Mapping course text to hyperaudio

Niels Seidel [1]

Abstract: Informal knowledge transfer through podcasts, audiobooks, and radio documentaries enjoys great popularity. However, informal learning settings at schools or universities auditory learning resources are rarely used. This is due to the high cost of producing and updating auditory resources. Furthermore, some learning resources can hardly be presented auditorily. In this paper, we present an approach to convert rich-text course material from LaTeX or MS Word format into interactive hyperaudio documents. For text to speech (TTS) conversion, we make use of common TTS web services. Text-specific design is mapped to audible effects and enriched by time-based interactive visual material. In a dedicated Moodle plugin we provide a basic hyperaudio player to be used within a course or as a stand-alone web application. With this approach, existing course texts can be transferred into hyperaudio within minutes.

Keywords: Hyperaudio; Multimedia Learning; Text to Speech

1 Introduction

Informal knowledge transfer through podcasts, audiobooks, and radio documentaries enjoys great popularity.However, in formal learning settings at schools or universities auditory learning resources are rarely used. On the one hand, this may be caused by the high production and update costs. On the other hand, visual learning content can not be presented auditorily. In teaching, the text is likely to be the dominant learning medium due to a long tradition. Modern textbooks or scripts or comparable e-books contain not only text but also illustrations, tables, and, depending on the discipline, formulas and program code. In principle, text-based media are just as legitimate as video-based or auditory learning media. From the learner's point of view, there are several advantages to offering auditory learning media: In addition to primarily text-based learning resources, learners receive an additional representation in the form of enriched audio documents. The use of the auditory channel can be a relief for learners who were previously confronted with a lot of visual information (e.g. after screen work). The auditory information can be received and processed in parallel to other activities, e.g. during sports as well as while sitting on a bus or train). Through auditory communication, technical terms and their correct pronunciation are conveyed.

To be able to use the audio to its full extent for the presentation of rich-text learning resources, it is advisable to prepare it as a hypermedia document so that figures, links, etc. can be integrated. This particular type of audio-based hypermedia is called hyper-

[1] FernUniversität in Hagen, Chair of Cooperative Systems, Universitätsstr. 1, 58097 Hagen, Germany Land
niels.seidel@fernuni-hagen.de, https://orcid.org/0000-0003-1209-5038

audio. In this paper, we present an approach to converting rich-text course material from LaTeX to interactive hyperaudio documents. The goal of this paper is to share insights into (i) the auditory design of longer speeches, (ii) the selection and configuration of text to speech (TTS) systems, and (iii) the design of a hyperaudio player to be used in Learning Management Systems (LMS).

2 Related works

This section summarizes related work regarding the auditory design of text-based resources and hyperaudio players. Different TTS systems will be discussed later in section 4. Donker; Blenn [DB07] laid a foundation for the use of audio in hypermedia applications with the Hyperaudio Encyclopedia. In this project, unlike the use of screen readers, articles in the encyclopedia were not just recorded and read to the listener. The users were able to interact with the audio document. For instance, links, headings, and other salient test passages were highlighted by auditory markers and recognizably presented to the listener. In terms of learning, Reinmann [Re09] encouraged receptive processes for listening rather than equating them with passivity. Receptive processes, such as listening to podcasts, must be contrasted with productive learning processes. Reinmann suggests improving the quality of storytelling and listening through modern technologies. In a study, [ZS14] compared textually and auditorily represented information in linear and non-linear forms. The non-linear representation increased cognitive load compared to linear representations. [Mo10] showed in a study with 100 participants that offering podcasts and assessments in addition to the usual lecture materials have a positive learning effect compared to a control group without these additional materials. In terms of hyperaudio technology, existing frameworks such as popcorn.js, waivsurfer.js, or timesheet.js [CQR11] can be used for implementation. Also, several commercial services like Soundcloud or YouTube offer the possibility to markup continuous audio documents similar to hyperaudio players. To the best of our knowledge, no hyperaudio player is yet available that enables audio design through effects and augmented content.

3 Audio design

For the conversion of text into speech, it would be insufficient to merely transfer the written sentences and words into spoken words, because the semantics expressed in the design of the text should be translated using acoustic means. Therefore, possibilities of designing the auditory space are described as audio design Raffaseder [Ra10].

Using non-fiction texts as an example, five design areas can be identified, each of which is considered in different ways in audio design: (i) text passages such as bulleted lists, examples, tasks, definitions, or quotations; (ii) text insertions such as footnotes, literature

references, expressions in parentheses, and marginalia texts; (iii) text representation such as tables, figures, formulas, program code; (iv) navigational support such as table of contents, page numbers, outline levels, cross-references, keyword indexes; and (v) highlighting (e.g., bold, italic, underline, small caps, color backgrounds, font color). There are several possibilities for implementing these textual design elements using acoustic means. However, text representations and elements do not have a direct auditory counterpart and in some cases can only be conveyed via the visual channel.

The accentuation when reading aloud can partly be derived from the written word based on punctuation marks, paragraphs, and headings. However, how individual words or parts of sentences are accentuated cannot be deduced from the text. The emphasis depends on the person reading the text. Rhetorical skills, voice training, and also subject knowledge of the content influence the pronunciation and emphasis when reading aloud. For synthesized speech production this contextual information is not available, but there are other design possibilities. Pauses and variations in pitch are among the simplest, but at the same time most effective, auditory effects. Beyond the spoken language, additional sounds and acoustic cues, so-called audio cues, can be integrated, which, however, can only be experienced for a short time compared to background music.

An audio cue is an acoustic indication. This can take the form of a tone, a sequence of tones, a melody, music, noise, speech, or a change in volume, frequency, timbre, or the addition of effects such as echo, reverb, modulation, and so on. In addition, the time and duration of the fade-in can be varied. An audio cue has semantics that must be conveyed to the listener. Although there are many design possibilities, listeners will only be able to differentiate between a limited number (< 7) of different audio cues. Audio cues must therefore be sufficiently distinct from each other. They should not interfere with or even overlay the speech output. At the same time, audio cues should be perceptible and not overheard. Also, “no sounds should be used that could also be heard in the user’s environment” [DB07]. In the literature, three types of audio cues are distinguished: (i) Auditory Icons use (nonverbal) sounds and noises from the user’s natural environment [BF11], (ii) Earcons as nonverbal, abstract, synthetic sounds [VA03], and (iii) Hearcons as the smallest unit of an auditory environment that represents an object through a sound characterized by volume, position, and extent in space [BG95]. The acoustic perception of audio cues and other acoustic design elements is always subject to the perception in space. Spatial effects can be generated artificially and used purposefully.

With regard to Audio Augmented Reality [Ba19] spatial hearing helps people to orient themselves in space. Level and frequency differences of direct and indirect sound sources enable localization of sound sources as well as an idea of the size and shape of the surrounding space. Sound sources can also be artificially generated in a virtual room in this way, so that a spatial sense of sound is created, at least by using stereo headphones. For the creation of a spatial listening experience one can make use of certain psychoacoustic properties like especially (i) interaural time differences, (ii) interaural level differences, (iii) reverb, and (iv) auto-pan effects. One way to realize

these effects provides the Resonance Audio SDK from Google[2].
In Tab. 1 a proposal of auditory representations for the text-based design elements is presented. The implementation of the auditory design was done on the one hand by the TTS system and on the other hand by audio cues and sound effects during the playback of the resulting audio document as well as by complementary interactive elements of the hyperaudio player.

Text design element	Hyperaudio representation
Headline	Short break before, pronounced with lower pitch
Chapter/section start	Moving audio source from bottom to top
Lists numbered/unnumbered	Switching pan from right to left per item
Blocks, e.g. examples, definitions	Audio cues per type of block
Quotes	Higher pitch
Footnotes	Audio cue
References	Audio cue
Expressions in brackets	Converted to speech with pronounced brackets
Cross-references	Audio cue
Tables, figures formulas, code	Audio cue indicating visual representation
Marginalia texts, table of contents, page numbers, index	(not considered)

Tab. 1: Mapping text design to auditory representations

4 Text to speech to hyperaudio

Hyperaudio is a special type of hypermedia in which hyperlinks and multimedia content are associated with the carrier medium audio. In our case, the basis for the carrier media is the course texts. These texts are transformed into an artificial speech output with the help of a TTS system. The conversion requires a special markup language, the Speech Synthesis Markup Language (SSML). Therefore, the conversion is performed in two steps: First, generation of the SSML from the text source. Second, generation of the audio from the SSML. In addition, the annotation of time markers will be automated to allow navigation and augmentation of the audio document with figures, tables, and links.

For the creation of the SSML document, a parser for LaTeX was written. Indirectly, Word or Writer documents can also be processed after a conversion to LaTeX using the LibreOffice extension Writer2LaTeX[3] in the Terminal. The task of the parser was on the one hand to transfer semantic and structural information like headings, citations and literature references to SSML and on the other hand to remove superfluous syntax commands from LaTeX. The BibTeX keys contained in the text could be converted into a readable form of author names using the Node.js module biblatex-csl-converter.

[2] See https://resonance-audio.github.io/resonance-audio/ (retrieved 04/14/2022).
[3] See https://extensions.libreoffice.org/en/extensions/show/writer2latex-1 (accessed 2022-06-14)

Speech output of figures, tables, formulas, and program code was not pursued in favor of a visual representation in the hyperaudio player. From the LaTeX sources, these elements could be extracted and converted to PNG image files using pdflatex.

Common cloud services considered as TTS were Google's TTS, Amazon Web Services (AWS) Polly, and Microsoft Azure. Microsoft Azure could not be used due to a lack of sufficient documentation. Google TTS supported far fewer SSML commands and especially fewer intonations than AWS Polly at the time of development (2021). For example, AWS Polly could be used to mark up English language citations in SSML so that they could these parts could be pronounced in English instead of German. The audio documents created with AWS Polly were of much better quality, both sonically and linguistically, than the audios created with Google TTS.

Worth noting is the <mark> markups, which are not included in the speech generation but allow the generation of a JSON file with timestamps of each sentence, word, or customized marks. This eliminates the need to manually identify the temporal positions of particular text elements. Cross-references to tables, figures, and text sections contained in LaTeX sources could also be declared as customized marks for later integration in the hyperaudio player. Using the <mark> markups, it was possible to set triggers for acoustic playback effects. Finally, word and sentence level marks enable precise navigation between the written words in the text and the corresponding speech.

The Hyperaudio Player was developed in an activity plugging for the Moodle LMS. The player allows teachers to upload audio files and associated SSML and mark files from AWS polly. Learners can play, pause and adjust the playback speed of the audio file. During playback, the corresponding sentences in the text are highlighted, and if necessary, the text scrolls to the position. Visual material that could not be represented as audio is displayed in the text. Audio cues provide users with an acoustic cue to available visual content and allow them to view visual content and use hyperlinks on-demand using the display. By clicking on a sentence in the text, the user can navigate to the corresponding position in the audio document. The horizontal timeline known from audio players is thus replaced here by the vertical scroll bar available in the browser. In summary, the audio design has been realized as proposed in Tab. 1.

5 Summary and outlook

In this paper, we shared insights on how to design auditory representations of text-based learning material in higher education. For the audio design features of the TTS systems, audio cues, and audio effects have been used to create a spatial listening experience. Beyond the audio, a hyperaudio player was developed that incorporates interactive and multimedia elements. The audio documents generated with AWS Polly have been used in the winter semester 2019/20 and 2020/21 Moodle courses for 82 and 67 students, respectively, of a master's program in computer science. Learners were asked in the

course forum and also in the virtual classroom sessions to assess the linguistic quality and individual usefulness of the audio course texts. The feedback was largely positive. Only a few people were bothered by the occasional imprecise pronunciation and the comparatively limited tonal variation. The Hyperaudio Player will be used in the coming winter semester. Until then, we are goint to improve the usability and user experience as well as enable a collaborative use.

Acknowledgement: This research was supported by the Research Cluster Digitalization, Diversity and Lifelong Learning – Consequences for Higher Education (D^2L^2) of the FernUniversitaet in Hagen, Germany.

Bibliography

[Ba19] Bauer, V.; Nagele, A.; Baume, C.; Cowlishaw, T.; Cooke, H.; Pike, C.; Healey, P.G.T.: Designing an Interactive and Collaborative Experience in Audio Augmented Reality. In (Bourdot, P.; Interrante, V.; Nedel, L.; Magnenat-Thalmann, N.; Zachmann, G., eds.): Virtual Reality and Augmented Reality. Springer International Publishing, Cham, pp. 305–311, 2019.

[BF11] Brazil, E.; Fernström, M.: Using and Creating Auditory Icons. In (John G. Neuhoff Thomas Hermann, A. H., ed.). Logos Publishing House, 2011.

[BG95] Bölke, L.; Gorny, P.: Direkte Manipulation von akustischen Objekten durch blinde Rechnerbenutzer. In (Böcker, H.-D., ed.): Software-Ergonomie ’95 Mensch-Computer-Interaktion Anwendungsbereiche lernen voneinander. B.G.Teubner, Stuttgart, pp. 93–106, 1995.

[CQR11] Cazenave, F.; Quint, V.; Roisin, C.: Timesheets.js: When SMIL Meets HTML5 and CSS3. In: Proceedings of the 11th ACM Symposium on Document Engineering. ACM, New York, NY, USA, pp. 43–52, 2011.

[DB07] Donker, H.; Blenn, N.: Gestaltung von Hyperlinks in einer Hyperaudio-Enzyklopädie. In (Gross, T., ed.): Mensch & Computer 2007: Konferenz für interaktive und kooperative Medien. Vol. 7, Oldenbourg Verlag, München, pp. 139–148, 2007.

[Mo10] Morris, N. P.: Podcasts and Mobile Assessment Enhance Student Learning Experience and Academic Performance. Bioscience Education 16/1, pp. 1–7, 2010.

[Ra10] Raffaseder, H.: Audiodesign. Hansa Verlag, München, 2010.

[Re09] Reinmann, G.: iTunes statt Hörsaal? Gedanken zur mündlichen Weitergabe von wissenschaftlichem Wissen Mündliche Weitergabe wissenschaftlichen Wissens. E-Learning – Lernen im digitalen Zeitalter 51/, pp. 256–267, 2009.

[VA03] Vargas, M. L. M.; Anderson, S.: Combining speech and earcons to assist menu navigation. In. 2003.

[ZS14] Zumbach, J.; Schwartz, N.: Hyperaudio learning for non-linear auditory knowledge acquisition. Computers in Human Behavior 41, pp. 365–373, 2014.