Evaluating speech synthesis in a mobile context: Audio presentation of Facebook, Twitter and RSS

Mathew Aylett[1], Yolanda Vazquez Alvarez[2], Lynne Baillie[3]
[1]*CereProc Ltd. Edinburgh, UK.*
[2]*Department of Computer Science, Glasgow University, Glasgow, UK*
[3]*Glasgow Caledonian University, School of Engineering, Glasgow, UK*
matthewa@cereproc.com, yolanda@dcs.gla.ac.uk, l.baillie@gcu.ac.uk

Abstract. *This paper presents an evaluation of a podcast service that aggregates data from Facebook, Twitter and RSS feeds, using speech synthesis. The service uses a novel approach to speech synthesis generation, where XML markup is used to control both the speech synthesis and the sound design of a resulting podcast. A two-phase evaluation was carried out: 1) participants listening to the podcasts on desktop computers, 2) participants listening to the podcasts while walking. Our findings show that participants preferred shorter podcasts with sound effects and background music, and were affected by the surrounding environmental noise. However, audio advertising which is part of the service did not have a significant negative effect. Another finding was that the advantage of using multiple voices for content segmentation may have been undermined by difficulties in listener adaptation. The work is part of a new approach to speech synthesis provision, where its style of rendition forms a part of the application design and it is evaluated within an application context.*

Keywords. Speech synthesis, evaluation, mobile systems, auditory interfaces.

1. Introduction

The Nooz Client application aggregates feeds from a users news and social networks and plays the highlights to them whilst they listen to music on their mobile phone. In order to do this we use speech synthesis. Speech Synthesis is a key enabling technology for pervasive and mobile interfaces. However, very little previous work has looked at evaluating speech synthesis in a mobile context. Typically, synthesis is evaluated independently, often using a 5-point mean opinion score based on how 'natural' a listener found the speech. Furthermore, very little work has investigated how audio design considerations for a mobile context can support or degrade an audio experience based on synthetic speech. In this paper we present an evaluation of a novel approach to speech synthesis provision, where its style of rendition forms a part of the application design and we evaluate it in context. The evaluation presented us with three significant challenges:

1. The evaluated application needs to have reasonably large chunks of synthesised speech so that the subject can experience the application without too many unnatural interruptions.
2. The evaluation should be carried out in the field and not the lab. This is so that natural conditions apply such as noise from traffic etc. However, carrying out such an evaluation is more difficult than in a controlled lab environment as level of traffic noise, weather etc cannot be known in advance, nor their impact on the users reaction to the application.
3. We wished to test whether the users choice of accompanying music tracks could affect the impression of the synthesised speech content that occurs between the tracks.

We therefore wished to answer the following research questions:
RQ1: What is the user acceptance of the amalgamation of news and social media updates played back on a music player via speech synthesis?
RQ2: How do we evaluate such a application correctly in context?

In conclusion our goal is to produce as good a user experience as possible whilst using our application. We are therefore interested in evaluating ways to improve the acceptance of speech synthesis both by determining which problems are affecting the experience, and how other design considerations can make the best use of the synthesis technology available to us.

2. Background

In the speech synthesis domain, previous research, such as the The Blizzard challenge [4], has focused on the evaluation of synthesis without reference to very specific contexts. In contrast, within HCI there is a tendency to view speech synthesis as a black box and rather than engage with its limitations, instead use pre-recorded prompts or alerts which avoid the use of dynamic content, for example Dingler et al [2] where different non-speech sounds are used to notify a mobile user of the presence of Twitter, Facebook and RSS feeds. This use of non-speech auditory feedback can be very effective. In Nomadic Radio[7] it was reported that the use of non-speech auditory cues was preferred when users were multitasking. However, if we wish to automatically present dynamic content in an audio context, speech synthesis is a requirement. As we will show in this paper, this is not an either/or choice. The use of non-speech audio together with speech synthesis when appropriate is a promising approach, even though it is a challenge to evaluate.

3. System Design

Our application was designed as follows:

- An automatic podcast lasting from 30-60 seconds is generated using speech synthesis from the user's personal online information (Facebook, Twitter, selected RSS feeds).

- The podcasts are downloaded to a mobile device, and, when possible, stored and played between audio tracks while the user is listening to his personal music.

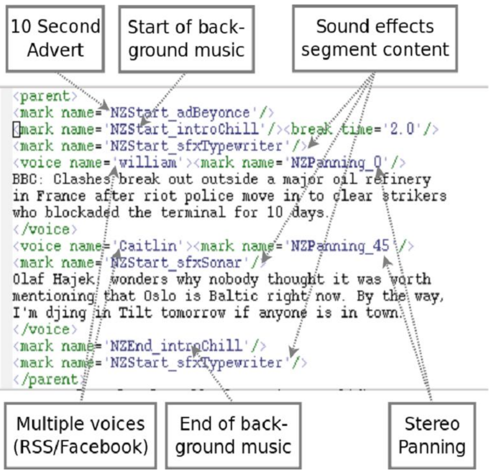
- Audio advertising is used to support the podcast service, background music and sound effects are added to improve the audio experience, and multiple synthetic voices, together with stereo panning, is used to support content segmentation. For example, a voice placed at a location in the panning space might render RSS news compared to a different voice at another location for Facebook status updates.

Previous applications have used synthesis to present such aggregated online information. Twuner http://mashable.com/2009/08/11/twuner) for example has an audio twitter feed for iPhone, and TweetRadio. Twuner did not use non-speech audio to create a richer audio experience and was viewed very much as a 'just being a spoken Twitter feed'. TweetRadio in contrast separated tweets with radio like noise, to give the effect of turning a radio dial, creating an unusual and pleasing audio effect.

3.1 Speech Synthesis

The engine used in the application presented here was developed by CereProc Ltd.[1] and uses a concatenative approach[3]. The CereVoice front end takes text and generates a series of XML objects we term spurts. The spurt is a section of speech surrounded by pauses. XML markup can be inserted into the input text and is maintained in the spurt output. The CereVoice system allows for a very wide variety of XML markup to control synthesis. Industry standard SSML markup is converted by the front end into a 'reduced instruction set' of XML with a clear functional specification.

In addition, a set of XML markup can change the selection process in the system, for instance the ability to alter pitch targets. Tags used to alter selection are used in conjunction with tags which cause a change in the speech using digital signal processing to create different speech styles. Tags can also be used to add non-speech audio content and control stereo panning. Figure 1 shows an example of the XML input required to generate a short podcast using multiple synthetic voices, including adverts, background music and audio sound effects.



```
<parent>
<mark name='NZStart_adBeyonce'/>
<mark name='NZStart_introChill'/><break time='2.0'/>
<mark name='NZStart_sfxTypewriter'/>
<voice name='william'><mark name='NZPanning_0'/>
BBC: Clashes break out outside a major oil refinery
in France after riot police move in to clear strikers
who blockaded the terminal for 10 days.
</voice>
<voice name='Caitlin'><mark name='NZPanning_45'/>
<mark name='NZStart_sfxSonar'/>
Olaf Hajek: wonders why nobody thought it was worth
mentioning that Oslo is Baltic right now. By the way,
I'm djing in Tilt tomorrow if anyone is in town.
</voice>
<mark name='NZEnd_introChill'/>
<mark name='NZStart_sfxTypewriter'/>
</parent>
```

10 Second Advert
Start of back-ground music
Sound effects segment content
Multiple voices (RSS/Facebook)
End of back-ground music
Stereo Panning

Figure 1. XML control language for rich audio podcast using CereVoice.

The emphasis of the CereVoice system is to create a voice with character, one that sounds less neutral, and is more appropriate for listening to content in a recreational environment. The result we hope is natural and engaging synthesised speech (see the online demo at http://www.cereproc.com, for examples).

4. Methodology

The objective was not to evaluate the speech synthesis in isolation, but to evaluate it within the application context, thus we investigated:

- To what extent different synthetic voices and different podcast content affect user experience?
- Does the addition of stereo panning, together with sound effects and background music improve the experience?
- Can multiple voices be used to help content segmentation and produce a newsroom feeling?
- Does the type of music played before and after the news feeds affect the user experience?
- How does the user response differ when listening to the podcast in a static environment, compared to listening on the move, outside, on a mobile device?

4.1. Materials

Participants listened to four synthesised podcasts surrounded by faded in and faded out music tracks (30 seconds in length). Four podcasts were created:

Long: Two podcasts were approximately 1 minute long and contained 2 BBC RSS feed headlines, 1 Facebook item, and 2 Twitter Feeds from the US movie critic Roger Ebert. Two short podcasts that were approximately 30-seconds and shortened by removing the Twitter information.
Then each podcast was altered to be:
- With and without a 10 second advert or short sponsorship message.
- With and without multiple voices and audio effects. For example: a typewriter noise to signal the start and end of each podcast, a sonar noise to signal Facebook information.

This resulted in 4 varieties of each podcast giving 16 different audio files.

4.2 Participants and Contexts

We carried out the evaluation in two phases. First, a static evaluation using participants from Mechanical Turk, a crowd sourcing service. Second, in a mobile environment while walking in a nearby park, using participants studying and working at The University of Edinburgh, UK. This two-phase approach had the advantage of offering a fast, resource efficient evaluation which allowed more subjects, while at the same time allowing us to validate those results with a more resource intensive in-field evaluation.

Participants in both phases were presented with the same materials and questions:
1. Initial questions on age range, gender, and how often they listened to music on digital devices.
2. Asked to listen to four audio tracks, created from a podcast sandwiched between four different musical styles (Classical, Pop, Motown, Salsa). After listening to each audio track they were then asked to:

- Fill out a multiple choice content question to ensure they had listened to the audio, followed by a feedback score for each podcast between 1 (hated it) to 100 (loved it).

3. After listening to all of the podcasts they were then asked to comment on the adverts and sponsorship, for example did they find them annoying? Did they like multiple voices in the same podcast? Did they like sound effects and background music in the podcasts?
4. Finally they were asked to provide any informal feedback.

4.3 Static Context Evaluation

In Mechanical Turk [5] different materials form different Human Intelligence Tasks (HITS). We required 16 HITS, each with 2 participants (total of 32), all residents in the United States. The experiment could not be done more than once. After each participant had completed the task, the answers to the content questions were checked. Only one subject failed to answer these questions correctly and was rejected and replaced with another. On average, the task lasted 10 minutes. Each participant received $1. No hearing difficulties were reported.

4.4 Mobile Context Evaluation

Eight participants took part in this phase. All questions were presented on a clip board held by the experimenter. The four audio tracks were loaded into a media player on a Nexus 1 phone running Android. After answering the initial questions, participants walked in a “calm and relaxed manner” in a nearby park (George Square Gardens), while listening to each audio track on Sennheiser H350 headphones, see Figure 2. The experimenter accompanied each participant.

Figure 2. Setup for mobile context experiment.

After each audio track the subject paused and answered questions for each podcast. Finally each participant answered the follow-up questions. No hearing difficulties were reported.

5. Results

5.1.Static Context Results

Participants answered content questions almost 100% correctly (n=5/132 errors, including the rejected participant from the static evaluation) confirming they had listened to the audio. The feedback scores varied widely, for example 3,10,3,8 for one participant who commented “Disjointed, unpleasant voice, just noise without any information or entertainment”, to 80,75,60,80 for another who commented “The music doesn’t fit the podcast”. The median across all feedback scores was 60 - a mild positive preference across all podcasts. We carried out a non-parametric Friedman test of the feedback scores, applied to three different groupings: podcast length (long, short), podcast rendition (+/- adverts, +/- multiple voices and sound effects), and by the music type in the podcast (Classical, Pop, Motown and Salsa).

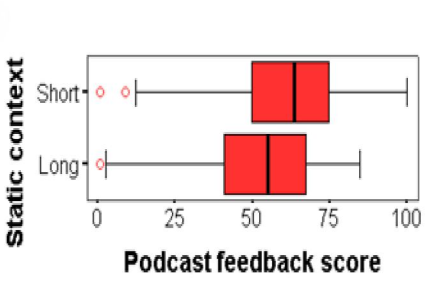


Figure 3. Variation in podcast feedback (1 = Hated, 100=Loved) by podcast length, (Long = 1 minute, Short = 30 seconds.

Results showed a significant result for podcast length, with the 30-second podcasts receiving a higher ranking than those 1-minute long (-2=12.448, df=1, p<0.001, N=32), see Figure 3. There was no significant result for audio rendition.

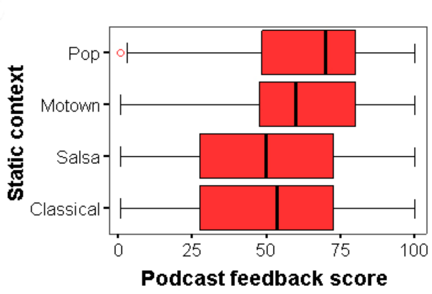


Figure 4. Variation in podcast feedback caused by surrounding music type.

There was also a significant effect by music type. Podcasts surrounded by pop and Motown music were ranked higher than those surrounded by salsa and classical music (2=8.546, df=3, p=0.036, N=32), see Figure 4. However a pair-wise Wilcoxon signed ranks test only showed a significant difference between salsa and Motown music after Bonferroni correction (Z=-2.775, p=0.036).

In contrast to feedback scores, follow up questions showed a very high expressed preference for podcasts which contained background music and sound effects (n=28/32, 87.5%, sign test:p<0.001). However, participants were split on the use of multiple voices (n=17/32, 53% preferred single voice renditions, sign test: not significant), and although a majority did not find the adverts and sponsorship annoying this was not significant (n=21/32, 66%, sign test:p=0.0551).

5.2 Mobile Context Results

Participants who walked in a park while listening to the podcasts had much more difficulty answering the questions on content (n=10/32 errors, 69% correct). However, on examination it was noted that most of these errors were for the first podcast suggesting participants found it hard to get used to the listening environment and the dual tasks of walking and listening. Variation in the feedback scores was less dramatic than in the static evaluation, but the median of scores was the same (60 - a mild positive preference across all podcasts). Participants reported that it was hard to concentrate on the speech part of the podcasts while walking.

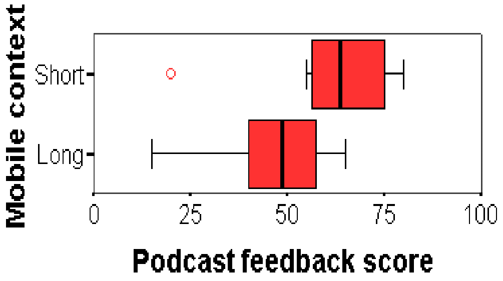


Figure 5. Variation in podcast feedback (1 = Hated it, 100=Loved it) by podcast length (Long = 1 minute, Short = 30seconds).

Friedman tests over the same three groupings as in the static context, showed similar results for podcast length, with a preference for the shorter podcasts (2=4.5, df=1, p=0.034, N=8), see Figure 5. No significant effect was found for music type although with only 8 participants it was not possible to cover all content music combinations so this result could be due to the smaller sample size.

Results to follow-up questions were also similar to the static context results. A high preference was expressed for background music and sound effects (n=8/8, 100%, sign test: p<0.001), no clear preference for multiple voices in the same podcast (n=4/8, 50% preferred single voice renditions, sign test: not significant), and no clear dislike of adverts (n=4/8, 50% did not find adverts and sponsorship annoying, sign test: not significant).

6. Discussion and Conclusions

Addressing our research questions, the two phase evaluation strategy was successful. Results matched closely enough between the static and the mobile contexts to support the use of Mechanical Turk for further refinement of the podcast application, while reserving the resource intensive mobile context evaluation for confirmation of more finalized designs. The three most important results were the confirmation that background music and sound effects improved the audio experience, that the style of advertising (10 seconds or so) did not have a significant negative effect on the results, and that participants preferred the shorter (30 second) podcasts. Furthermore these results held for both static and mobile evaluations.

The results for multiple voices were also interesting. It has been shown that listeners adapt to new voices [6], finding it easier to understand a voice the longer they are exposed to it. However, our participants did not have a chance to adapt to the different speakers. Thus, although multiple voices could help content segmentation, multiple voices also put more strain on the listener. Given this, and the different reactions of individuals to different voices (e.g. “male voice clearest”, “was a little hard to understand the accent”), allowing users to customise the service to use different voices as well as the number they wish, is an important design requirement.

Ultimately, to effectively implement interfaces based on speech synthesis, the audio presentation should respond to design requirements. Speech synthesis should not be a box you bolt on with preconceived constraints. Customising technology for individual applications is not an indication of a failure of that technology, rather it is an indication of its flexibility and utility.

In relation to the size (2 participants in the static evaluation and 8 in the field) and length (short period of time that they listened to the podcasts) of the study there are off course limitations as to the strength that can be ascribed to the results. We were aware of this weakness and as a result have undertaken a further study in which the participants use the Nooz Client application over the course of a week.

7. References

[1] Aylett, M. P., and Pidcock, C. J. The CereVoice characterful speech synthesiser SDK. In AISB (2007), 174–8.

[2] Dingler, T., Brewster, S. A., and Butz, A. Audiofeeds – a mobile auditory application for monitoring online activities. In ACM Multimedia (2010).

[3] Hunt, A., and Black, A. Unit selection in concatanative speech synthesis using a large speech database. In ICASSP, vol. 1 (1996), 192–252.

[4] King, S., and Karaiskos, V. The blizzard challenge 2010. In The Blizzard Challenge 2010 workshop (2010).

[5] Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with mechanical turk. In Proc. CHI '08 (2008), 453–456.

[6] Nygaard, L., and Pisoni, D. Talker-specific learning in speech perception. Attention, Perception, & Psychophysics 60 (1998), 355–376.

[7] Sawhney, N., and Schmandt, C. Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. In ACM Trans. On Computer-Human Interaction, vol. 7(3) (2000), 353383.