Designing Immersive Audio Experiences for News and Information in the Internet of Things using Text-to-Speech Objects

Mark Lochrie, Robin De-Neef, John Mills, Jack Davenport
Media Innovation Studio
University of Central Lancashire
PR1 2HE
{mlochrie, rde-neef,jmills,jdavenport1}@uclan.ac.uk

The use of personal Internet-enabled objects is on the rise. The use of voice interactions and audio-based delivery of content are the preferred approach for these screen-less objects. In terms of the news industry, Text-To-Speech (TTS) is an obvious choice for repurposing readily available written articles of text for such platforms. Currently through flash briefing skills on services such as Amazon Echo, news publishers are seeking innovative new ways to distribute their content to new audiences. This paper continues the research of NewsThings and uses one of its prototypes, RadioThing, to evidence the problems that are currently present in TTS, how these challenges affect publishers and outlines seven recommendations when designing TTS objects for news and information content. These recommendations are suggested to improve the authenticity, immersion and production of audio-based journalism. Consequently, prompting the use of TTS as a viable form of media for local and regional news organisations.

News, Objects, Text to Speech, Alexa, IoT, Speech, Interaction, Media

1. INTRODUCTION

This position paper explores the use of TTS in IoT to deliver news and information. With the recent advances of technology and machine learning, TTS services now provide content designers with a new means of communicating their message (Moon, 2016). However, human-sounding voices reading textual information to users is not yet perfect. The following article addresses the uses and problems of TTS and outlines some key recommendations when designing news and information content for TTS.

Although incorporating user-generated content has been a rising trend for over a decade (Thurman, 2008) news organisations are typically set up to broadcast in one-way interaction models. Publishers are being forced to adapt to seemingly ever-increasing technological advances in order to survive in the changing landscape of news and information medias. Engagement with quality journalism was on the decline amongst younger demographics and no clear solution is on the horizon (Kobo, 2017; Lopatovska 2018). Smart speakers present a possible solution for addressing this challenge as they offer new methods for audience engagement. Publishers are beginning to explore voice-controlled devices and how they could be used to broadcast news in novel ways to reach news audiences. This is underpinned by journalism searching and exploring for new and legitimate revenue streams in the digital age in light of these changes (Picard, 2014).

2. BACKGROUND

During the most recent News:rewired conference (March 2018) Robert McKenzie, editor of BBC News Labs, presented findings from their work with voice and TTS. Of particular interest were the findings of TTS synthesis (Ciobanu, 2018), “editors are not dependant on studios or recording equipment, single editor produces multiple voices - scalability “. McKenzie highlights (Gabay, 2018a) how these devices should not be dismissed as potential futures for the sector as well as what challenges persist for news outlets when launching their own news platforms. He argues that users could find the novelty of consuming news from voice-commands be less intrusive “It requires a lot less effort to interact with something verbally than reading it,”, it’s on-demand (unlike search or browsing the internet) you are presented with news directly, current analysis of the uses of voice and audio-based journalism have shown positive results. “the week-on-week growth has been really impressive for us. Stats have also shown that the audience is loyal.” Rob Owers, The Telegraph (Gabay, 2018a). Mckenzie (Gabay, 2018a) does, however, highlight the challenges faced, infancy of such platforms, natural language processing and discoverability and interaction to name a few. Unfortunately, for this

sector, people are not using these smart devices for news (Edison 2017; Lopatovska, 2018). "There's quite a lot of work that we need to do to come up with, propositions that will be really engaging and exciting and will make people want to consume their news on voice-controlled devices." Robert McKenzie (Gabay, 2018a). Publishers may need to capture audiences the moment they set up these smart devices as research shows owners set up their devices with the content they are familiar with and tend not to update their preferences (Lopatovska, 2018). Furthermore, the use of audio-based journalism to engage audiences, peaks at 7am (weekdays) and 8am (weekends) for the breakfast hour and similarly with early evenings (6pm) (Gabay, 2018a; Scribblelive, 2018). These trends show the spikes for when people are most likely to consume such content (Edison, 2017; Lopatovska, 2018). Unfortunately, for local and regional news outlets, creating content creates resourcing challenges some are not able to meet.

In order to address the problems and uses of TTS, our research is grounded in the case study of current research into NewsThings (Mills, 2017). This project explores how news and information content can be delivered to users through individually crafted IoT objects. These objects will provide the foundations into understanding how users experience objects within their own home, receive and convey emotion and meaning from them. It also aims, from postmodern and post-structuralist perspectives, to understand how connected objects may create their own cultural syntax and agency, and how this agency may impact on content and interactions with it.

This paper's contribution is grounded in the creation and study of the use of TTS content used within RadioThing. A news object from NewsThings (Mills, 2017) developed in order to provoke and evaluate the current issues when designing for TTS. Exploring the current issues with TTS for screen-less devices to deliver news and information content and what design characteristics should be considered when creating audio packages to distribute news content over the IoT? Prompting discussions around the duration of audio items and how to mitigate the meta from being read out (i.e. image copyright), limits on translated text, mis-pronounced words and the lack of regional accents. Through its iterative development, various TTS engines were adopted to test their authenticity of speech. Amazon's Polly was adopted due to its natural sounding voices and customisability. However, this approach did not deliver the type of experience imagined when the project was first envisioned. This was partly due to when testing these engines with short example phrases listeners did not experience the real world uses of the platform. When large amounts of text is converted to TTS, users exhibit traits of boredom, losing interest and concentration. Due to the nature of synthetic sounding voices. Our research explores what challenges persist in TTS and what do users require from internet enabled objects when delivering content as audio. Until machines can perfect human voice, we will continue to observe these problems.

3 DESIGNING CONTENT FOR TTS

McLuhan suggests that the medium on which media is conveyed affects the human sense of the message in contains. The message is therefore important when designing human interactions with TTS over the IoT. As such, RadioThing will be deployed to engage audiences with news content as an intervention to study IoT and news and generate further insights for the news industry as it explores IoT as a content dissemination platform. As RadioThing uses audio as the chosen medium, work was required in order to fully understand the issues of repurposing text articles into streams of audio. This paper outlines the development of a hybrid approach to TTS (Fig. 1), which has been designed using the same methodology as outlined for the project. The hybrid audio output was designed to tackle the challenges of TTS and studied in this paper. The approach included identifying key themes for repurposing text articles. In the first instance, it was deemed appropriate to break down the article into sections of statements, background information, quotes and story to use difference voices (Moon, 2016). To create a sense of interviewer and interviewee conversational content (Gabay, 2018b). Furthermore, to bridge the gap between TTS and podcasts, the use of pre-recorded sound bites from the author would be interjected into the content, again breaking down the repetitive nature of TTS. Then through the use of ambient sounds a background soundtrack was applied to create a sense of immersion in the content. Moreover, the use of emphasising summary points generated by textual analysis, highlighted what parts of the story was most important to convey. It is these design characteristics that enabled the researchers to create a hybrid version of the audio content that would be used to compare and evaluate against the full and summarised texts of news articles.

A questionnaire was designed in order to explore the uses and problems that persist with TTS and provided three examples of how TTS audio could be designed. This data was collected using NASA's Task Load Index (NASA, 1986) rating scale to collate evidence how people find listening to TTS audio. The questionnaire was split into eight sections; current TTS/IoT/news applications, interactions with three types of content, personalisation and personal information. In the interaction of TTS content, users were presented

with a selected full article of four minutes in length followed by a summarised one-minute version of a different article and, finally, a "hybrid" solution designed by the researchers. The hybrid content consisted of key design considerations to address the issues observed in the previous two pieces of content. These three articles were studied to investigate the mental and temporal demands (Lopatovska, 2018), frustration, effort and performance of listening to the content. The questionnaire then took the same approaches and investigated which version of the same article provided the most knowledge, yielded the highest level of engagement, was most difficult to follow and what/if any of the designed content in the custom audio created a higher sense of immersion.

At present, 13 respondents from a mixed age group (17-50) engaged with the pre-sampled content in order to address the design considerations of TTS. While realistic, natural sounding voices with intonation and accents were listed as one of the main challenges for TTS, only 27% indicated it as one of the most important aspects of TTS. This could indicate that the human-like voices aren't really a priority for TTS. As predicted the summarised content received the lowest mental demand and the highest overall performance. Furthermore, the rise in IoT use does not correlate with the innovation and applications of content that reside on these platforms. The initial study reinforced this claim, resulting in users rating current TTS as average amongst mobile devices and in-home IoT objects (Alexa and Google Home). Voice interactions were deemed most important mode of interaction and that TTS and audio/podcasts are similar in their importance for minimal screen intersections and screen-less devices.

It was deemed important to understand the motivations of audiences from how they interact and consume news and information content. Participants reported that the use of notification sounds to be used to breakdown content should be considered when delivering multiple pieces or lengthy content as seen in The Inspection Chamber, capturing user interactions (Cooke, 2017).

The type of accents and conversational interviews should also be considered to assist in design of TTS content. Length of content was also a main focus of the study as typically, a news article is written 500 - 800 words (Ferne, 2017). Participants in the study do not think that this evidence relates to audio-based journalism. Personalisation of content also scored highly, with many participants suggesting that they want the ability to store content for when it's convenient to consume, provide a handover from other services and know what they are interested in and make suggestions. It was found that participants do not rate quality of audio to be important. This was surprising as we perceive digital to mean reliable and high definition. This question was asked to explore whether journalists could interject ambient and user generated content into broadcasts. Similar to live breaking events on TV when you see mobile phone footage used or quotes from witnesses as phone call voice messages.

Moreover, when interacting with the samples of audio, the hybrid approach provided the most knowledge (63.6%), had most engaged (72.7%) and was the preferred audio format (54.5%). Whereas the full article was the most difficult to follow (54.5%) due to its duration and linearity of information. The summary method had the lowest mental demand and the highest overall performance. The hybrid method noted the lowest frustration levels and effort. While only having a slightly higher demand and a slightly lower overall performance than the summary. The full article has a significantly higher mental demand, frustration level and effort than the other two which brings down the overall performance. One thing that is interesting with these findings is that temporal demand is the lowest on the full articles and the highest on the summary, even when both voices were configured in the same tempo.

These findings indicate the issues with the current innovations in TTS for IoT, repurposing existing text content that is not fit for the platform and the short briefings do not really offer much in the way of informing audiences. News content that was summarised was less mentally demanding but not as pleasant to hear or gain much more information that the flash briefings. Whereas, the hybrid method seemed more appropriate to listen (due to its nature of breaking down the content into various voices, pre-recorded audio and ambient sounds. Participants indicated that they prefer the hybrid method (54.5%) since it is more engaging and provides additional knowledge whereas the full article is the most difficult to follow. The summary method was the easiest to follow but provided less knowledge, but 36.4% of the participants choose the summary as their preferred method of content delivery. This highlights that content providers need to consider the trade-off between the crafted hybrid method and the summarised content. Whilst the hybrid method might be the preferred method with the most engagement will take more time to create in comparison to the summary. This has prompted the creation of a toolkit for designing for TTS.

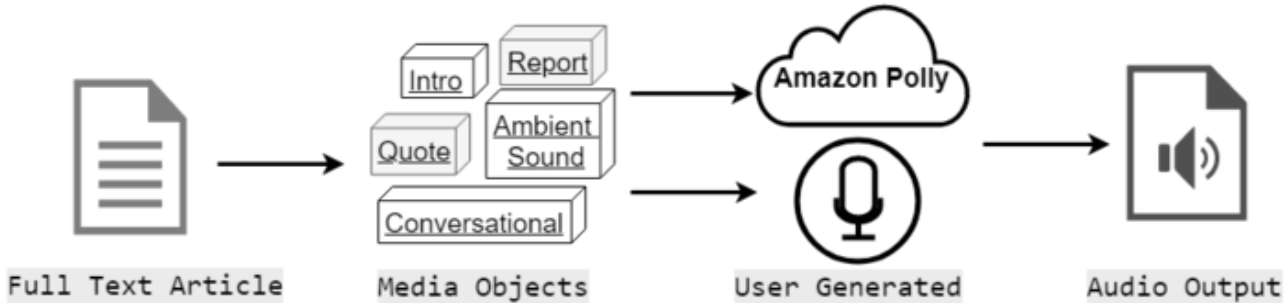

Figure 1. Diagram to show the object based creation of the automated audio from original text article to the final audio output

4. DISCUSSIONS

Through the study and findings obtained, the researchers propose seven key recommendations when designing with TTS for news and information outlets.

1. The creation of object-based bitesize chunks of information. Moving away from linear stories to objects (Cooke, 2017; BBC 2017) of information that could be reused across a range of platforms.
2. Emphasising summary points within an article.
3. The use of specialised actors within stories (Moon, 2016). For example, "over to Rachel with the weather". This could be achieved through assigning certain categories to particular voices. However, this raises challenging questions around authenticity when applied to direct quotation. As long as the narrator clearly defines these voices are not the originals when converting quotes as TTS or should attempt to use recorded audio samples taken from the original source.
4. Utilising platform specific meta tags. Polly supports standard and specific SSML tags
    a. breaths.
    b. 'drc' - mid-range "loudness".
    c. phonation.
    d. vocal-tract-length.
    e. Whispered.
5. Responsive journalism: change writing style for IoT TTS. Writing for purpose whether it's for Twitter, Facebook, a poster or a news article is commonplace. Writing styles need to change to fit the specified platform. The industry has a track record of platform adaptations: radio to TV, print to digital.
6. The use of embedded pre-recorded dialects into content. For example, as we are exploring the use of local and regional news, audiences feel a greater sense of trust and authenticity when 'real' voices are heard.
7. Ambient and environmental sound added into the background audio. Creating immersive content to break down the non-humanistic sounds of TTS.

If journalism is to acknowledge this platform as a potential outlet for news, then trade-offs must be made from an organisational resources perspective, writing responsive content and understanding the end users environments (user's experience). Therefore, the need for further work to understand how journalists could create new pieces of content is needed. However, as indicated by the participants, this new content does not need to be of professional quality (recorded in studio or tweaked for best performance). Therefore, we propose the need for a toolkit to assist in the creation process of designing for audio-based journalism. Bespoke toolkits for journalist are not new (BBC NewsLab, 2016a; BBC NewsLab, 2016b; Norton, 2017), though a platform for crafting content for TTS is novel. The collaborative nature, rich media and text analysis of such toolkit needs further exploration. From the research thus far, we suggest the toolkit is multi-platform, enabling journalists to retrieve the article, highlight sentences for tagging content (with their own audio, user generated audio, environmental sounds etc). The toolkit would then produce meta code which accompanies to the textual article and can be merged as objects when delivering as audio content (BBC R&D, 2017).

Our initial study demonstrates the richness of TTS exploration when placed within a real-world context of both users and industry requirements, and that it could form an intriguing new way of generated news and information among smart home audiences. It also argues that journalists will again need to adapt their approaches for emergent editorial platforms: to think about both the medium and the message.

5. ACHNOWLEDGEMENTS

We acknowledge Google Digital News Initiative for funding the NewsThings project. The users involved in the human cantered design process, participants in the studies presented in this paper and to the partners of the project for their participation and support; Thomas Buchanan's Tom Metcalfe, Peter Bennett, Chloe Foy and David Haylock, and Trinity Mirror's Paul Gallagher, Adam Walker, Alison Gow and Mandy Brain.

6. REFERENCES

BBC NewsLab. 2016a. Alto – A Multilingual Journalism Tool. http://bbcnewslabs.co.uk/projects/alto/ (retrieved on 16 April 2018).

BBC NewsLab. 2016b. Online Content Toolkit. http://bbcnewslabs.co.uk/projects/octo/ (retrieved on 16 April 2018).

BBC, R&D. 2017. Object Based Media. http://www.bbc.co.uk/rd/object-based-media (retrieved on 16 April 2018).

Ciobanu, M. 2018. Slides and audio – Virtual voiceover translation in news production. https://www.newsrewired.com/2016/03/18/slides-and-audio-virtual-voiceover-translation-in-news-production/ (retrieved on 16 April 2018).

Cooke, H. 2017. The Inspection Chamber. http://www.bbc.co.uk/rd/blog/2017-09-voice-ui-inspection-chamber-audio-drama (retrieved on 16 April 2018).

Edison 2017. The Smart Audio Report. https://nationalpublicmedia.com/wp-content/uploads/2017/06/The-Smart-Audio-Report-from-NPR-and-Edison-Research-2017.pdf (retrieved on 11 May 2018).

Gabay, A, L. 2018a. The benefits of using voice-controlled devices for news distribution. https://www.newsrewired.com/2018/03/07/benefits-using-voice-controlled-devices-news-distribution/ (retrieved on 16 April 2018).

Gabay, A, L. 2018b. 'Look for the sweet spot between news and context' – How to make your live streams more engaging. https://www.newsrewired.com/2018/03/07/create-engaging-live-videos-peter-stewart/ (retrieved on 16 April 2018).

Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q. and Martinez, A., 2018. Talk to me: Exploring user interactions with the Amazon Alexa. Journal of Librarianship and Information Science, p.0961000618759414.

Mills, J., Lochrie, M., Metcalfe, T., and Bennett, P. 2018. NewsThings: Exploring Interdisciplinary IoT News Media Opportunities via User-Centred Design. In Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '18). ACM, New York, NY, USA, 49-56. DOI: https://doi.org/10.1145/3173225.3173267.

Moon, Y., Kim, K.J. and Shin, D.H., 2016, July. Voices of the internet of things: An exploration of multiple voice effects in smart homes. In International Conference on Distributed, Ambient, and Pervasive Interactions (pp. 270-278). Springer, Cham.

NASA Ames Research Centre. 1986. Task Load Index.

Norton, A. 2017. From speech to text: four applications of automated transcription in the newsroom. https://medium.com/bbc-news-labs/from-speech-to-text-5fff6abf4df1 (retrieved on 16 April 2018).

Picard, R, G. "Twilight or new dawn of journalism? Evidence from the changing news ecosystem." Digital Journalism 2.3 (2014): 273-283.

Scribblelive. 2018. https://embed.scribblelive.com/Embed/v7.aspx?Id=2754615&ThemeId (retrieved on 16 April 2018)

Thurman, N. "Forums for citizen journalists? Adoption of user generated content initiatives by online news media." New media & society 10.1 (2008): 139-157.