Navigating the Spoken Wikipedia

Marcel Rohde, *Timo Baumann*

Universität Hamburg, Department of Informatics, Natural Language Systems Group, Germany

{2rohde,baumann}@informatik.uni-hamburg.de

Abstract

The Spoken Wikipedia project unites volunteer readers of encyclopedic entries. Their recordings make encyclopedic knowledge accessible to persons who are unable to read (out of alexia, visual impairment, or because their sight is currently occupied, e. g. while driving). However, on Wikipedia, recordings are available as raw audio files that can only be consumed linearly, without the possibility for targeted navigation or search. We present a reading application which uses an alignment between the recording, text and article structure and which allows to navigate spoken articles, through a graphical or voice-based user interface (or a combination thereof). We present the results of a usability study in which we compare the two interaction modalities. We find that both types of interaction enable users to navigate articles and to find specific information much more quickly compared to a sequential presentation of the full article. In particular when the VUI is not restricted by speech recognition and understanding issues, this interface is on par with the graphical interface and thus a real option for browsing the Wikipedia without the need for vision or reading.

Index Terms: accessibility, eyes-free interaction, voice user interface, Wikipedia, hyperlistening

1. Introduction

Accessibility on the web is primarily established through valid and semantically meaningful markup that can be rendered by web agents regardless of the presentation format. An auditory rendition of the web is available to persons who cannot read with screen readers which provide spoken access and rely on text-to-speech and speech synthesis. One of the problems of general text-to-speech is the broad variety of text that it has to deal with, whereas domain-restricted technology can perform better.

For Wikipedia, one of the 10 most heavily accessed websites on the web[1], there is a specific webservice (the Pediaphon[2] [1]) which offers to read out encyclopedic articles, without requiring any screen-reading software. However, while both the quality of speech synthesis itself (i.e., the process of producing artificial speech sound) and of text-to-speech technology (the process of inferring how some text should be spoken, e.g. wrt. abbreviations, phrasing, intonation, etc.) have advanced considerably in the past years [2], the quality of artificial speech still lacks compared to natural speech, even for read-out text [3]. Text-to-speech mostly performs sentence-by-sentence and hence is unable to adequately cover discourse and information structure (with some notable exceptions, e.g. [4]). Humans in contrast, do very well at presenting the information structure and this is crucial for understanding with little effort [5].

The Spoken Wikipedia[3] is a project in which volunteers read out articles from Wikipedia to provide high-quality aural access to Wikipedia for people who cannot read. Roughly a thousand articles for each of English, German and Dutch are available, each totalling around 300 hours of speech (with smaller amounts in another 25 languages). This data has recently been made accessible by Köhn et al. [6][4] who automatically aligned the audio recordings to their respective article texts using speech recognition technology. Using these alignments, we are able to relate what parts of the article are spoken at any moment in the recordings. While the resource can be useful for fostering speech technology research (e.g. training acoustic models for open-source speech recognition), we want to make the material more accessible for its original purpose, to bring natural speech to those who prefer speech over text but do not necessarily want to linearly listen to full recordings.

2. The Written and Spoken Wikipedia

Wikipedia is accepted as the standard source for encyclopedic knowledge on the web and comes in the form of a strongly interlinked *hypertext*. Hypertext adds to traditional text the means for reading along a self-chosen reading path (i.e., non-linearily), called *hyperreading* [7]. Wikipedia provides indices, extensive structural information, and – most importantly – associative links to enable hyperreading. A common strategy in hyperreading Wikipedia is *leaping* between sections of articles and between articles based on links or structure [7]. The recent advent of *find as you type* in most browsers has made *text*

[1] http://www.alexa.com/topsites

[2] www.pediaphon.org

[3] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Spoken_Wikipedia

[4] https://nats-www.informatik.uni-hamburg.de/SWC/

Table 1: Comparative statistics of spoken and written versions of the German and English Wikipedia.

		German	English
Written	# articles	1,950,022	5,174,458
	— distinguished	6,283	29,189
	average text size	5.3 kB	6.2 kB
Spoken	# articles	916	1,344
	— distinguished	314	213
	average text size	25.8 kB	26.0 kB
Spoken	articles	0.047 %	0.026 %
Coverage	— distinguished	5.0 %	0.73 %
	est. speech time	0.22 %	0.11 %

search a frequently used strategy to find information in web pages, including for users with disabilities [8].

The Spoken Wikipedia has previously only been available as a linear audio recording, omitting all the positive aspects of hypermedia and making navigation or search impossible. Our software sets out to change this.

As also mentioned by Zhang [7], a disadvantage of hyperreading is the possibility of getting lost due to the flexibility of what to read next. Getting lost may be of particular concern when *hyperlistening*, as speech is such an inherently linear medium. Our experiments below will hence focus on whether participants are able to leap through speech without getting lost (too much), by assessing whether they are successful in navigating to key information in the article.

Wikipedia contains millions of articles on all sorts of topics in the major languages, inviting the question of whether the Spoken Wikipedia's meager thousand articles per language (at least for English, German and Dutch) are of any practical relevance when browsing for spoken information, or whether a screen reader is needed in all practical use-cases anyway.

To address this concern, we compare the composition of the written and spoken collections for German and English in Table 1. As can be seen in the table, both language versions consist of several million articles each, with a small proportion of *distinguished* articles.[5] We estimate the average length of written articles on a random sample of 1,000 articles for both languages (using their size in bytes as a proxy for text length). We find that articles selected for being spoken are (a) much longer than average articles (4-5 times as long), and (b) more often come from one of the distinguished article categories. In the German Wikipedia, some 5 % of distinguished articles have been read. Nevertheless, only a tiny proportion of the full Wikipedia is available as a naturally read version (0.11–0.22 %) and we estimate that a fully read Wikipedia would have an audio duration of several decades – indicating the infeasibility of full coverage.

While high-quality synthetic voices can be rated as more natural than amateur speech [9], naturalness ratings have been shown to degrade when listening to synthesized speech for an extended period [10], making the advantage of natural speech particularly relevant for long and complex articles from the distinguished categories.

Distinguished articles also tend to be more stable with fewer relevant changes, and hence their recordings remain up-to-date for longer. Thus, while we have equipped our software with the ability to synthesize articles on-demand, our experiments reported below focus on natural speech and we focus on relatively long articles of around one hour of speech.

3. Implementation

We first explain how we postprocess the SWC to re-align text and HTML markup. We then describe the graphical and voice user interfaces of our application.[6]

3.1. Data model

The Spoken Wikipedia Corpus [6] contains per-article alignments of plain text to audio. Unfortunately, those alignments do not take into account the article structure (in terms of the HTML DOM). In addition, the text has partially been altered to ease alignment and does not fully match the text (and other elements) contained in the HTML version. We overcome this issue by using fuzzy matching to produce a document that contains all of:

- the structural hierarchy of the article,
- the timing of all time-aligned words in the article,
- the sentence segmentation from the corpus, and
- the hyperlinks contained in the article.

This enables the application to

- leap (by sentence, paragraph, or section),
- identify links close to the current timing in the article audio (and follow these links), and
- identify timings for all words (for searching).

Both the time-alignment and matching occasionally go astray or are missing some data. Our method is nevertheless robust to such errors and provides timings whenever possible. We synthesize the table of contents based on the observed article structure, as this is not spoken by the readers; other material that is not spoken by readers (e.g. tables, lists, bibliographies) remains left out.

3.2. GUI

The *graphical user interface* consists of multiple parts that can each be hidden for experimentation. It is implemented in JavaFX and depicted in Figure 1. It offers multiple ways of accessing and leaping the structure of the article, as well as access to close-by links.

[5] English articles can be distinguished as either 'good' or 'featured', where the corresponding German categories are 'lesenswert' (worth reading) and 'exzellent'.

[6] Available at `http://github.com/hainoon/wikipediareader`.

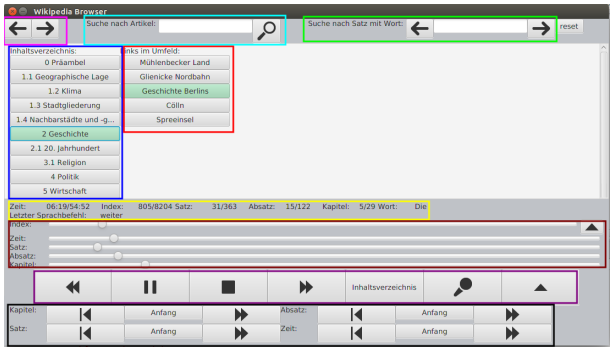

Figure 1: The full application GUI, including forward/backward jumps between articles (magenta), article search (cyan) and within-article search (green), the responsive table of contents (blue), a responsive list of currently relevant links (red), some status information (yellow), sliders indicating the relative position in the article (brown), buttons for standard audio navigation (forward/backward/pause), for listening to the table of contents, and for voice-based interaction (purple), and finally buttons to navigate the article structure: by chapter, paragraph, sentence, or jumping ahead/back by 10 seconds per click (black). In the experiments, only parts of the interface are available to users.

3.3. VUI

The *voice user interface* for navigating spoken articles consists of speech activation, recognition and rule-based language understanding with the aim of offering similar functionality as the graphical interface.

The user presses and holds down the only button in the interface to activate speech recognition. When the button is released, we decode the recording using Google's freely available Speech API [11][7].

Language understanding makes use all returned (n-best) hypotheses using a hierarchy of patterns. For robustness, patterns only need to match parts of what was spoken, allowing the user the freedom to add material such as "show me" or "now, go to". The hierarchy of rules is important as multiple rules may match a given input. N-best results are useful to deal with Google's variability in returning numbers (and other material). Users may say (variations of) the following:

- "[show me the] [table of] contents",
- "next/previous chapter/section/paragraph/sentence",
- "[go back to the] beginning of the chapter/section/paragraph/sentence" (or simply "repeat"),
- "[go to] chapter/section/subsection N",
- "*section name*" to go to the named section,
- "article *name*" to follow a link or search an article.

Our language understanding (as well as other parts of the software) currently work for English and German and would be easy to port to other languages.

[7] https://cloud.google.com/speech/

Figure 2: Setup of the user study: the experiment participant (right side) and the experimenter/wizard (left side) are separated by a dividing wall.

4. User Study

We conducted a user study to gain insight into the preferred modality for interaction, to see whether targeted navigation works as expected, and to learn about the overall usability of our software. For our experiment we disabled the search and link-following options in order to force users to stay within the article and to focus on structural navigation within the article.

Participants were given a choice of two articles so as to increase interest in the article in question. Participants where first allowed 2 minutes of 'free browsing' in the article. Afterwards, they were asked to use targeted navigation to answer three factual questions about the article in question. The facts were positioned anywhere in the article and sometimes required some combination (such as aggregation of denominations for the full proportion of religious affiliation). We compare three conditions:

GUI Users interacted using the graphical user interface as described in Subsection 3.2 above.

VUI Users interacted by speaking voice commands to the system described in Subsection 3.3. They were given a schema for possible commands.

Wizard-control As in the the VUI setting, users interacted by speaking, but were instructed to use commands as they saw fit for the task (lead to believe that this was a 'better' system). In this condition, the experimenter followed the Wizard of Oz paradigm and navigated the article according to how the speech interface *should* act absent of recognition (and ensuing understanding) errors.

12 participants (normally sighted, not regular TTS or screen reader users) took part in the study. Each participant used the system in all three conditions and we balanced for ordering effects. The first 6 participants were allowed no more than 2 minutes for each question, the remaining 6 participants were allowed a total of 15 minutes for the questions with gentle reminders to move on after 5 minutes per question. As participants were given a free choice of 2 articles for each condition, we could not balance the usage of every article.

In all conditions, users wore a headset to listen to the recording. The headset's microphone was used only in

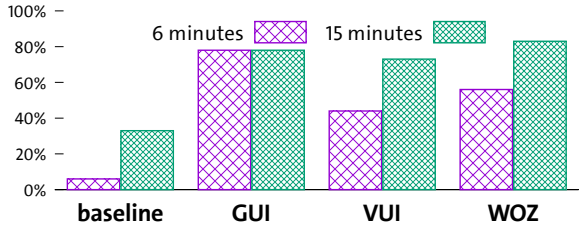

Figure 3: Proportion of questions answered after 6/ 15 minutes for the experimental conditions and a non-interactive baseline.

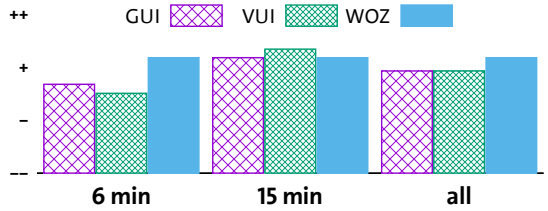

Figure 4: Average user ratings of overall interaction quality for the interaction conditions.

the VUI condition, whereas the wizard directly heard the speaker and performed commands using the GUI from a separate computer. See Figure 2 for a picture of the setup.

We asked the participants to fill out a questionnaire after the initial 'free browsing' and after targeted navigation for each interaction condition.

5. Results

We analyze our user study with respect to the participant answers to the given questions, their ratings in the questionnaire and the logged interaction behaviour. Given the low number of participants and the free choice of the read article, we do not expect results to be significant; they are, however, clearly indicative of general tendencies.

5.1. User Success

Figure 3 shows the proportion of (fully or partially) correct answers under the three experimental conditions for the first group (2 minutes per question, 6 in total) and second group (15 minutes). We add a baseline condition in which the user would not be able to navigate (and hence only be able to give answers that have occurred after a maximum of 6 and 15 minutes, respectively). As can be seen, targeted navigation greatly improves over linear listening. We find that voice-based navigation profits from longer interactions, then reaching results on par with the GUI. We want to add that a few questions were never answered correctly because the information was very hard to find given just structural navigation.

5.2. User Feedback

Figure 4 shows the overall interaction quality as reported in the questionnaires. All versions are rated as 'usable' with a slight tendency towards spoken interaction (possibly because there is no modality change between output and input as commented by one user). Users tend to rate better when they had more time to interact, indicating that only 2 minutes per question result in stress, whereas 5 minutes are sufficient. Stress could be lower in the WOZ condition in which interaction was more successful.

Users often commented that they would have liked to search by keywords, a functionality that we had excluded from the experiment. We believe that voice-based interaction will further improve when search is included.

5.3. User Behaviour

All participants interacted heavily (hyperlistened) in all conditions rather than listen linearly. In particular, they (a) navigate to sections, (b) skip ahead one section, paragraph or sentence, (c) go back one sentence when they notice that they found the desired information, or (d) pause playback. The GUI condition also shows interesting use of skipping words (presumably to save time), and in voice-based interactions users often call the table of contents (before then calling for a section). Unfortunately, we did not record statistics of whether participants prefer to call sections by name or number.

Users often pressed the push-to-talk button too late (and/or released it too early) which hindered recognition. This could easily be solved by voice activity detection. Likewise, while speech recognition worked well for some, VUI performance was greatly restricted by errors. This as well could be solved by better technology.

6. Summary and Conclusions

We have described a system for aural access to Wikipedia articles: spoken articles can be navigated via their structure, or searched by keywords and links can be followed to voice-browse the full Wikipedia (with articles synthesized if not available in a naturally spoken version). Our software enables *hyperlistening*, i.e. making use of the crucial hypertextuality of modern encyclopaedia usage without the need for reading.

We find that users are able to navigate to information in articles much quicker than if they had to listen linearly, and their usage patterns as well as comments indicate that they easily stay on top of things even without feedback about the current position in the article.

Both the graphical as well as the voice-based mode of interaction work well, at least when speech recognition error is low and enough time is available. This indicates that hyperlistening fits well with voice-based navigation and can hence be useful for persons without vision available for browsing.

Finally, while our interfaces enable browsing naturally read articles, the full Wikipedia experience includes user participation such as adding links and contents [7], or commenting on the 'talk' pages. Thus, ours are just initial steps towards a full eyes-free and speech-only access to Wikipedia.

Acknowledgments We wish to thank all Wikipedia authors and speakers for creating and maintaining the written and spoken data, as well as the participants in our experiments. We also thank Michael Blesel for selecting test articles and devising the questions used in the experiments. Finally, we wish to thank the reviewers for their helpful comments and pointers.

7. References

[1] A. Bischoff, "The Pediaphon-speech interface to the free Wikipedia encyclopedia for mobile phones, PDA's and MP3-players," in *18th International Workshop on Database and Expert Systems Applications (DEXA 2007)*. IEEE, 2007, pp. 575–579.

[2] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[3] P. Taylor, *Text-to-Speech Synthesis*. Cambridge Univ Press, 2009.

[4] F. Kuegler, B. Smolibocki, and M. Stede, "Evaluation of information structure in speech synthesis: The case of product recommender systems," in *Speech Communication; 10. ITG Symposium; Proceedings of*, Sept 2012, pp. 1–4.

[5] J. Hirschberg and J. Pierrehumbert, "The intonational structuring of discourse," in *Proceedings of the 24th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1986, pp. 136–144.

[6] A. Köhn, F. Stegen, and T. Baumann, "Mining the spoken wikipedia for speech data and beyond," in *Proceedings of LREC 2016*, 2016.

[7] Y. Zhang, "Wiki means more: hyperreading in Wikipedia," in *Proceedings of the seventeenth conference on Hypertext and hypermedia*. ACM, 2006, pp. 23–26.

[8] L. Spalteholz, K. F. Li, and N. Livingston, "Efficient navigation on the world wide web for the physically disabled." in *WEBIST (2)*, 2007, pp. 321–327.

[9] K. Georgila, A. Black, K. Sagae, and D. R. Traum, "Practical evaluation of human and synthesized speech for virtual human dialogue systems." in *LREC*, 2012, pp. 3519–3526.

[10] E. Pincus, K. Georgila, and D. Traum, "Which synthetic voice should i choose for an evocative task?" in *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, vol. 105, 2015.

[11] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Your word is my command: Google search by voice: A case study," in *Advances in Speech Recognition*. Springer, 2010, pp. 61–90.