlgren, and Ian Soboroff, editors, *Position paper for the AudioCHI Workshop at CHIIR*, 2022. URL https://speechretrievalworkshop.github.io/Bimodal.pdf.

Martha A Larson, Maria Eskevich, Roeland Ordelman, Christoph Kofler, Sebastian Schmiedeke, and Gareth JF Jones. Overview of mediaeval 2011 rich speech retrieval task and genre tagging task. In *MediaEval*. Citeseer, 2011.

Douglas W. Oard. Talking with the planet. In Gareth J. F. Jones, Maria Eskevich, Ben Carterette, Joana Correia, Rosie Jones, Jussi Karlgren, and Ian Soboroff, editors, *Position paper for the AudioCHI Workshop at CHIIR*, 2022. URL https://speechretrievalworkshop.github.io/Talking%20with%20the%20Planet.pdf.