VoiceCogs: Interlocking Concurrent Voices for Separable Compressed Browsing with Screen Readers

Jeongwon Choi
choijw@postech.ac.kr
POSTECH
Pohang, Gyeongbuk, South Korea

Inseok Hwang
i.hwang@postech.ac.kr
POSTECH
Pohang, Gyeongbuk, South Korea

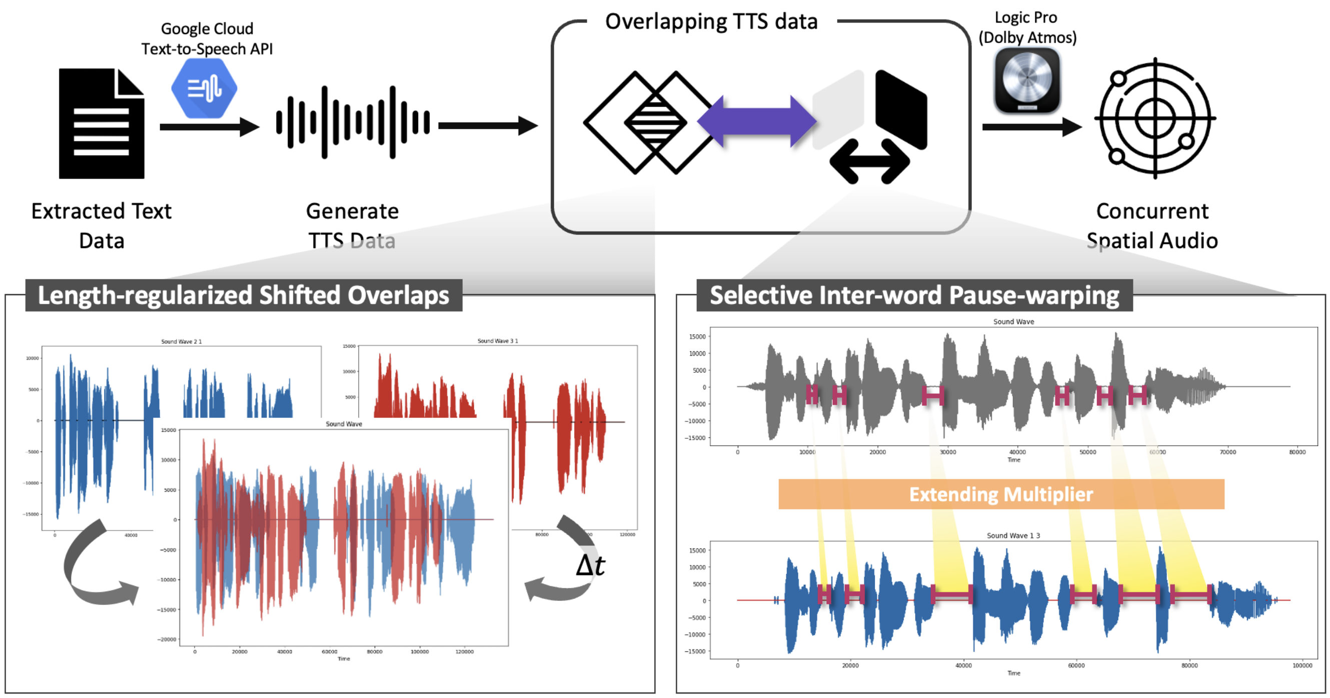


Figure 1: VoiceCogs System Flow Chart

ABSTRACT

Ensuring universal accessibility to information cannot be overstated. Unlike visual readers, however, screen reader users are given inefficient and restricted channels to acquire the given information. In particular, we focus on the initial step of information acquisition – *quickly* scanning the overall structure of a textual document so that the reader makes an informed decision about where to jump and read the details. While this step is inherently quick for visual users, screen reader users passively listen to the slow, sequential list of items read aloud. To close this gap, we call for a technique that accelerates screen reader users' scanning process. Our system, VoiceCogs, takes multi-itemed text sources and synthesizes audio that concurrently plays multiple text-to-speech from a respective text source while facilitating the discernibility of individual sources. To this end, we devise and implement two interlocking techniques to minimize phonetic interferences between concurrent speeches.

CCS CONCEPTS

• **Human-centered computing → Accessibility systems and tools**; *Sound-based input / output*; *Interactive systems and tools.*

KEYWORDS

concurrent speech; spatial audio; screen reader; accessibility

ACM Reference Format:

Jeongwon Choi and Inseok Hwang. 2023. VoiceCogs: Interlocking Concurrent Voices for Separable Compressed Browsing with Screen Readers. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct), October 08–12, 2023, Cancun, Quintana Roo, Mexico.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3594739.3610716

UbiComp/ISWC '23 Adjunct, October 08–12, 2023, Cancun, Quintana Roo, Mexico

ACM ISBN 979-8-4007-0200-6/23/10.
https://doi.org/10.1145/3594739.3610716

1 INTRODUCTION

Upon seeing a newspaper or webpage, in which order would you browse the contents? Perhaps you would capture the page structure by quickly scanning the titles and headlines, then jump on the text block that interests you. Here, your vision gets the initial scanning done very fast; the concurrent presence of spatially distributed multiple visual items hardly interferes with each other [22]. The proliferation of VR is making the viewing space even larger [7]. You can even control the scanning order by leveraging the spatial proximity of text blocks.

Unfortunately, people with visual impairment who rely on screen readers neither have such fast scanning nor controllability of the scanning order. In general, listening is much slower than viewing for grasping the content [8, 9, 16, 17, 19]; it takes much more time for screen reader users to go through the list of titles and headlines than visual readers to do the same by eyes. Furthermore, screen reader users have little freedom to deviate from the order that the screen reader reads aloud along the page structure. Eventually, screen reader users have inherent disadvantages in terms of time-efficiently browsing structured texts.

We propose VoiceCogs, a screen reader system that facilitates *compressed browsing* of structured texts by overlapping multiple text phrases while keeping the aural discernibility high. As it is well-known that simultaneously speaking while another is speaking is discouraged for social and intelligibility reasons [14, 15, 18], we develop VoiceCogs to address the intelligibility challenges originating from overlapping.

VoiceCogs interlocks multiple text phrases in a way that *minimizes their syllabic overlaps*. In particular, we devise two techniques: (1) length-regularized shifted overlaps, and (2) selective inter-word pause-warping. VoiceCogs firstly adjust the time offset of each text phrase to be read aloud so that the text phrases are overlapped in time but little coincide in the syllabic levels, while regularizing the time offset to control the total length of the overlapped speech. To reduce syllabic collision even fewer under a given total length limit, we selectively warp inter-word gaps where otherwise the syllables from different phrases momentarily overlap. To further improve the discernibility of overlapped voices, VoiceCogs additionally employs spatial audio [1, 6, 11, 12, 20] and varying individual voices.

In this paper, we present the key algorithms and the implementation of VoiceCogs, followed by the evaluation results from a pilot study with 15 participants.

2 RELATED WORK

Screen readers are widely used to provide an alternative user interface to textual contents for people with visual impairment [10, 21].

The human cognitive system has the ability of selective attention, i.e., focusing on the specific content while simultaneously accommodating multiple external sounds at the same time, also known as Cocktail Party Effect [4, 5] where people can selectively listen to and understand one talker in a multi-talker situation. Additionally, it is known that the ability to seperate concurrent speeches can be improved by learning [23]. The design rationale of VoiceCogs is analogous to such inherent human ability of selective attention in multi-talker situations, on top of which we add microscopic temporal alterations to each independent speech source.

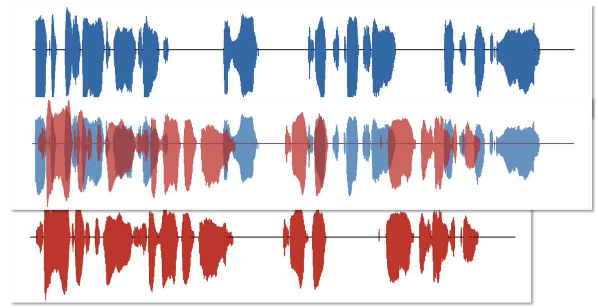


Figure 2: Length-regularized Shifting Example

3 VoiceCogs

3.1 Design

Figure 1 depicts the overall flow of VoiceCogs system. First, we extract the target page's titles and headings, followed by TTS generating [2] the audio that reads aloud individual text phrases. Then, we apply multiple iterations of *length-regularized shifted overlaps* and *selective inter-word pause-warping* to reach an optimal balance between the syllabic overlaps and the total length of the overlapped speech. Finally, we use a DAW (Digial Audio Workstation) to synthesize the spatial audio effect that mimics the independent voices spatially separated as desired.

3.2 Optimizing Voice Overlaps

In the process of overlapping multiple voices, our algorithm finds an appropriate time offset Δt for each voice source to minimize syllabic overlaps with other concurrent voices, so that the discernibility of each voice can be improved. Formally, our algorithm solves the following optimization problem, Eq. (1):

$$\underset{\Delta t}{\text{argmin}} \left[\frac{1}{N} \sum_{t=0}^{N} \{E_1(t) + E_2(t+\Delta t)\} + \lambda|\Delta t| \right]$$
$$\text{where} \begin{cases} N & \text{: total length of audio} \\ s_i(t) & \text{: audio signal of i-th voice} \\ E_i(n) = \sum_{t=0}^{n} s(t)^2 & \text{: energy of i-th voice} \end{cases} \quad (1)$$

To further increase the effectiveness of the algorithm mentioned in Eq. (1), we run multiple iterations of alternating 'length-regularized shifted overlaps' and 'selective inter-word pause-warping'. As shown in the Figure 2, 'length-regularized shifted overlaps' is a shifting method that minimized the overlap between the concurrent audios. And 'selective inter-word pause warping' functions to warp the pause between words without losing context and identification. We empirically set the lower- and upper-bounds of inter-word pauses. Then, at each iteration, we apply small perturbations to the inter-word pauses within the bound, followed by re-running the optimization problem (1).

Condition	# of subpage titles	# of unique voices	# of concurrent audio	Spatiality	Length-regularized shift	Inter-word pause warping	mean Likert Scale (μ)	std (σ)
C1.1	2	2	2	X	X	X	3.87	1.06
C1.2	2	2	2	O	X	X	4.47	0.74
C1.3	2	2	2	O	O	X	3.87	0.83
C1.4	2	2	2	O	X	O	3.93	1.10
C1.5	2	2	2	O	O	O	4.33	0.72
C2.1	4	4	4	X	X	X	1.00	0.00
C2.2	4	4	4	O	X	X	1.73	0.70
C2.3	4	4	4	O	O	X	3.00	0.85
C2.4	4	4	4	O	X	O	2.47	0.83
C2.5	4	4	4	O	O	O	3.20	0.86

Table 1: Conditions for tasks browsing subpage titles.

Condition	# of contents	# of unique voices	# of concurrent audio	Spatiality	Length-regularized shift	Inter-word pause warping	mean play time (μ)	std (σ)
C3.1	8	2	0 (linear)	X	X	X	16.00	7.41
C3.2	8	2	2	O	X	X	9.07	4.13
C3.3	8	2	2	O	O	O	10.8	5.17

Table 2: Conditions for tasks browsing an auto-attendant phone system's menus.

3.3 Implementation

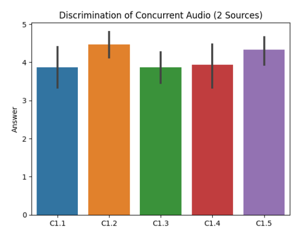
We use Google Cloud Text-to-Speech API as the TTS. With Speech Synthesis Markup Language (SSML), we control the detailed elements of generated voice, such as pronunciation, volume, pitch, emphasis, and rate. We employ Apple Logic Pro [3] as our DAW (Digital Audio Workstation), where we apply Dolby Atmos plug-in – one of the most popular surround sound technologies. It provides a virtual space where each independent sound source can be placed at a specific location.

4 PILOT STUDY

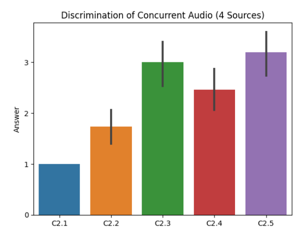
To evaluate the discernibility of the VoiceCogs-generated concurrent voices, we designed a pilot study with a group of voluntary participants and itemized lists of text highlights (e.g., section headings, subpage titles) from a structured text. We recruited a total of 15 university students (F=3, M=12) with an age range of 20 to 29 (mean=23.5, std=2.7). Based on the literature that reported no significant difference in discriminating concurrent audio between the visually impaired and the non-impaired [12], we conducted the pilot study with the participants who are visually non-impaired.

We designed two sets of sample tasks – (1) browsing the subpage titles listed in a university bulletin page, and (2) browsing the menus of an auto-attendant phone system. In each task, we varied the number of unique voice tones and concurrent voices, as well as the ablations of individual techniques that we proposed. Table 1 and 2 orgnaize the detailed experimental conditions and results in each task.

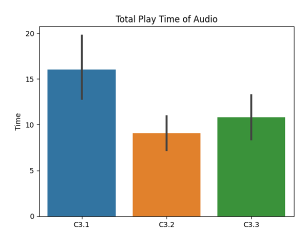
We received responses about the intelligibility of the concurrent voices on a 5-point Likert scale. The mean value in Table 1 implies the degree of discrimination of overlapping voices, and Table 2 implies the average time spent browsing. The detailed results for each task are shown in the Figure 3. The pilot study has shown that hearing multiple menus concurrently with VoiceCogs reduced the average time required to find the desired menu. Also, applying overlapping techniques increases the discrimination, especially in four audio source situation.



(a) Discrimination of overlapping voices: 2 subpage titles situation (5-point Likert Scale)



(b) Discrimination of overlapping voices: 4 subpage titles situation (5-point Likert Scale)



(c) Total Play time of Audio: browsing an auto-attendant phone system's menus (total 8 menus)

Figure 3: Pilot Study Result

P11 said, "*The sounds don't feel completely overlapping. It felt more listenable than the previous ones.*" after hearing audio that applied full techniques: pitch of voice, spatiality, length-regularized shift, and inter-word pause-warping. Also, P9 mentioned improving their ability to identify simultaneous speech through practicing.

5 CONCLUSION AND FUTURE WORK

In this research, we developed VoiceCogs – a screen reader system that facilitates *compressed browsing* of structured texts by interlocking multiple voices of text phrases in a way that minimizes their syllabic overlaps and thereby keeping the aural discernibility high. In this light, we devised two techniques in VoiceCogs, namely length-regularized shifted overlaps and selective inter-word pause-warping. Throughout a pilot study, we obtained the preliminary evaluation results that shed light on the usability and time efficiency of VoiceCogs.

For future works, we can enhance VoiceCogs by taking account of additional dimensions in controlling the syllabic overlaps and evaluating their discernibility, such as phonetic or prosodic metrics. We may also introduce a real-time adaptation of independent voice sources' virtual spatial placement to reflect the user's potential real-time interest changes. When there are too many items to be compressed into one overlapping group, we may adopt a probabilistic model to find a tradeoff between compression density and intelligibility [13]. Also, we need to diversify the demographics of our participant pool, e.g., to include the visually impaired population.

6 DEMO PLAN

We expect that a screen reader generating concurrent speeches might not be familiar to most users. For easy understanding, listening directly to our VoiceCogs system is essential. In order to guarantee a clear experience, we will prepare headphones with noise canceling functions.

For a demo, we will provide an interface for participants to compare the degree of discrimination of the synthesized overlapping voices. First, participants will be asked to select sample sentences to be overlapped. Based on these samples, our system synthesizes speeches applying each technique one by one. Participants will then listen to the synthesized speech using headphones and evaluate the improvement in scanning efficiency and the intelligibility of the speech between outputs.

ACKNOWLEDGMENTS

This research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021R1A2C200386612). This research was also partly supported by the IITP grants by the MSIT, Korea (IITP-2020-0-01778, RS-2023-00229171) and by the Alchemist Project NTIS1415187366(20025752) funded by the Ministry of Trade, Industry & Energy(MOTIE, Korea).

REFERENCES

[1] 2012. Dolby Atmos - Official Site. Retrieved June, 2, 2023 from https://www.dolby.com/technologies/dolby-atmos/

[2] 2014. Text-to-Speech. Retrieved June, 2, 2023 from https://cloud.google.com/text-to-speech

[3] 2023. Overview of the Dolby Atmos plug-in in Logic Pro for Mac. Retrieved June, 2, 2023 from https://support.apple.com/guide/logicpro/dolby-atmos-plug-in-overview-lgcpad99a338/mac

[4] Adelbert W Bronkhorst. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica* 86, 1 (2000), 117–128.

[5] E Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America* 25, 5 (1953), 975–979.

[6] Sungjae Cho, Yoonsu Kim, Jaewoong Jang, and Inseok Hwang. 2023. AI-to-Human Actuation: Boosting Unmodified AI's Robustness by Proactively Inducing Favorable Human Sensing Conditions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–32.

[7] Sungjae Cho, Jungeun Lee, and Inseok Hwang. 2022. TouchVR: A Modality for Instant VR Experience. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–3.

[8] Woohyeok Choi, Jeungmin Oh, Taiwoo Park, Seongjun Kang, Miri Moon, Uichin Lee, Inseok Hwang, Darren Edge, and Junehwa Song. 2016. Designing interactive multiswimmer exergames: a case study. *ACM Transactions on Sensor Networks (TOSN)* 12, 3 (2016), 1–40.

[9] Woohyeok Choi, Jeungmin Oh, Taiwoo Park, Seongjun Kang, Miri Moon, Uichin Lee, Inseok Hwang, and Junehwa Song. 2014. MobyDick: an interactive multi-swimmer exergame. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*. 76–90.

[10] William Grussenmeyer and Eelke Folmer. 2017. Accessible touchscreen technology for people with visual impairments: a survey. *ACM Transactions on Accessible Computing (TACCESS)* 9, 2 (2017), 1–31.

[11] João Guerreiro and Daniel Gonçalves. 2014. Text-to-speeches: evaluating the perception of concurrent speech by blind people. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. 169–176.

[12] João Guerreiro and Daniel Gonçalves. 2016. Scanning for digital content: How blind and sighted people perceive concurrent speech. *ACM Transactions on Accessible Computing (TACCESS)* 8, 1 (2016), 1–28.

[13] Inseok Hwang, Qi Han, and Archan Misra. 2005. MASTAQ: a middleware architecture for sensor applications with statistical quality constraints. In *Third IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 390–395.

[14] Inseok Hwang, Youngki Lee, Chungkuk Yoo, Chulhong Min, Dongsun Yim, and John Kim. 2019. Towards interpersonal assistants: next-generation conversational agents. *IEEE Pervasive Computing* 18, 2 (2019), 21–31.

[15] Inseok Hwang, Chungkuk Yoo, Chanyou Hwang, Dongsun Yim, Youngki Lee, Chulhong Min, John Kim, and Junehwa Song. 2014. TalkBetter: family-driven mobile intervention care for children with language delay. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1283–1296.

[16] Bumsoo Kang, Chulhong Min, Wonjung Kim, Inseok Hwang, Chunjong Park, Seungchul Lee, Sung-Ju Lee, and Junehwa Song. 2017. Zaturi: We put together the 25th hour for you. create a book for your baby. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1850–1863.

[17] Haechan Lee, Miri Moon, Taiwoo Park, Inseok Hwang, Uichin Lee, and Junehwa Song. 2013. Dungeons & swimmers: designing an interactive exergame for swimming. In *Proceedings of the 2013 ACM conference on Pervasive and Ubiquitous Computing adjunct publication*. 287–290.

[18] Youngki Lee, Chulhong Min, Chanyou Hwang, Jaeung Lee, Inseok Hwang, Younghyun Ju, Chungkuk Yoo, Miri Moon, Uichin Lee, and Junehwa Song. 2013. Sociophone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. 375–388.

[19] Meera Radhakrishnan, Darshana Rathnayake, Ong Koon Han, Inseok Hwang, and Archan Misra. 2020. ERICA: enabling real-time mistake detection & corrective feedback for free-weights exercises. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 558–571.

[20] Rishi Vanukuru. 2020. Accessible Spatial Audio Interfaces: A Pilot Study into Screen Readers with Concurrent Speech. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–6.

[21] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W White. 2019. Verse: Bridging screen readers and voice assistants for enhanced eyes-free web search. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 414–426.

[22] Chungkuk Yoo, Inseok Hwang, Seungwoo Kang, Myung-Chul Kim, Seonghoon Kim, Daeyoung Won, Yu Gu, and Junehwa Song. 2017. Card-stunt as a service: Empowering a massively packed crowd for instant collective expressiveness. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 121–135.

[23] Benjamin Rich Zendel and Claude Alain. 2009. Concurrent sound segregation is enhanced in musicians. *Journal of Cognitive Neuroscience* 21, 8 (2009), 1488–1498.