Human-Computer Interaction

Navigation of Audio Files

Authors
Rouven Anderer, uipno@student.kit.edu
Andreas Hangg, andreas.hangg@etu.univ-grenoble-alpes.fr

Date:
Grenoble, January 15, 2026

1 Introduction

The fundamental challenge of human-computer interaction is dealing with the limited bandwidth of data that the human perception can process. Biological studies show that visual processing is the most efficient and optimized, followed by audio. [8]

In recent years, voice messages have become more and more common, reaching 13 voices messages per user and day on WhatsApp in 2014. [7] This can mainly be explained by the fact that voice messages combine the possibility of asynchronous communication of text messages with the natural expression of a direct call. [1] Haas et al. have determined four further reasons in a user study, which are convenience, para-linguistic features, situational constraints, and receiver preferences or handicaps. [4]

However, voice messages pose two major challenges to users. Firstly, they only make use of the second most efficient human processing channel, the ears, and not the most efficient being the eyes. We argue that this issue is of lower priority, since the processing rate of human brains is much lower than the information extraction of human senses anyways, as has been famously suggested by Zheng et al. [8] The second problem, the lack of navigability, is the more important one. When users try to find information for a specific topic in an audio message, for instance when revisiting it for information they have forgotten, there is no possibility of skimming it like a text message. This leads to slow information extraction and possible user frustration.

Other studies have done research into the same problem, but with a focus on podcasts. Park et al. compared keyword search, topic segmentation and image segmentation, all of which were created using generative artificial intelligence, in a user study with 12 participants. [6] In a similar approach, Zhi et al. developed a user interface allowing for audio content browsing, topic-based and keyword-based navigation, communication of transcript and speaker information in real time, and content-based query of podcasts. [9] While these papers study a similar challenge, the circumstances of a podcast are slightly different in that the creator is usually more considerate about the structure of the audio file before creating it and also willing to invest more effort in the specification of timestamps, topics etc. after its creation.

In this work, we conduct a user study in which we apply similar techniques to voice messages. We compare two different approaches:

1. The first one is similar to the before-mentioned methods, where the voice message segmented by topic and the user is given a list of titles, small descriptions and the according voice message segments (referred to as *by-topic* in this work).

2. Additionally, we propose a new method, in which the user is given key points in the voice message with the according timestamps (referred to as *by-keypoint* in this work). Other studies have shown that such key points can be extracted automatically. [3] [5] In this work, we focus on the human-computer interaction and, therefore, synthesize this data.

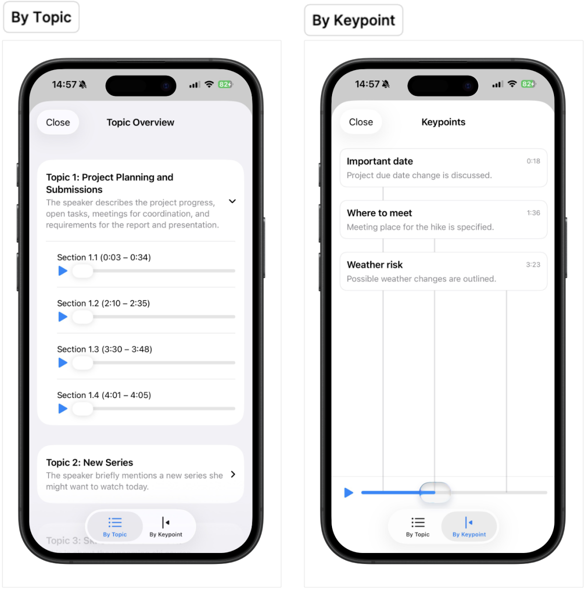
In summary, this leaves us with the following research question: Can users reach a higher information extraction performance and/or user satisfaction using keypoint audio file navigation in comparison to by-topic navigation?

2 Methodology

As the basis for our user study, we created two voice message with a duration of about 4:30 minutes containing several different topics, topic switches and information.

2.1 User Interface

To navigate this audio file, we developed an IOS app providing the two different navigation approaches by-topic and by-keypoint. Figure 1 depicts the structure of this user interface.



Topic 1: Project Planning and Submissions

The speaker describes the project progress, open tasks, meetings for coordination, and requirements for the report and presentation.

Topic 2: New Series

The speaker briefly mentions a new series she might want to watch today.

Important date

Project due date change is discussed.

Where to meet

Meeting place for the hike is specified.

Weather risk

Possible weather changes are outlined.

Figure 1: **The left side shows the by-topic navigation, where segments are sorted into topics and can be started individually. The right side shows the by-keypoint navigation, where the user can jump to keypoints in the whole audio file. Keypoints and topics are example data.**

2.2 User Study Setup

In this section, we describe the pre- and postconditions as well as the experiment setup.

2.2.1 Pre-Experiment Survey

For later evaluation, we surveyed the following properties of all participants:

1. Age
2. Gender

Additionally, every participant was given an explanation of the experimental setup, the quantitative and qualitative criteria as well as the chance to bring forward any questions.

2.2.2 Training

To get an understanding of the user interface, the participants were given a short introduction to both methods and were able to try them out on the target device. Given the simplicity of the user interface, no detailed training was required and all participants confirmed their basic comprehension of the system within less than a minute.

2.2.3 Experiment

The audio files are crafted to differ in topics, but to follow the same pattern of topic switches and similar types of unnecessary and required information (Structure shown in Figure 2). This allowed us to let users try out both methods without already knowing the content of the second file while still keeping the results comparable.

Each participant was given the same Iphone 16 Pro with the application opened. After a brief demonstration of how the navigation fundamentally works, a timer was started, which the participant was made aware of. Once the participant had finished the assignment, they could stop the timer themselves and subsequently hand in the results without further changes.

To equalize the potential training bias, we separated the participants in four groups, where each group starts with a different audio file and method first.

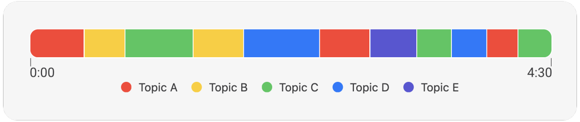


Figure 2: **The visualization maps the inherent structural complexity of the voice messages, illustrating the 'thematic chaos' that participants had to navigate during the study.**

2.2.4 Objective Criteria

The main goal of the participant is the retrieval of certain information following the following quantitative goals:

1. Accuracy
2. Speed

The speed is a simple measurement from the handout of the phone to when the user hands in their results. The accuracy is measured based on the users responses to the synthesized questions, which we provide more details in the next section.

2.2.5 Subjective Criteria

After each experiment, we handed out another survey to the participants to measure the following criteria. For these, we used a subset of the suggested user satisfaction survey by Chin et al.[2]:

1. Quantitative: Overall reaction to the software (terrible vs. wonderful; difficult vs. easy; frustrating vs. satisfying; dull vs. stimulating; confusing vs. clear organization of information) on a scale of 1 to 9 per method

2. Qualitative: *Did you encounter any difficulties with either of the methods?*

3. Qualitative: *Describe in your own words if you preferred one of the navigation methods and if so, why?*

2.2.6 Termination Criteria

For the termination criteria, we decided to continue asking participants until for both methods, the average time performance change is below 3% for three consecutive new experiments while also having equal group sizes. This ended up being the case after 16 participants.

2.2.7 Success Criteria

Only results fulfilling the following criteria were counted as valid:

1. The participant has confirmed to having understood all information about the experiment in the pre-experiment survey without further questions.

2. The experiment phase was executed without any technical issues or other interruptions.

3 Results and Discussion

In this section, we lay out and illustrate the results and biases of the experiments.

3.1 Quantitative Results

Figure 3 shows that for each case (audio file 1, audio file 2 and combined), the by-topic method performed better on average and has a smaller spread of results. In all three cases, the by-keypoint method achieves the best but also the worst result.

Across all participants, only 3 out of 192 answered questions were incorrect, resulting in an overall accuracy of 98.4%. Two of these errors occurred using the by-keypoint method and one using the by-topic method. Given the extremely high accuracy across both methods, these errors are considered outliers rather than indicative of systematic performance differences between the navigation approaches.

Similarly to the performance results, Figure 4 indicates that the participants perceive the by-keypoint method slightly worse than the by-topic one across the board. Both the averages and the extrema are equal or worse for all four dimensions. For the by-topic method, all averages lie between 8 and 9.5, whereas for the by-keypoint one they are between 5 and 8.

3.2 Qualitative Results

Within the "faced difficulties" question, all five responses we received concerned the by-keypoint method. One participant thought his initial assumption that not every information may be contained in the keypoints was actually wrong after the experiment. This assumption was indeed correct, though, which we could immediately clarify. All four other responses mention the disadvantage of the keypoint method that it follows the chaotic structure of the audio message, meaning that keypoints belonging to the same topic may be spread all over the audio message.

For the "preferred method" question, we received nine responses. Five of these preferred the by-topic method for its clearer structure and for the security of not missing any information regarding a selected topic. Two participants mention the same advantage of the topic method, but also express that they see the keypoint method as potentially more effective, especially if the audio message has been listened to before. The last two participants preferred the by-keypoint method, one of them because of the simpler user interface and the other because the audio message felt like a more natural conversation to them.

3.3 Biases

Out of 16 participants, 13 were male and 3 were female. Twelve out of 16 participants were 30 years old or younger, the rest was older. This shows a bias towards male and younger participants in the experiment. Unfortunately, due to time and resource restrictions, a representative selection of participants was not achievable. However, we ensured that every participant estimated their level in English to be at least B2 and that they were experienced in using the target device.

4 Summary and Conclusion

In summary, the by-topic method performed better in the objective and subjective quantitative as well as the qualitative results. We came up with the following main observations explaining this:

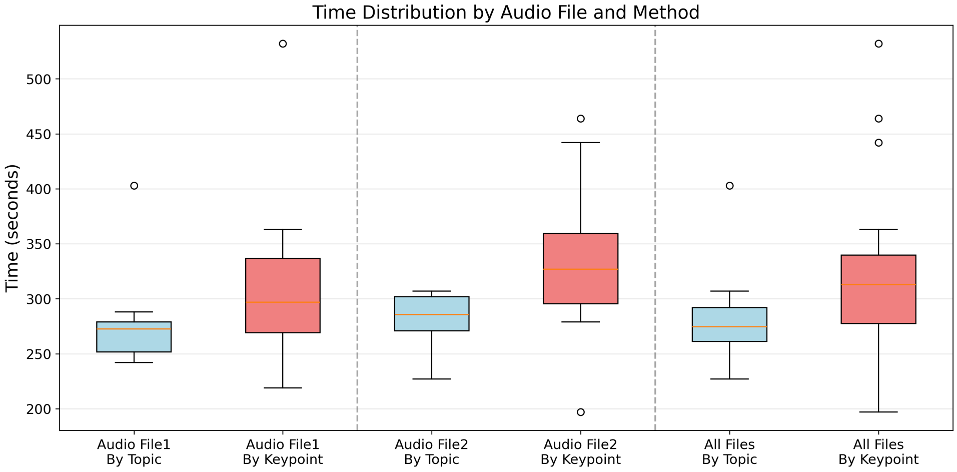


Figure 3: **This figure shows the time it took participants to complete all tasks. Each pair of two boxplots show the difference by method, where the first pair belongs to audio file 1, the second to audio file 2, and the third to the combined result. (Within box: 25th-75th percentile; whiskers: 1.5 times interquartile range below 25th and above 75th percentile; orange line: median; points: outliers; n=16)**

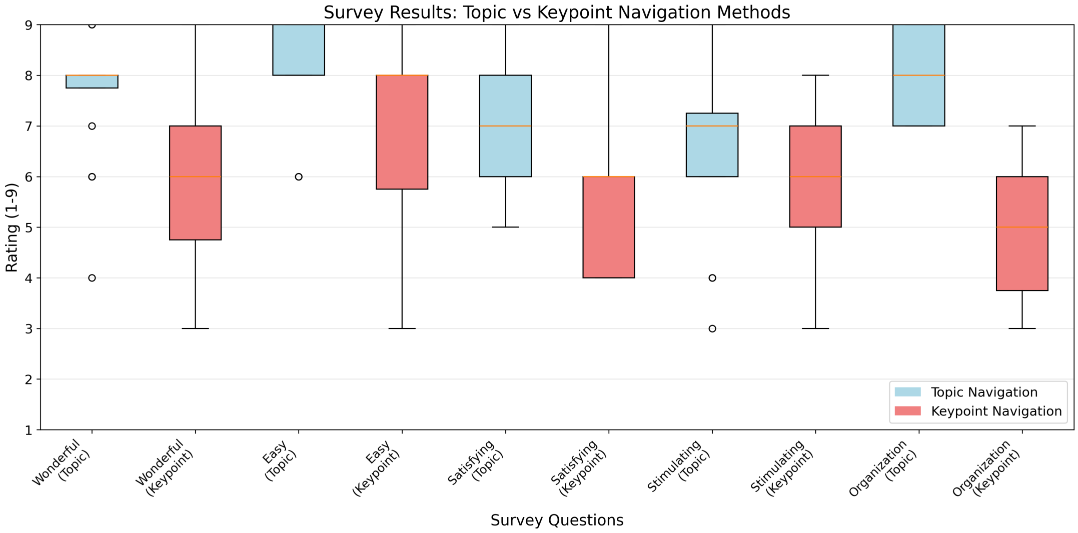


Figure 4: **This figure shows the quantitative results of the user survey regarding both methods. (Within box: 25th-75th percentile; whiskers: 1.5 times interquartile range below 25th and above 75th percentile; orange line: median; points: outliers; n=16)**

1. Clearer structure: The approach that all the participants took using the by-topic method was very similar and straightforward. First, they checked which topics were relevant for the questions, then they went through them one-by-one. From our observations and the qualitative feedback, this leads to a smaller spread in results and a higher user satisfaction due to a reduced amount of stress and confusion.

2. Experience: The by-topic user interface was mentioned to be more intuitive by multiple users, even though the by-keypoint is simpler from an objective point of view. Therefore, we suspect the former to be something users are more used to.

Nevertheless, there are some points to make for the by-keypoint method as well:

1. Higher potential: The spread and the better maximum results in the performance results of the by-keypoint method indicate the potential to perform better, even if the average results are worse in our experiments.

2. Experimental setup: In hindsight, we suspect the experimental setup to slightly favor the by-topic method. Using this approach, users are

never faced with the chaotic nature of the original audio file, which is different for the by-keypoint method. In future research, another experiment setup could let users listen to the audio file without knowledge of the questions, and then let them use the methods 15 minutes later to respond to the questions. In this setup, the higher maximum potential of the by-keypoint method could become more evident.

To leap back to our research question, our experiment could not show a better performance in information extraction on average, nor a higher user satisfaction when using the by-keypoint method. However, having the best maxima values, it did show the potential for a higher performance, which could be investigated in further, slightly adjusted experiments, especially in less fragmented voice messages where the structural advantage of the by-topic method fades.

5 Outlook

Another interesting method could be the combination of both methods. The clearer structure of the by-topic method could be extended by keypoints within the regarding topic, therefore offering a defragmented environment for the user to work with keypoints.

Additionally, an adaptive approach could automatically detect the structural complexity of voice messages. For coherent messages with few topic switches, the topic segmentation overhead could be skipped entirely, allowing users to navigate directly via keypoints. This would leverage the efficiency potential of the by-keypoint method while reserving the defragmentation benefits of the by-topic approach for structurally chaotic messages like those tested in our experiments.

References

[1] Barbara L. Chalfonte, Robert S. Fish, and Robert E. Kraut. "Expressive richness: a comparison of speech and text as media for revision". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '91. New Orleans, Louisiana, USA: Association for Computing Machinery, 1991, pp. 21–26. ISBN: 0897913833. DOI: 10.1145/108844.108848.

[2] J. P. Chin, V. A. Diehl, and L. K. Norman. "Development of an instrument measuring user satisfaction of the human-computer interface". In: *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '88*. the SIGCHI conference. Washington, D.C., United States: ACM Press, 1988, pp. 213–218. ISBN: 978-0-201-14237-2. DOI: 10.1145/57167.57203.

[3] Carlos-Emiliano González-Gallardo et al. *Audio Summarization with Audio Features and Probability Distribution Divergence*. Apr. 2, 2020. DOI: 10.48550/arXiv.2001.07098. arXiv: 2001.07098[cs].

[4] Gabriel Haas et al. ""They Like to Hear My Voice": Exploring Usage Behavior in Speech-Based Mobile Instant Messaging". In: *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI '20: 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services. Oldenburg Germany: ACM, Oct. 5, 2020, pp. 1–10. ISBN: 978-1-4503-7516-0. DOI: 10.1145/3379503.3403561.

[5] S Swaroop Kaushik and Sanjit Kangovi. "Automated Extraction and Augmentation of Key Information from Audio using Speech Recognition and Text Summarization". In: 8.10 (2023).

[6] Jimin Park et al. "Enhancing the Podcast Browsing Experience through Topic Segmentation and Visualization with Generative AI". In: *ACM International Conference on Interactive Media Experiences*. IMX '24: ACM International Conference on Interactive Media Experiences. Stockholm Sweden: ACM, June 7, 2024, pp. 117–128. ISBN: 979-8-4007-0503-8. DOI: 10.1145/3639701.3656324.

[7] Felix Richter. *An Average WhatsApp User Sends ¿1,000 Messages per Month*. 2014. DOI: https://www.statista.com/chart/1938/monthly-whatsapp-usage-peruser/.

[8] Jieyu Zheng and Markus Meister. *The Unbearable Slowness of Being: Why do we live at 10 bits/s?* Nov. 15, 2024. DOI: 10.48550/arXiv.2408.10234. arXiv: 2408.10234[q-bio].

[9] Qiyu Zhi et al. "VisPod: Content-Based Audio Visual Navigation". In: *Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces*. IUI'18: 23rd International Conference on Intelligent User Interfaces. Tokyo Japan: ACM, Mar. 5, 2018, pp. 1–2. ISBN: 978-1-4503-5571-1. DOI: 10.1145/3180308.3180318.