


Immersive Pedagogy: Converting AI-Driven Podcasts into Virtual Reality Learning Objects

Tomas Blažauskas 

Department of Software Engineering, Kaunas University of Technology, Lithuania

Eglė Butkevičiūtė 

Department of Software Engineering, Kaunas University of Technology, Lithuania

Filippo Sanfilippo 

Department of Software Engineering, Kaunas University of Technology, Lithuania

Patrikas Armalis 

Department of Software Engineering, Kaunas University of Technology, Lithuania

Ugnė Auksoraitytė 

Department of Software Engineering, Kaunas University of Technology, Lithuania

Abstract

Educational content creation is being transformed by artificial intelligence (AI) tools that can generate, process, and deliver learning materials across different environments. This paper presents a novel approach for converting AI-generated podcasts into immersive virtual reality (VR) learning experiences. We use the NotebookLM platform to create educational podcast audio recordings, which are then transformed into VR learning objects through automatic transcription, speaker diarization, and integration into a VR environment. The resulting VR learning objects address multiple learning preferences - supporting auditory learners through audio and visual learners through virtual presenters, and kinesthetic learners through immersive interaction. We evaluated this approach within a software engineering course, demonstrating its practical applicability and educational effectiveness.

2012 ACM Subject Classification Software and its engineering → General programming languages

Keywords and phrases Artificial Intelligence, Virtual Reality, NotebookLB, Learning objects, content generation

Digital Object Identifier 10.4230/OASICS.ICPEC.2025.16

Funding This research was funded by the Erasmus programme under the grant number 2023-1-ES01-KA220-VET-000153652.

Acknowledgements We want to thank all partners from the VR4INCLUSIVE Erasmus project whose contribution made this article possible.

1 Introduction

Generative artificial intelligence (AI) is transforming how we create educational content and learning objects across various fields. These technologies enable more personalized learning by efficiently generating educational materials tailored to different learner needs and learning styles [7]. NotebookLM represents a successful application of generative AI in educational content creation. This AI platform offers dynamic content generation features that allow real-time customization and adaptation, helping instructors create personalized learning experiences [8]. The platform uses Large Language Models (LLMs) to summarize source materials in various formats and generate educational content such as summary overviews, quiz questions, glossaries, and more. NotebookLM helps reduce the cognitive load on educators, allowing them to focus more on teaching rather than administrative tasks [4].



© Tomas Blažauskas, Eglė Butkevičiūtė, Filippo Sanfilippo, Patrikas Armalis, and Ugnė Auksoraitytė; licensed under Creative Commons License CC-BY 4.0

6th International Computer Programming Education Conference (ICPEC 2025).

Editors: Ricardo Queirós, Mário Pinto, Filipe Portela, and Alberto Simões; Article No. 16; pp. 16:1–16:12

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

One of its most interesting features is the ability to create podcasts based on provided material. These podcasts feature two AI-generated hosts (a man and a woman) who discuss various aspects of the content in a conversational format. While this audio format works well for learners who prefer auditory learning, it may be less effective for those who learn better through other modalities [1].

This paper presents our approach for transforming AI-generated audio podcasts from NotebookLM into virtual reality (VR) learning objects. Our method is designed to support learners with reading and visual learning preferences, and to some extent, kinesthetic learners as well. We have applied this methodology in a software engineering course with over 200 students and conducted a comprehensive survey to evaluate its effectiveness. The results provide insights into both the technical implementation and the educational impact of converting AI-generated audio content into immersive VR learning experiences.

2 Methods to Support Virtual Show Creation

Recent research offers several promising technological approaches for transforming educational audio content into immersive VR experiences. This section examines methodologies that can enhance our pipeline for creating VR TV shows from generated podcasts.

2.1 Advanced Speaker Diarization Approaches

Recent work by Wu [11] demonstrates that text-based speaker diarization methods can perform comparably to – and in some cases outperform – traditional audio-based systems by 2.5% to 10%, particularly for shorter educational content under 15 minutes. The Multiple Prediction Model (MPM) framework provides particularly robust speaker change detection by making multiple predictions about speaker changes at every transition point and using advanced aggregation mechanisms to determine the result.

Sentence-level speaker change detection has shown advantages over traditional word-level approaches by maintaining semantic coherence and capturing richer contextual information. This is especially important for educational content where complete thoughts and explanations must remain attributed to the correct virtual speaker.

For evaluation, specialized text-based metrics like Text-based Diarization Error Rate (TDER) and Diarization F1 (DF1) provide more comprehensive quality assessment by considering both transcription accuracy and speaker attribution in a unified framework. These metrics are particularly valuable for educational content where both accuracy and correct attribution are essential.

2.2 AI-Generated Visual Content Integration

Xu et al. [12] presents a structured approach for transforming educational content into immersive experiences. Their three-stage pipeline for converting slides into “generated instructional videos in under 10 minutes” involves generating lecture text, creating digital characters, and producing instructional videos through the integration of multiple AI technologies.

For creating authentic virtual presenters, they found that “dynamic video input over static photos captures more natural facial expressions and blinking, including those above the mouth,” resulting in more engaging presentations. This approach aligns with advanced lip-syncing and facial animation techniques that create convincing virtual hosts in VR environments.

Regarding learning effectiveness, studies indicate “significantly higher retention in the AI-generated instructional video condition compared to recorded videos,” suggesting that multimodal VR presentations can enhance information retention by engaging multiple sensory channels simultaneously. Importantly, users of AI-generated videos “reported lower social presence when interacting with virtual resources, but simultaneously experienced lower cognitive load,” indicating that immersive content can offer cognitive benefits while providing engaging learning experiences.

2.3 Cross-Lingual Adaptation Challenges

Pérez et al. [9] highlight both the promise and limitations of cross-lingual voice technologies in educational settings. “The rapid progress of modern AI tools for automatic speech recognition and machine translation is leading to a progressive cost reduction to produce publishable subtitles for educational videos in multiple languages.” Their research suggests that “text-to-speech technology is experiencing large improvements in terms of quality, flexibility and capabilities,” enabling “lecturer’s voice cloning in languages she/he might not even speak.”

However, significant challenges remain for less-resourced languages. While cross-lingual adaptation is promising, the main obstacle is the quality of the generated audio in Lithuanian. Existing models are not good enough to produce natural-sounding speech in Lithuanian. Despite this limitation, Pérez et al. found that “TTS technology is not only mature enough for its application at the UPV, but also needed as soon as possible by students to improve its accessibility and bridge language barriers.”

These methodological approaches provide a foundation for enhancing our pipeline for transforming audio podcasts into VR TV shows, while also highlighting areas where future research and development are needed, particularly for languages with fewer resources.

2.4 Procedural Generation for Educational VR Content

Recent advances in Large Language Models (LLMs) have opened new possibilities for automated generation of VR content from textual inputs. The Text2VRScene framework described by Yin et al. [13] presents “a pioneering LLM-based automated system framework for increasing reliability and optimizing the utilization of LLMs” specifically designed for generating immersive VR experiences. This approach is particularly relevant for educational content transformation, where reliability and content fidelity are paramount.

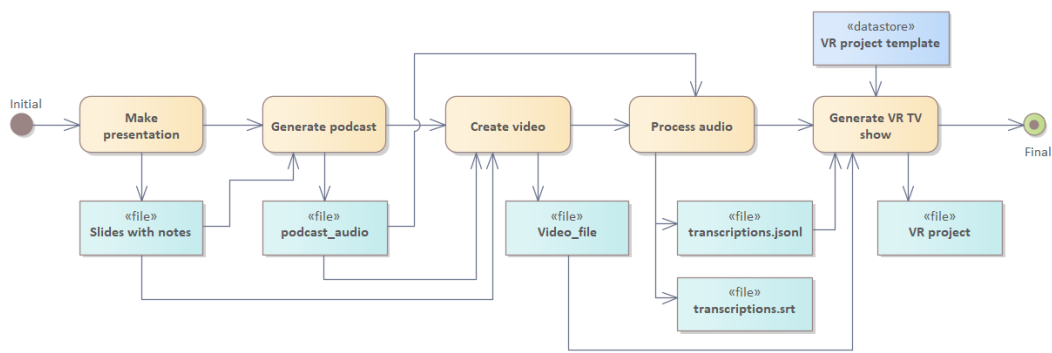
The Text2VRScene system employs a sophisticated multi-stage processing pipeline that divides complex VR generation tasks into manageable subtasks. This structured approach aligns closely with our proposed workflow for podcast-to-VR transformation, providing a formal framework for organizing the process. As Yin et al. demonstrate, LLMs “can be utilized to comprehend the user prompt about the desired VR scene, instruct the generation of digital assets through text prompts, and generate the source code that organizes different digital contents with reasonable storytelling.” This capability could be adapted to transform educational podcast content into coherent, engaging VR experiences.

A significant challenge in developing reliable LLM-based systems is maintaining context coherence throughout multiple processing stages. The researchers note that “LLMs may suffer from forgetting or conflating previous dialogue history in long-term conversations,” which could impact the accurate representation of educational content. Their proposed solution includes “Key Information Message Passing” which “aims to alleviate the risk of memory loss by distilling crucial information from the previous conversation and integrating it into the subsequent prompt.” This technique could be particularly valuable for maintaining the semantic integrity of educational content when transformed into VR format.

The system also addresses the challenge of dynamic content generation through specialized processing stages. Similar to our approach of separating diarization and transcription, the Text2VRScene framework includes distinct stages for scene description, asset generation, and animation. Their “Animation Adding” step, which “polish[es] the source code by adding animation to each 3D model according to the description of the scene,” could be adapted to create more engaging representations of speaker interactions in educational VR environments.

3 Approach

To generate VR learning objects from podcast audio files, we propose a comprehensive processing pipeline as illustrated in Figure 1. Our methodology transforms AI-generated educational podcasts into immersive VR experiences through a series of automated processing stages.



■ **Figure 1** Overview of the podcast-to-VR transformation pipeline.

We provide pseudo-code to illustrate the implementation of our solution in Algorithm 1.

3.1 Content Preparation and Podcast Generation

We recommend using presentation slides with detailed speaker notes as the primary source material. The presentation slides should follow best practices in visual design, particularly minimizing on-screen text content. The speaker notes, in contrast to the visual elements, should contain comprehensive material intended for podcast narration.

These presentation materials are uploaded to the NotebookLM platform to generate the audio podcast featuring two AI-generated hosts. The resulting audio file serves as the foundation for subsequent processing stages. At this stage, a preliminary video can be created by synchronizing the presentation slides with the generated audio using tools such as iSpringSuite, which facilitates narration timing management and supports video export functionality.

3.2 Audio Processing Workflow

The core audio processing pipeline consists of five distinct phases as shown in Figure 2. Each phase addresses specific challenges in transforming conversational audio into structured, speaker-attributed content suitable for VR integration.

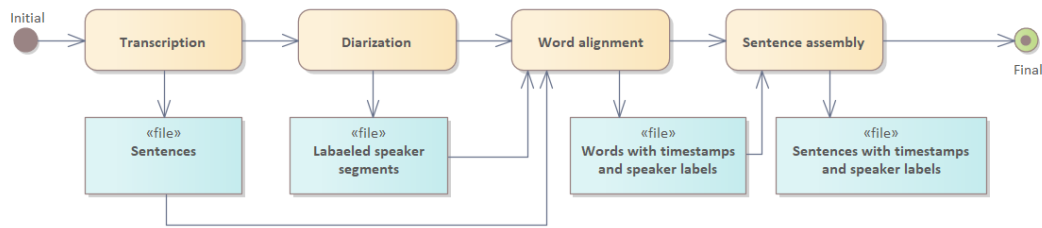
Algorithm 1 Audio to VR TV Show Transformation.

```

1: Phase 1: Audio Transcription
2: TranscribeAudio(audio_file):
3:   Load ASR model
4:   transcription  $\leftarrow$  model.transcribe(audio_file)
5:   language  $\leftarrow$  transcription.language || "en"
6:   return transcription, language
7:
8: Phase 2: Speaker Diarization
9: IdentifySpeakers(audio_file):
10:  Load diarization model
11:  diarization_result  $\leftarrow$  model.diarize(audio_file)
12:  speaker_segments  $\leftarrow$  []
13:  for each segment, track, speaker in diarization_result:
14:    Add {start, end, speaker} to speaker_segments
15:  return speaker_segments
16:
17: Phase 3: Word-Speaker Alignment
18: AlignTranscriptionWithSpeakers(transcription, speaker_segments, audio_file):
19:  Load alignment model for detected language
20:  aligned_words  $\leftarrow$  align(transcription.segments, audio_file)
21:  for each word in aligned_words:
22:    Find matching speaker in speaker_segments based on timestamp
23:    Assign speaker to word
24:  return aligned_words
25:
26: Phase 4: Sentence Assembly
27: AssembleSentences(aligned_words):
28:  speaker_sentences  $\leftarrow$  {}
29:  current_speaker  $\leftarrow$  null
30:  current_sentence  $\leftarrow$  {start: null, end: null, text: ""}
31:  for each word in aligned_words:
32:    speaker  $\leftarrow$  word.speaker
33:    // Speaker change detection
34:    if speaker  $\neq$  current_speaker:
35:      Save current_sentence to speaker_sentences[current_speaker]
36:      Start new sentence with current word
37:      current_speaker  $\leftarrow$  speaker
38:    else:
39:      Add word to current_sentence
40:      // Sentence boundary detection
41:      if word ends with period, question mark, or exclamation:
42:        Save current_sentence to speaker_sentences[current_speaker]
43:        Start new empty sentence
44:  return speaker_sentences
45:
46: Phase 5: Export to VR Environment
47: ExportToVRFormat(speaker_sentences):
48:  for each speaker, sentences in speaker_sentences:
49:    Export to JSONL format for VR integration
50:  return vr_ready_data
51:
52: Main Pipeline
53: ProcessAudioForVR(audio_file):
54:  transcription, language  $\leftarrow$  TRANSCRIBEAUDIO(audio_file)
55:  speaker_segments  $\leftarrow$  IDENTIFYSPEAKERS(audio_file)
56:  aligned_words  $\leftarrow$  ALIGNTRANSCRIPTIONWITHSPEAKERS(
57:    transcription, speaker_segments, audio_file)
58:  speaker_sentences  $\leftarrow$  ASSEMBLESENTENCES(aligned_words)
59:  vr_template  $\leftarrow$  LOADTEMPLATE(template_file)
60:  vr_data  $\leftarrow$  EXPORTTOVRFORMAT(speaker_sentences, vr_template)
61:  return vr_data

```

16:6 Converting AI-Driven Podcasts into VR Application



■ **Figure 2** Detailed audio processing pipeline.

3.2.1 Phase 1: Audio Transcription

We employ the WhisperX model from OpenAI for high-quality transcription across multiple languages. While NotebookLM currently generates podcasts exclusively in English, the transcription system is prepared for multilingual support. The transcription process produces a complete textual representation of the spoken content without speaker attribution.

3.2.2 Phase 2: Speaker Diarization

Speaker diarization identifies distinct speakers and their temporal boundaries within the audio stream. We utilize a speaker diarization model from Hugging Face that segments the audio by speaker identity, producing labeled data indicating when each speaker is active. This step is crucial for maintaining conversational structure in the final VR presentation.

3.2.3 Phase 3: Word-Speaker Alignment

The alignment phase synchronizes transcribed words with speaker segments using timestamp information. We again employ WhisperX to perform word-level alignment, matching each transcribed word with its corresponding speaker based on temporal overlap with diarization results.

3.2.4 Phase 4: Sentence Assembly

Sentence assembly organizes aligned words into coherent sentences attributed to specific speakers. The system iterates through transcribed words, using punctuation markers and speaker change detection to determine sentence boundaries. This process results in clean, speaker-attributed text segments that preserve the conversational flow of the original podcast.

3.2.5 Phase 5: Export and Integration

The processed content can be exported in multiple formats. For traditional video platforms, we generate subtitle files (.srt format) with differentiated speaker attribution, enabling clearer dialogue identification compared to standard auto-generated captions. For VR integration, the data is exported in JSON Lines (.jsonl) format for seamless integration into VR applications.

3.3 Virtual Reality Environment Implementation

To address the technical barrier of VR development for educators, we developed a template project within the Delightex (formerly CoSpaces) environment. This template includes:

- A pre-configured studio environment with appropriate lighting and staging
- Virtual speaker avatars with distinct visual characteristics
- Integrated code framework for processing exported dialogue data
- Interactive elements supporting various learning modalities



■ **Figure 3** VR TV show implementation.

Educators can adapt the template by pasting their processed .jsonl data into designated sections of the codebase. While some manual adjustments remain necessary (such as video file replacement), the template significantly reduces the technical expertise required for VR content creation.

The resulting VR learning object supports multiple learning preferences: auditory learners benefit from the original podcast audio, visual learners engage with the virtual presenters and environment, reading learners can access synchronized transcriptions, and kinesthetic learners interact with the immersive VR space through natural movement and gesture. The example of a VR show, created from PowerPoint slides, is shown in Fig. 3.

3.4 Multilingual Adaptation Challenges

Translation into target languages presents both opportunities and limitations. We investigated two primary approaches: retrofitting translated audio to existing video timelines and regenerating audio in the target language. Initial studies revealed that direct timeline synchronization is problematic due to linguistic differences in sentence structure and length between languages.

The alternative approach of generating new audio in the target language shows promise but faces quality constraints, particularly for languages with limited training data such as Lithuanian. Current text-to-speech models produce inconsistent quality for less-resourced languages, though this limitation may be addressable as language model capabilities continue to improve.

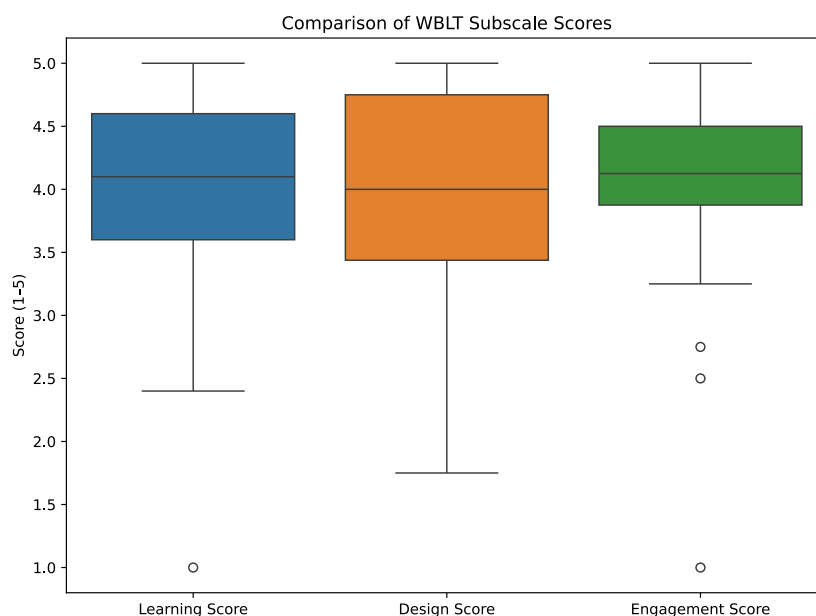
4 Experimental evaluation

To evaluate the VR show as an interactive learning object, a scale Web-Based Learning Tools (WBLT) [6] was used. Also, we evaluated the usability of the learning object using SUS (System Usability Scales) scales [3]. The learning object was used for the students to prepare for the mid-term exam. Around 200 “Software Engineering” course students used this tool for studying material. 28 students participated in a voluntary survey by answering questionnaire questions.

4.1 Evaluation of a generated VR Show as a learning object

The WBLT evaluation scale is used to subjectively assess learning tools based on three components: learning, design, and participation. In general, the scale consists of 12 statements which are grouped according to the aforementioned components. Respondents evaluated each statement using a 5-point Likert scale ranging from 1 to 5, where 1 indicates strong disagreement with the statement and 5 indicates strong agreement. The scores for each component are calculated by averaging the scores of the statements that belong to that specific category.

The internal consistency of the WBLT evaluation subscales was assessed using Cronbach's alpha, a commonly used measure of scale reliability. The Learning subscale demonstrated high reliability with an alpha of 0.827, indicating that the items used to assess learning outcomes were well-correlated and measured the same underlying construct. The Design subscale exhibited similarly strong internal consistency with an alpha of 0.852, suggesting that the questions evaluating usability, layout, and instructional coherence formed a cohesive and dependable scale. The Engagement subscale yielded an alpha of 0.723, which, while slightly lower, still represents acceptable reliability in the context of exploratory research and small sample sizes.



■ **Figure 4** WBLT subscale evaluation results.

All the subscales received similar scores - there is no significant difference. The mean scores for all five questions related to learning subscale range from 3.86 to 4.07, suggesting that most respondents agreed or strongly agreed that the tool helped them learn. The standard deviations for separate questions range from 0.94 to 1.18, indicating moderate variability in responses, with slightly greater dispersion observed for Q2 ("The feedback from the learning object helped me learn"). The interquartile range (IQR) confirms that the majority of responses fell between 4 and 5, demonstrating a consistently positive evaluation across the subscale. The learning object was generally perceived as effective in supporting learning, with particularly strong agreement on visual and conceptual support (Q3 and Q4).

The slightly lower average and higher variability in Q2 may point to an area for potential improvement in the clarity or helpfulness of feedback mechanisms. Overall, the learning subscale data suggests the tool successfully facilitates learning for most users.

The Design subscale of the WBLT instrument evaluates the structural and usability aspects of the learning object using four items (Q6–Q9). The mean scores range from 3.79 (Q9: organization) to 4.18 (Q7: clarity of instructions), suggesting respondents found the design features to be well-implemented and supportive, particularly in terms of guidance and ease of use. The standard deviations show moderate dispersion, with Q9 (organization) displaying the highest variability ($SD = 1.20$), suggesting more mixed opinions regarding how well the content was structured. All items have a median score of 4.0, with the 75th percentile reaching 5.0, indicating that many users rated the design features at the highest level. The interquartile ranges (25% to 75%) for Q6–Q8 span from 3 to 5, showing that most participants rated the design as good or excellent. We can conclude that the learning object was perceived as well-organized, easy to follow, and user-friendly by the majority of users. However, the slightly lower mean and higher variability for Q9 (“The learning object was well organized”) suggest there may be room for improvement in how the content is structured or presented.

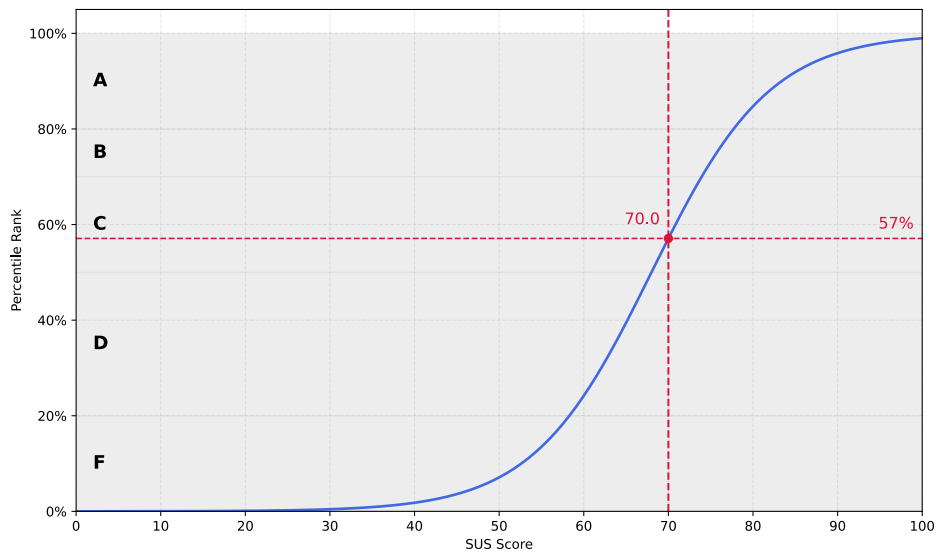
The Engagement subscale of the WBLT instrument (comprising Q10–Q13) evaluates learners’ emotional and motivational responses to the learning object. The mean scores range from 3.89 (Q10 and Q11: liking the theme and perceived engagement) to 4.11 (Q12: learning was fun), showing that learners found the tool appealing and enjoyable to use. All items have a median of 4.0, and the 75th percentile reaches 5.0 for all questions, indicating that at least a quarter of participants gave the highest possible rating. Standard deviations range from 1.03 to 1.40, showing moderate variability in responses, with the most variability in Q10 (theme appreciation). All items received the full range of responses (1–5), suggesting that while many learners were highly engaged, some did not find the tool engaging or enjoyable. The results demonstrate that most users found the learning object motivating, enjoyable, and worth reusing. The high scores on Q12 and Q13 especially indicate that the tool succeeded in making learning fun and repeatable. However, the wider spread in Q10 and Q11 may reflect differing preferences regarding the theme or emotional resonance of the tool’s content.

4.2 Generated learning object usability evaluation

The SUS (System Usability Scale) is a tool used to determine a system’s usability score, derived from users’ subjective evaluations. The scale consists of 10 statements that alternate between positive and negative phrasing, and users rate the system based on these statements using a 5-point Likert scale, where 1 indicates strong disagreement, and 5 indicates strong agreement. The final SUS score is calculated by subtracting 1 from the rating of each positively worded item, and inverting the ratings of negatively worded items by subtracting the given score from 5. All adjusted values are then summed and multiplied by 2.5, resulting in a final score ranging from 0 to 100. The SUS usability score allows the system or its features to be compared with other systems in terms of usability or with previous iterations of the same system.

The obtained System Usability Scale (SUS) score for the tool was 70.0, indicating a moderate level of perceived usability. When converted into percentile ranks based on normalized SUS benchmark data, this score corresponds to approximately the 50th to 60th percentile. This means that the tool performed on par with or slightly above average, demonstrating a usable system with room for improvement.

16:10 Converting AI-Driven Podcasts into VR Application



■ **Figure 5** SUS evaluation results.

In terms of grading, the SUS score of 70.0 falls within the C grade range, according to the curved grading scale developed by Sauro and Lewis [10]. This suggests acceptable but not exceptional usability performance. While the system meets baseline usability expectations, refinements could enhance user satisfaction and efficiency.

The adjective rating associated with this SUS score is “Good”, consistent with the findings of Bangor et al. [2], who mapped SUS scores to qualitative usability descriptors. A “Good” rating suggests that users found the tool generally usable and functional for its intended purpose.

From the perspective of acceptability, the score of 70.0 places the system at the beginning of the “acceptable” range. According to the SUS interpretation framework, scores above 70 are considered clearly acceptable, those between 50 and 70 are marginally acceptable, and scores below 50 are deemed unacceptable. Thus, the learning object demonstrates sufficient usability for practical usage.

5 Discussion and conclusions

In this paper, we presented an innovative approach for converting AI-generated educational podcasts from PowerPoint slides into immersive virtual reality (VR) learning experiences. Currently, the process is not fully automated. Manual work is still required to prepare the video file, and some coding in the Python programming language is necessary to add custom questions. Our ongoing research is focused on completely eliminating manual work.

Videos can be created automatically from slides. The main challenge lies in automatically detecting which slide is being discussed by the podcast hosts. This could be addressed using large language models (LLMs). Alternatively, more efficient algorithms may be available for audio-text classification, matching the spoken content with the corresponding slide notes.

Another issue we need to address is the environment used for our prototype. The Delightex platform is easy to use and allows for rapid creation of prototypes. However, its execution performance is not sufficient. Some students reported lag on lower-end devices. Additionally, Delightex lacks interoperability – it cannot report student activities to learning management

systems such as Moodle. Therefore, we plan to develop our own VR studio template using game engines that support XR in web browsers (such as Godot or Unity). We will use the SCORM standard to report student activities to learning management systems. This approach will enable a fully automated pipeline and address the custom coding limitation by providing an interface for question creation.

Besides technical challenges, the adoption of AI technologies in educational content generation raises important ethical concerns. The use of Large Language Models (LLMs) presents questions regarding the sourcing of training material and potential copyright infringement in AI-generated educational content [5]. Additionally, content validation by content creators remains essential, even when AI-generated material involves summarization of provided content, translation, or reformulation. Generative AI models produce content based on statistical patterns from massive datasets but lack true understanding, which may lead to factual inaccuracies and fabricated sources.

While there are challenges to overcome, the practical application of this approach in a software engineering course confirmed both its feasibility and pedagogical value. The implemented system supports a range of learning preferences – auditory, visual, reading, and kinesthetic (when used in a VR headset). Evaluation results using the Web-Based Learning Tools (WBLT) instrument indicated satisfaction in the learning and design dimensions, and strong engagement. However, expanding interactivity and improving feedback mechanisms could further enhance learning effectiveness. Although the overall System Usability Scale (SUS) score of 70.0 places the system within the “acceptable” usability range, this is another area that can be significantly improved.

References

- 1 Saadia Ayub, Asiya Karim, and Amina Laraib. Learning styles of medical students. *The Professional Medical Journal*, 30(09):1214–1218, 2023. doi:10.29309/TPMJ/2023.30.09.7650.
- 2 Aaron Bangor, Philip T. Kortum, and James T. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, 2009.
- 3 John Brooke. *SUS – a quick and dirty usability scale*, pages 189–194. Taylor and Francis, January 1996.
- 4 R.V. Mahendra Gowda. Education 5.0: Evolution of promising digital technologies – a comprehensive review. *International Journal of Advanced Science and Engineering*, 10(2):3442–3448, 2023. doi:10.29294/ijase.10.2.2023.3422–3448.
- 5 Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79, 2023.
- 6 Robin Kay. Evaluating learning, design, and engagement in web-based learning tools (wblts): The wblt evaluation scale. *COMPUTERS IN HUMAN BEHAVIOR*, 27(5):1849–1856, September 2011. doi:10.1016/j.chb.2011.04.007.
- 7 Subhankar Maity and Aniket Deroy. Generative ai and its impact on personalized intelligent tutoring systems. *arXiv preprint arXiv:2410.10650*, 2024. doi:10.35542/osf.io/kawr5.
- 8 Kazunori Matsumoto. Exploring patterns of generative ai utilization in education. *IIAI Letters on Informatics and Interdisciplinary Research*, 4, 2023. doi:10.52731/liir.v004.134.
- 9 A. Pérez, G. Garcés Díaz-Muniñ, A. Giménez, J. A. Silvestre-Cerdà, A. Sanchis, J. Civera, M. Jiménez, C. Turró, and A. Juan. Towards cross-lingual voice cloning in higher education. *Engineering Applications of Artificial Intelligence*, 105:104413, 2021. doi:10.1016/j.engappai.2021.104413.
- 10 Jeff Sauro and Jim Lewis. How to interpret a sus score, 2023. Accessed: 2025-05-17. URL: <https://measuringu.com/interpret-sus-score/>.

16:12 Converting AI-Driven Podcasts into VR Application

- 11 Peilin Wu. Beyond audio: Advancing speaker diarization with text-based methodologies and comprehensive evaluation. Honors thesis, Emory University, 2024. URL: <https://etd.library.emory.edu/concern/etds/rb68xd387>.
- 12 T. Xu, Y. Liu, Y. Jin, Y. Qu, J. Bai, W. Zhang, and Y. Zhou. From recorded to ai-generated instructional videos: A comparison of learning performance and experience. *British Journal of Educational Technology*, 2024. doi:10.1111/bjet.13530.
- 13 Z. Yin, Y. Wang, T. Papatheodorou, and P. Hui. Text2vrscene: Exploring the framework of automated text-driven generation system for vr experience. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 701–711, 2024. doi:10.1109/VR58804.2024.00090.