Designing a Synthesized Content Feed System for Community Radio

KRISTEN M. SCOTT, Madeira Interactive Technologies Institute, Portugal
SIMONE ASHBY, ITI / LARSyS, Portugal
ADRIANA STAN, Technical University of Cluj-Napoca, Romania

The use of text-to-speech to generate radio content is largely unexplored, despite the importance of radio in remote parts of the world, where TTS offers a robust means of transforming data into media for low-literate audiences and those without regular internet access. How suitable are TTS voices for meeting the expectations of radio listeners and what type of content are these voices best suited to deliver? We present an application for generating automated daily synthesized weather forecasts for selected locations and language varieties, based on the provision of a regularly updated weather data service. We present results from a pilot listener study aimed at exploring people's reactions to this and other synthesized audio content, as we begin to explore best practices around the design of a synthesized content feed system for community radio.

CCS Concepts: • **Human-centered computing** → *Empirical studies in ubiquitous and mobile computing*; **Accessibility technologies**; • **Applied computing** → *Media arts.*

Additional Key Words and Phrases: synthetic speech; text-to-speech; voice; radio; Romania; ubiquitous computing

ACM Reference Format:
Kristen M. Scott, Simone Ashby, and Adriana Stan. 2020. Designing a Synthesized Content Feed System for Community Radio. XX, X, Article XXX (X 2020), 6 pages. https://doi.org/XXXXXXXXX

[Good morning] [location]. The forecast for this **morning** from **[6:00]** is **[clear sky]**, with a temperature of **[10]** degrees, with wind of **[8,2]** meters per second in the **[easterly]** direction. The forecast for the **[afternoon]** from **[12:00]** to **[18:00]** is **[cloudy]**, with a temperature of **[10]** degrees, with wind of **[5,8]** meters per second in the **[south-west]** direction. Weather forecast provided by YR.no application, the Norwegian Meteorological Institute and NRK.

1 INTRODUCTION

As DIY synthetic speech voices become easier to generate - for example using free, open source toolkits such Idlak [8] - we can expect the use of text-to-speech (TTS) to expand to a greater number of use cases over the coming years. Through our interactions with Alexa, Siri and other personal digital assistants, we are gradually becoming aware of some of the possibilities and limitations of interacting with high-quality TTS voices [1, 6, 7] and the complex social implications of speech technology voice choices [3, 4]. We are also encountering an increasing array of uncanny valley

Authors' addresses: Kristen M. Scott, kristen.scott@m-iti.org, Madeira Interactive Technologies Institute, Funchal, Madeira, Portugal; Simone Ashby, simone.ashby@m-iti.org, ITI / LARSyS, Funchal, Madeira, Portugal; Adriana Stan, adrianac.stan@gmail.com, Technical University of Cluj-Napoca, Cluj-Napoca, Romania.

phenomena around the use of TTS voices in varying contexts, such as the recent Google Assistant haircut appointment demo, described by one YouTube commenter as "[a]n AI with near perfect vocal fry" [1].

However, personal digital assistants are far from being the only application of synthetic speech. Apart from [10], much remains to be explored for uncovering listener needs and preferences around the use of TTS in broadcasting, and radio in particular. And yet TTS is poised to play a pivotal role in the success and longevity of both commercial and community stations. As part of a European consortium dedicated to expanding and augmenting an open technology stack for low-power FM community radio stations [5], we are particularly focused on the needs of our community partners (e.g. in rural Romania) for ensuring that stations are sustainable and do not violate their state granted licensing agreements. For our Romanian stations, this means providing non-stop, 24/7 broadcast content. Thus, given the task of designing and integrating purpose-built TTS applications within our FM radio technology stack, we start with the initial assumption that *some* amount of TTS will be deemed acceptable, and even necessary, by our local partners and their listening audiences. However, questions remain as to what other types of TTS content would be appropriate, as well as the type of synthetic voice (or voices) to use.

We present an automated synthesized audio weather forecast generator, which was designed through collaborating stations managers and volunteers and which we plan to extend to further content types for broadcast. We include results of our preliminary listening study of the generated and proposed content with Romanian speakers residing in Romania.

2 AUTOMATED WEATHER FORECAST APPLICATION

In conversations between the Romanian station managers and the fishermen chief in one of the communities, it was determined that weather and wind speed and direction forecast information are crucial for them. Current practice is to view this data from a variety of more weather web-sites, however, not everyone has access to computers and internet, and a regularly updating forecast heard over the radio would be valuable.

Forecast data is pulled from Yr.no, an open source weather data service. The application is scheduled to update three times daily. On update it pulls the most current forecast data from the service and inserts it into a pre-written script which is then converted to synthetic speech using the Cerecloud API [2]. The resulting audio file is saved to a server location which is configured to be accessed as an RSS feed, allowing the updated audio file to be added to the schedule of station software such as RootIO [5].

[Buna Dimineata] [Sfântu Gheorghe]. Prognoza de **[dimineață]** până la ora **[6:00]**, astăzi **[Cer senin]**, cu o temperatură de **[10]** grade, cu vânt de **[8,2]** metri pe secundă din direcția **[est]**. Prognoza de **[la noapte]** de la ora **[24]** până la ora **[06:00]**, este **[Înnorat]**, cu o temperatură de **[10]** grade, cu vânt de **[5,8]** metri pe secundă din direcția **[sud-vest]**. Prognoza meteo furnizată de aplicația yr.no, a Institutului Meteorologic din Norvegia și a NRK.

Fig. 1. An example weather forecast script with variable values in bold. Translation: *Good morning Sfântu Gheorghe. The forecast for this morning from 6:00 is clear sky, with a temperature of 10 degrees, with wind of 8,2 meters per second in the easterly direction. The forecast for the afternoon from 12:00 to 18:00 is cloudy, with a temperature of 10 degrees, with wind of 5,8 meters per second in the south-west direction. Weather forecast provided by YR.no application, the Norwegian Meteorological Institute and NRK*

[1] https://www.youtube.com/watch?v=yDI5oVn0RgM)

3 LISTENING TEST

In order to assess the viability of the generated content and to better understand potential preferences around broadcasting different types of TTS content, we conducted a listener test using clips of synthesized voices reading out content generated by our synthesized weather feed application. In addition to the weather clips, we also presented listeners with clips of synthesized news and cultural content as an initial effort to explore different TTS voice and content pairings, and for guiding future design decisions with respect to leveraging TTS for community radio.

> *I looked for it in flowers but their sophisticated geometry was inaccessible to me. I looked for her in love but it was so ephemeral. I have sought her wisely, but I have wandered among many paths. I looked for her in the newborn baby but I forgot the purity. I searched for it in monasteries, but the mystery was terrifying to me.*

Fig. 2. A translated excerpt of one of the cultural content scripts. Original taken from http://casapentrucultura.ro/

3.1 Methodology

A total of 17 participants, nine female and eight male ages 22 to 56 took part in an online listening test. All were Romanian nationals who had a background in engineering or speech engineering. Four participants, additionally participated in a 15-minute follow-up interview, which were conducted in English. Participants were asked to assess each audio clip in terms of four adjectival descriptors (coherent, suitable for radio, trustworthy, attractive) [2], as measured on a seven-point Likert scale. All audio and written survey content was presented in Romanian using a modified version of the webMUSHRA audio survey software [9].

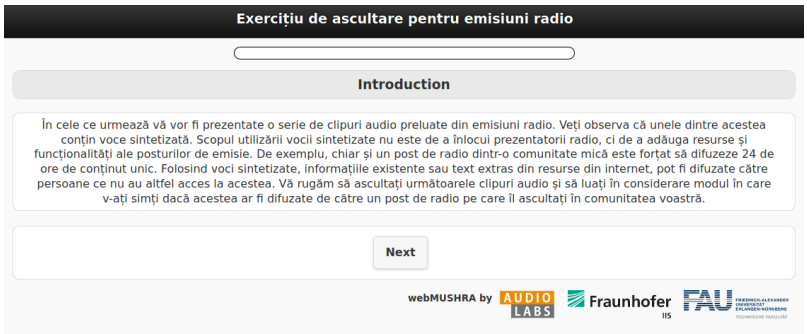

Fig. 3. Introduction to the online listening test. Translation: *In the following you will be presented with a series of audio clips of radio shows. You will notice that some of them contain synthesized voice. The purpose of using synthesized voice is not to replace radio presenters, but to add resources and functionality to radio stations. For example, even a radio station in a small community is forced to broadcast 24 hours of unique content. Using synthesized voices, existing information or text extracted from Internet resources, can be broadcast to people who have no other access to them. Please listen to the following audio clips and consider how you would feel if they were broadcast by a radio station that you listen to in your community.*

Participants rated a total of 15 audio clips, each of which featured ambient radio noise, such as tuning static, in-between clips. The clips distributed represent three content types - news, weather (see figure 1) and culture (see figure 2) - and were presented in a randomized order. Three different TTS voices were included in the test: two SWARA voices

[2]In Romanian: *coerent adecvat pentru radio, de încredere, atractiv.* - with the Romanian words in italics

[11], including a male (IPS) and a female (BAS) voice; and a high-quality 'characterful' female voice, created by Cereproc (CER) [2]. The two SWARA voices were generated using an HMM-based statistical parametric speech synthesiser, while the Cereproc voice was derived using a proprietary unit selection method that is generally perceived to yield more human-sounding TTS voices.

3.2 Results

3.2.1 Ratings by Content Type. The graph in Figure 4a shows the differences in ratings between clips of different content types. The differences were then analyzed using one-way repeated measure ANOVA tests through the statsmodels package in Python (version 3.7.5). Differences between content type ratings were found to be significant for the descriptors "coherent", "suitable for radio" and "attractive", while differences among "trustworthy" ratings were not statistically significant. A post-hoc analysis comparing ratings by content type was conducted for the descriptors "coherent", "suitable for radio" and "attractive" using pairwise t-tests. Cultural clips were rated as significantly less coherent (M=4.44; SD=1.52) than news clips (M=4.88; SD=1.52). Cultural clips were also rated as significantly less suitable for radio (M=3.94; SD=1.62) than news (M=5.03;SD=1.65) or weather (M=4.96; SD=1.59) clips. There was no significant difference between ratings of news and weather on ratings of coherence. Cultural and weather clips were rated as significantly less attractive (cultural M=3.41, SD=1.77; weather M=3.73 SD=1.64) than news clips (M=4.33 SD=1.63).

(a) Mean score, with standard error, on four measures (coherent, suitable for radio, trustworthy, attractive) for clips of each of the three content types (news, weather, culture

(b) Mean score, with standard error, on four measures (coherent, suitable for radio, trustworthy, attractive) for clips of each of the three voices (BAS, LIS, CER

Fig. 4. Results of listening test by content type and voice

3.2.2 Ratings by voice. We additionally used one-way repeated measure ANOVA tests were used to examine the differences in ratings per synthesized voice; the results are shown in Figure 4b. Differences between TTS voices were found to be significant for the descriptors "suitable for radio" and "attractive", while differences for the descriptors "coherent" and "trustworthy" were not statistically significant. Again, a post-hoc analysis comparing ratings by synthetic voice was conducted for the descriptors "attractive" and "suitable for radio" using pairwise t-tests. Clips featuring the BAS voice (i.e. the lower quality SWARA female voice) were rated significantly lower on both attractiveness (M=3.36; SD=1.66) and suitability for radio (M=4.19; SD=1.76) than clips featuring the CER (attractive M=4.03, SD=1.80; suitable M=4.73, SD=1.70) and IPS (attractive M=4.0, SD=1.67; suitable M=4.71, SD=1.64) voices. There was no significant difference distinguishing clips with the CER and IPS voices on measures of attractiveness and suitability for radio.

3.3 Discussion

The results indicate that the category of content being read out by our three synthetic voices can have an impact on the audio clip's perceived coherence, suitability for radio and attractiveness. However, we saw no significant effect of content type on the perceived trustworthiness of audio clips. Cultural content was rated below weather and news in terms of the clips' relative coherence and suitability for radio. With respect to perceived attractiveness, news ranked higher than both weather and culture. The overall acceptability of synthetic speech for broadcasting weather information has been documented in previous work. However, the current study appears to show that though it may be considered acceptable, it may not be considered 'attractive'. Our results further suggest that the use of TTS voices to broadcast news content may be deemed as good as or better than using TTS to broadcast weather information. The lower ratings observed for cultural clips indicates that additional research is needed to explore appropriate uses of synthesized content that go beyond "informative" to conveying more general interest or esoteric types of information.

When interviewed about their radio listening habits, participants identified a wide variety of preferred content (music, current events debates, listener call-in shows on politics, culture, and etc.). However, all four interviewees stated that they stop listening when the programming segues into content that they deem as being uninteresting. P1, for example, explained that they do not listen to political content because "the way politics is covered ... is not relevant to my life right now, I can't relate to it." We expect this dynamic will pose even more challenges for the appropriate design of synthesized radio content.

In terms of the effects the different TTS voices had on our listener judgments, the only significant difference we observed concerned clips read by the female SWARA 'BAS' voice, which participants rated as less attractive and less suitable for radio than the other two TTS voices. In interviews, participants described the BAS voice as high-pitched, "annoying', and featuring unnaturally pauses (P1); and as "too rapid" and "hard to listen to for more than a few minutes."

Overall, participants rated the voices as sufficiently intelligible, as supported by the fact that no significant difference in coherence rating per voice was observed. Interestingly, interviewees provided some subjective and unexpected perceptions of the voices they heard during the listening test. Two of the interviewees reported hearing four or more different voices, while another interviewee made frequent mention of a "baby" or "child" voice, which they found disconcerting.

4 LIMITATIONS AND FUTURE WORK

Given the paucity of studies focused on TTS for broadcasting, and for non-English language use cases in general, this preliminary listener study offered some interesting initial observations on perceptions of long-form synthetic speech as media content. Our future work will focus on further examining the interaction between voice and content type, as we seek to better understand and document best practices in the design of widely acceptable forms of synthesized news and weather content. Additionally, the acceptability and usefulness of this kind of content within the context of a specific community radio stations needs to be examined further through research and design within the given communities.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of ITI/LARSyS and the European Commission's Horizon 2020 Research and Innovation Programme (H2020-ICT-2016-2017-780890).

REFERENCES

[1] Matthew P. Aylett, Benjamin R. Cowan, and Leigh Clark. 2019. Siri, Echo and Performance: You have to Suffer Darling. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19*. ACM Press, Glasgow, Scotland Uk, 1–10. https://doi.org/10.1145/3290607.3310422

[2] Matthew P. Aylett and Christopher J. Pidcock. 2007. The CereVoice Characterful Speech Synthesiser SDK. In *Intelligent Virtual Agents*, Catherine Pelachaud, Jean-Claude Martin, Elisabeth André, Gérard Chollet, Kostas Karpouzis, and Danielle Pelé (Eds.). Vol. 4722. Springer Berlin Heidelberg, Berlin, Heidelberg, 413–414. https://doi.org/10.1007/978-3-540-74997-4_65

[3] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content. *Proceedings of the ACM on Human-Computer Interaction* (2020), 13.

[4] Julia Cambre and Chinmay Kulkarni. 2019. One Voice Fits All?: Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–19. https://doi.org/10.1145/3359325

[5] C. Csíkszentmihályi and J. Mukundane. 2016. RootIO: ICT + telephony for grassroots radio. In *2016 IST-Africa Week Conference*. 1–13. https://doi.org/10.1109/ISTAFRICA.2016.7530700

[6] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, Santa Clara, California, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[7] Judith A Markowitz. 2017. Speech and Language for Acceptance of Social Robots: An Overview. *Voice Interaction Design* 2 (2017), 11.

[8] Blaise Potard, Matthew P. Aylett, David A. Baude, and Petr Motlicek. 2016. Idlak Tangle: An Open Source Kaldi Based Parametric Speech Synthesiser Based on DNN. 2293–2297. https://doi.org/10.21437/Interspeech.2016-1188

[9] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. 2018. webMUSHRA — A Comprehensive Framework for Web-based Listening Tests. *Journal of Open Research Software* 6, 1 (Feb. 2018), 8. https://doi.org/10.5334/jors.187 Number: 1 Publisher: Ubiquity Press.

[10] Kristen M Scott, Simone Ashby, and Julian Hanna. 2020. (accepted) "Human, All Too Human": NOAA Weather Radio and the Emotional Impact of Synthetic Voices. *Proceedings of the ACM on Human-Computer Interaction* (2020), 9.

[11] Adriana Stan, Florina Dinescu, Cristina Tiple, Serban Meza, Bogdan Orza, Magdalena Chirila, and Mircea Giurgiu. 2017. The SWARA speech corpus: A large parallel Romanian read speech dataset. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, Bucharest, Romania, 1–6. https://doi.org/10.1109/SPED.2017.7990428