Citizen Tagger: Exploring Social Tagging of Conversational Audio

Delvin Varghese(✉), Patrick Olivier, and Madeline Balaam

Open Lab, Newcastle University, Newcastle upon Tyne, UK
{d.varghese2, patrick.olivier, madeline.balaam}@newcastle.ac.uk

Abstract. This paper discusses Citizen Tagger (CT), a mobile application for tagging audio-based chat-show content. The application allows users to create audio and text tags (annotations). Through an iterative design process, CT was designed and deployed with 16 members of a faith-based community who tagged a panel discussion about 'faith and vocation'. Based on usage statistics, analysis of created tags, and other qualitative data, the user experiences of tag creation were assessed. Questions around how to configure tagging-related parameters were investigated, and diverse user motivations for creating tags were also explored. Tagging was discovered to be a subjective experience, with participants expressing a desire to customise their tagging setup. Furthermore, despite being instructed to tag for content organisation and retrieval, users utilised tagging as a tool for self-reflection.

Keywords: Social tagging · Multimodal interaction · Audio annotations · Assisted note-taking · Speech modality

1 Introduction

Human interaction with the world is inherently multimodal [20]. Schaffer et al. [18] looked at the state of the art in speech interfaces and found that despite Voice User Interfaces (VUIs) being part of many smartphones and navigational systems (and often saving interaction steps or time), Graphical User Interfaces (GUIs) are still the more common input modality. However, providing alternate input modalities accommodate a wider range of users, tasks, and environmental conditions [15]. While audio as an input modality has received much attention in the HCI literature [2, 18, 20], relatively little work has looked at it within the context of audio annotation or tagging of audio files. Anguera [1] presented a tagging application to annotate digital photos using speech input, which showed that such interfaces can harness this under-used input modality in everyday tasks. Cherubini et al. [4] have previously compared the differences between text and audio tags in the context of a mobile-based photo annotation and retrieval task. They found that participants took longer to tag text and in general, participants preferred voice-based tagging.

Social tagging [11] is defined as a community of users applying free-form tags to digital objects [23]. When applied to chat show content, which is conversational and unscripted in nature, serves as a way of giving voice to the (otherwise passive) users

R. Bernhaupt et al. (Eds.): INTERACT 2017, Part IV, LNCS 10516, pp. 116–125, 2017.
DOI: 10.1007/978-3-319-68059-0_7

who are part of radio shows, giving them a role in knowledge production [7, 11]. CT was designed to assess the feasibility of such a concept i.e. investigating the complexities of tagging resources such as conversational audio. This mobile application enables a user to add audio and text tags as they are listening to a piece of audio content. This technology was deployed with 16 members of a UK Christian community, and the application usage statistics and feedback that was received through post-usage surveys and interviews was analyzed. The primary contribution of this piece of work is to assess (i) the concept of audio and text-based tagging of audio content, and (ii) user experiences of audio-based tagging.

1.1 Note-Taking During Meetings

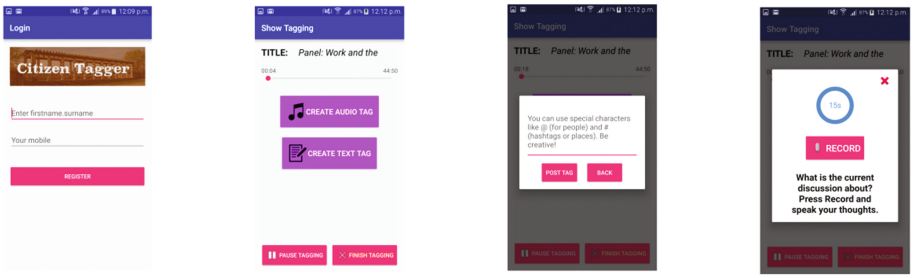
Chiu et al. have done previous work on capturing meetings and assisting note taking by using digital video and ink in a physical conference room. They augmented the meeting recording process by enabling tools to support indexing, accessing and browsing the captured meeting [5]. By noting various changes in the meeting room e.g. timestamp of user note creation, presentation slide changes etc. multimedia systems can support users in indexing and annotating interesting areas in the content [6]. Similarly, other work has explored the indexing of notes in an automated meeting capture environment [12, 14], where the index in the notes is populated with the relevant multimedia content i.e. presentation slide for which the notes are taken [12].

1.2 Social Tagging and Folksonomies

When tagging is studied in the literature, it is taken to refer to tagging of entire resources and not within resources i.e. music is tagged but tagging of indices or segments within music files (i.e. *intra*-content tagging) [21] is an under-researched area. Sack et al. [16] state that current tagging systems "produce a hit list that contains entire resources, although the tags describing these resources might refer to specific parts of those resources". SoundCloud (soundcloud.com), a technology that is a well-known social audio platform that utilizes timed comments within audio content, is an interesting technology to look at in relation to this work as it provides a framework for user-generated text-based comments (or tags) that are visible during playback. Text-based applications are common, and Singh et al. have commented how speech has been under-utilized as an input modality for tagging [19]. They utilized a narrative structure to help reconcile the technological challenges of speech-based applications and the challenges arising from unmet user expectations and needs.

2 Initial Design of CT

In our initial design iterations, it was considered important to ensure that a tagging user was not unduly influenced by the tags created by another user. A *blind tagging* system (where a tagging user cannot view tags assigned to the same resource by other users while tagging) was thus used to assist the tagging process [13]. Another key design consideration was to encourage free-form tagging, where the users are given the ability



a) The login screen. b) Show playback screen. c) Text tagging popup. d) Audio tagging popup.

Fig. 1. Citizen Tagger app screenshots

to express their tags in any keyword(s) as deemed relevant by them [22]. To ensure that tags are created on a regular basis, a timer was built into the application which displayed tagging popups. The popups prompted the user to create an audio tag by giving a brief instruction followed by a 'Record' button (seen in Fig. 1). Users could dismiss the tag if they did not want to create a tag at that moment. The frequency of the prompts can be configured in the *Config* menu tab, where other options such as logging out, clearing previous tags, changing the live show settings etc. are also available. Users were also able to add text and audio tags manually at their discretion.

The show is stored on the web server and can either be streamed by the user or downloaded for offline listening and tagging. The ability to pause/rewind/fast-forward playback was also present. It was important to not disrupt the flow of the audio playback when users were creating a tag. To ensure this, when time-sensitive input was required from the users, pop-up dialog boxes were used [8]. Additionally, they performed as a cue to prompt the user to tag. When the text or audio tagging popups appear, the phone provides haptic feedback to alert the user, who then has the option to either create and post a tag or to touch an area of the screen not used by the popup to hide the tag-creation popup. The system uses two kinds of popups: (1) text tagging popup features a box where users can type their text and then submit it, and (2) audio tagging popup shows a timer (that signals how long they have to record an audio tag), a Record button, and an instruction on tagging. A text tag contains the typed text, the author of the tag and a timestamp. The audio tag is similar except the content of the tag is an audio clip. The tags are automatically synced to the server when the user has finished tagging. The user can also access all their tags in the 'History' area of the app.

2.1 Tagging Frequency and Audio-Tag Length

One aim of this research was to explore how to configure the tagging frequency and audio-tag length for the tagging user. A number of frequencies were used to understand the effect of different time intervals on user perceptions of tagging. Over a four-week period, the initial application was designed and tested iteratively. Each test consisted of a group of five postgraduate students and researchers, who were asked to tag a radio show discussion. Through initial testing, values of 1 min and 2 min intervals emerged as ideal tagging frequencies, and these needed to be tested with other users. The application

assigned the tagging popup frequency/audio-tag length for the user. This was achieved by creating a set of configurations on the server, and when the application connects to the server, the server shuffles through the configurations set and sends the next configuration on the list to the user. Three tagging configurations were designed by the author to test the ideal set of audio-tag length and prompting frequency (how often the application asked them to create a tag) for the study. These parameters were chosen based on feedback received during the design phase, where 1 min (High Frequency, HF) and 2-minute (Low Frequency, LF) tagging frequencies were tested. A third frequency, in between, of 90 s (Medium Frequency, MF) was also used. As the frequency was decreased, the tag-length was increased proportionally to reflect the extra time that was being afforded to the user to reflect on their tag (HF: 5 s, MF: 10 s, LF: 15 s).

3 Deployment

The next phase involved testing CT with a community who would engage with audio content and would potentially be interested in the audio tagging concept. The author used an opportunistic sampling approach [17] to contact 35 individuals from 3 different Christian communities in the UK. 16 individuals opted to take part in the study (9 males and 7 females; aged between 20 and 32; $M = 23.13$, $SD = 3.07$, Response rate = 45.7%). Individuals in the communities are already encouraged to consume and engage with audio content, particularly in the form of sermons and panel discussions, and thus readily expressed interest in taking part in this study. The deployment was conducted over a two-week period, and a between-subjects study design was used. The participants were informed that their tags would be used by other listeners of the show in the future to find summaries of different sections of the show. This was done based on previous feedback in the design informing workshops, where several attendees stated that knowing how their tags would be used would motivate them to create the tags. The participants were entered into a raffle to win one of ten £20 vouchers for taking part in the study.

Participants were asked to listen to and tag a Christian audio show using the CT application, around the theme of Christian Vocation, lasting 45 min. The show was titled 'Redefining Work Panel Discussion' from the TGC13 Faith at Work Post-Conference [9]. The panel had 5 speakers who were having an unstructured free-form conversation about the concept of 'calling' or vocation, and how that impacts the way Christians work. The speakers were topic experts, but the conversational style of the discussion was informal.

After using the application, the participants were asked to fill in an online survey. The survey contained 4 sections: (1) general questions about their current audio listening habits, (2) questions about the tagging prompts and tag lengths, and how they would configure that experience, (3) general usability questions about the CT application itself, and (4) concluding questions about their likes/dislikes about the application and other comments. Semi-structured interviews were conducted with the participants to ask them about their experience. The interviews were audio-recorded, transcribed and thematically analyzed [3].

4 Findings

16 participants used CT to listen and tag the aforementioned show. 202 tags were generated in total using the application (152 audio tags, 50 text tags). Based on the usage statistics, post-usage survey, and interviews that were held, the themes that emerged are presented below.

4.1 Tagging-Prompt Frequencies and Tag Creation

The data from the participants shows that those in the HF tagging group created many more tags than those in the LF tagging group. An individual in the HF group created on average 4.67 audio tags manually and 8.17 tags through the tagging-prompts issued by the application every minute. The MF tagging group individuals however, still created 4.67 manual audio tags per person but the tags recorded as a response to application prompts fell to 5.17 tags on average, per person (39.24% of prompts shown were responded to). Furthermore, the LF tagging group's prompted tagging rate was similarly low at 2.75 tags per person.

In the survey, participants expressed mixed opinions about the tagging frequencies assigned to them. 3 out of the 6 HF participants felt, that the application prompted them too frequently to keep up. The other 3 felt that the frequency was appropriate. Similar results were received from the 6 MF participants: 3 of them felt the prompts were too frequent, 2 of them felt it was appropriate, and another was undecided. Participants were also asked whether they would prefer to be prompted to create tags, or would they rather create tags themselves without external prompts. Five out of the six HF taggers expressed a preference for manual tagging. In the MF tagging group, 4 participants wanted to only create manual tags, while one wanted to be prompted, while yet another wanted both types of tagging to be present. Finally, in the LF tagging group, 3 persons desired both kinds of tagging-creation to be present while one participant opined that they preferred manual tagging.

Those that expressed a desire for prompted-tagging justified it by saying that "*its easier and I won't forget to tag*" (Alan). While others who did not want automated prompting said, "*being prompted is quite distracting and disrupts my train of thought*" (Sam). By contrast, those eager for both mechanisms to be present opined, "*I like to be prompted in case I have lost concentration, but like the opportunity to create my own tags so that I can make extra notes*" (Nicola). Interestingly, one user who found the tagging prompts quite disconcerting (and quit the tagging process partly through because '*it drove me nuts'*), stated that they would only want manual tagging because "*it would make it more listenable, but then I'd probably never tag*" (Colin).

It was interesting to note that no single person was fully satisfied with the parameters of tagging set in CT. Regardless of the tagging group assigned to them, a lot of the users would have preferred a longer tag length. This was expressed by all but seven of the participants. Three users (one from the HF tagging group, and two from the MF group) expressed a desire for an "unlimited" tag length, so as to "*be concise when I need to be, but also not be restricted by an arbitrary time limit*" (Aisha). Others were a bit more cautious, and stipulated that a tag "*shouldn't be longer than 25–30 s region else it becomes an audio show in itself*" (Sam)". Another feature that was

requested in the post-show discussions, was the ability to pause a show while recording a tag. The application was built with the idea of continuous playback while tagging, which was meant to encourage rapid tag-creation as slowness in the process penalized the user by making them miss part of the show, which would be playing at a reduced volume in the background as they recorded a tag.

4.2 Text Tags vs Audio Tags

There was a diversity of opinion among the participants regarding their choice for text or audio. 5 users did 60% or more of their tags using text, while 9 users did 60% or more of their tagging via audio. Text tagging was not very popular, with only 50 of the 202 tags created using the system being in written format. A significant portion of these 50 text tags were done by a few individuals (68% of text tags were done by 3 individuals).

Furthermore, the tags were analyzed to see how the tag format affected the verbosity of the tags. To enable this comparison, all 152 audio tags were transcribed. The results of the comparison between audio and text tags is shown in the number of words used in the audio tags outnumbers the number of words used in the text tags. The transcribed tags were thematically analyzed, which resulted in a qualitative understanding of the tag contents. While most of the tags were descriptive tags that summarized current conversation, they were not succinct. An example audio tag that matches this description is given here (emphasis has been added to indicate words that are not necessary): "***the guy says about*** *language plays* ***such*** *a big part* ***for him***". (Anita). Some utilized tagging as more of a personal reflection and note-taking tool. Only 2% of the tags were self-reflective in nature. One user's tag contained: "need to teach people to love and interpret other people's actions as love instead of a negative way **and I personally might need to change my personality and working habits when dealing with customers or others**." (Andrea, emphasis added).

4.3 Tagging Motivations

An interesting theme that was uncovered lay around the tension between the perceived benefits and drawbacks of the experience of using CT. Many enjoyed the content of the show they listened to using the application, others stated that they didn't enjoy it as much and that this contributed to their tagging motivation being lower. Many tried to maintain a balance between the extra effort that they felt was required by them to create tags while listening, and the ability to reflect which they felt they received as a result. One participant stated they would have been more proactive in making tags '*if it was a topic I was a bit more personally interested in*' (Amit). Another person added that despite liking the quality of the show, their dislike of discussion panels in general meant they '*lost interest as the podcast went on*' (Alex).

Occasionally, participants were motivated despite their non-favorable perception of the content. Omar, for instance, created 8 audio tags (5 manually created, 3 when prompted to by the application). Conversely, Colin had a poor tagging experience and said that though he wanted to listen to the entire clip, "*the constant interruptions drove me nuts so I had to quit"*. Furthermore, he added that the concept of tagging is not very useful for someone like him who can '*only think about doing one thing at a time*'.

This, he explained, meant that tagging was '*not very good if you actually want to enjoy the listening experience*'. In one sense, this was backed up by Judith, the most prolific tagger (author of 29 tags), who stated in an interview that she didn't feel like she '*listened to a full show*' because *'it was broken down a lot*' by the prompts.

The survey results suggest that many participants felt that tagging required extra effort, as opposed to just passive listening. When asked the question, 'Did the process of creating tags require extra effort on your part?', one participant responded that it did require more effort than merely listening to the audio since '*tagging is an active process*'. They added: '*getting used to the new system also took a bit of time and effort.*' (Aisha). Others praised the positive side of this tagging demand from the application, stating that it helped them '*stop and think about what I was hearing*' (Sofia). Alan quipped that the tagging prompts '*presented an opportunity to articulate and summarize the subject matter*' and tagging '*enhanced the experience rather than having a negative effect*'.

5 Discussion

The results highlight that many participants felt that CT prompted them to tag too often, and that given the choice they would reduce the frequency of tagging prompts. Through content analysis of the audio and text tags, it was apparent that many tags that had timestamps near each other, were similar in terms of the gist of their tag, even if the wording of the tags varied. Building on Chiu et al.'s work [6], more work should be done to explore how context-aware content tagging can be done. For example, the tagging prompts could be based on cues in the audio e.g. pauses, change of speaker, advertisement breaks etc. The timer-based prompts, while appreciated by some users, were not preferred by others. This raises the question of whether more should be done to train users to increase their tolerance for 'still-evolving' speech-based mobile applications, as recommended by Singh et al. [19].

Another design consideration that has arisen from this study is the need to spread the content tagging across the tagging users, for example, by distributing different sections of the chat-show to different users to tag. Doing this would allow users to listen to most of the show without feeling obliged to tag, until their tagging contributions are required. On the other hand, cues in the audio could also be used to divide the content into separate sections. Tagging distribution could be particularly relevant if a third type of tagger (besides the audio and text-based ones) is utilized: a *section tagger*. This could be a tagging role which involves marking the start and end of segments (or points in the show when new themes/topic are introduced). The role would be appropriate for those who want to tag but are unsure where to start, or do not feel confident creating text/audio tags. This role can be deployed either during the real-time show listening stage, or during the post-show content playback stage. Once this tagger has adequately segmented the show into distinct sections, the system could simply ask for one tag from an audio or text-based tagger that best summarizes the content within. This would allow the app to prepare the user adequately in advance for when they need to be ready for interruptions, possibly increasing tagging usability.

5.1 The Purpose of Tagging

Despite the fact that participants were instructed to tag for content organization and ease of information retrieval, they started using the tagging feature in novel ways. Many saw CT as a tool to create personal notes and annotations for audio content. As stated in the literature, user motivations and incentives '*may influence the resultant tags*' [13]. This might be a reason why some chose to create long, self-reflective tags. This was also apparent in the post-usage feedback from some participants, where the positives of the system were framed against how they benefitted from it and then negative feedback was around what they wished the system would be able to do for them. Thus, only a few of the incentives and motivations for tagging mentioned in the literature [10, 11], were observed in the participants. This might be due to previous work focusing on the tagging of entire content (particularly music) and often conducted in non-blind tagging scenarios (i.e. user can see tags created by other users and may use tags as a way of social signaling). Future work needs to look at the possibility of training CT users, as it might be beneficial to undergo some tagging exercises to familiarize them with the system so that user expectations are matched with system capabilities. Frameworks such as Kustanowitz et al.'s annotation framework [10] need to be used to try different annotation technologies to lower the effort-barrier for users who found tagging difficult and to encourage them 'to spend time adding rich metadata'.

The design implications proposed in the literature, that when designing for mobile tagging applications, multiple modalities (i.e. both text and audio) should be presented to the user [4], still hold true. The design, content and community of a platform can influence the motivations of a user and affect the tags that are created [23]. In future studies, further comparisons need to be done on how each of these favors the annotation of audio content.

6 Conclusion

In this paper, a mobile application (CT) was designed, developed and deployed to assess experiences of audio and text-based social tagging of chat-show audio. The findings suggest that there is still much work to be done to refine the tagging experience for the user. Although plenty of rich, descriptive tags were generated using the CT application, majority of the participants saw tagging as a subjective experience and wanted to customize one or more of the tagging parameters like tag length, tagging frequency etc. Further work needs to be explored to identify the design principles for a generalizable audio-tagging interaction that also intelligently prompts the user for tag contributions, as the findings presented here are preliminary in nature due to the small sample size and exploratory study design.

References

1. Anguera, X., Xu, J., Oliver, N.: Multimodal photo annotation and retrieval on a mobile phone. In: Proceeding of the 1st ACM international conference on Multimedia information retrieval – MIR 2008, p. 188 (2008). https://doi.org/10.1145/1460096.1460127
2. Azenkot, S., Lee, N.B.: Exploring the use of speech input by blind people on mobile devices. In: Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility – ASSETS 2013, pp. 1–8 (2013). https://doi.org/10.1145/2513383.2513440
3. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006)
4. Cherubini, M., Anguera, X., Oliver, N., De Oliveira, R.: Text versus speech : a comparison of tagging input modalities for camera phones. In: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1:1–1:10 (2009). https://doi.org/10.1145/1613858.1613860
5. Chiu, P., Kapuskar, A., Reitmeier, S., Wilcox, L.: Room with a rear view. Comput.-Support. Coop. Work Room **7**, 48–54 (2000)
6. Chiu, P., Kapuskar, A., Wilcox, L., Reitmeier, S.: Meeting capture in a media enriched conference room. In: International Workshop on Cooperative Buildings, pp. 79–88 (1999)
7. Font, F., Serrà, J., Serra, X.: Audio clip classification using social tags and the effect of tag expansion. AES 53rd International Conference on Semantic Audio, pp. 1–9 (2014)
8. Jedrzejczyk, L., Price, B.A., Bandara, A.K., Nuseibeh, B.: On the impact of real-time feedback on users' behaviour in mobile location-sharing applications. Computer **38**(12), 14:1–14:12 (2010). https://doi.org/10.1145/1837110.1837129
9. Keller, T., Doll, B., Alsdorf, K.L., Kiersznowski, D., Forster, G.: Redefining Work (Panel) (2013). http://resources.thegospelcoalition.org/library/redefining-work-panel-discussion-tim-keller-bob-doll-katherine-leary-alsdorf-greg-forster-dave-kiersznowski. Accessed 18 Aug 2016
10. Kustanowitz, J., Shneiderman, B.: Motivating annotation for digital photographs: lowering barriers while raising incentives. HCIL-2004-18 (2004)
11. Lamere, P.: Social tagging and music information retrieval. J. New Music Res. **37**(2), 101–114 (2008). https://doi.org/10.1080/09298210802479284
12. Lee, D., Hull, J.J., Erol, B., Graham, J.: MinuteAid : multimedia note-taking in an intelligent meeting room. In: IEEE International Conference on Multimedia and Expo (ICME) (2004)
13. Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, pp. 31–40 (2006)
14. Moran, T.P., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., Van Melle, W., Zellweger, P.: "I' ll Get That Off the Audio": a case study of salvaging multimedia meeting records. In: Conference on Human Factors in Computing Systems, pp. 202–209 (1997)
15. Oviatt, S., Cohen, P.: Perceptual user interfaces: multimodal interfaces that process what comes naturally. Commun. ACM **43**(3), 45–53 (2000). doi:10.1145/330534.330538
16. Sack, H., Waitelonis, J.: Integrating social tagging and document annotation for content-based search in multimedia data. In: CEUR Workshop Proceedings p. 209 (2006)
17. Sada, A.N., Maldonado, A.: Research methods in education. Sixth Edition - by Louis Cohen, Lawrence Manion and Keith Morrison. Br. J. Educ. Stud. **55**(4), 469–470 (2007). https://doi.org/10.1111/j.1467-8527.2007.00388_4.x
18. Schaffer, S., Schleicher, R., Möller, S.: Modeling input modality choice in mobile graphical and speech interfaces. Int. J. Hum. Comput. Stud. **75**, 21–34 (2015). doi:10.1016/j.ijhcs.2014.11.004

19. Singh, A., Larson, M.: Narrative-driven multimedia tagging and retrieval: investigating design and practice for speech-based mobile applications. In: SLAM@ INTERSPEECH, pp. 90–95 (2013)
20. Turk, M.: Multimodal interaction: a review. Pattern Recogn. Lett. **36**(1), 189–195 (2014). https://doi.org/10.1016/j.patrec.2013.07.003
21. Yadati, K., Chandrasekaran Ayyanathan, P.S.N., Larson, M.: Crowdsorting timed comments about music: foundations for a new crowdsourcing task. In: CEUR Workshop Proceedings, p. 1263 (2014)
22. Yew, J., Gibson, F.P., Teasley, S.: Learning by tagging: the role of social tagging in group knowledge formation1. CEUR Workshop Proceedings, vol. 312, pp. 48–62 (2007)
23. Zollers, A.: Emerging motivations for tagging: expression, performance, and activism. WWW (2007). https://doi.org/10.1.1.118.7409