Concurrent Speech and Auditory Tag Clouds for Non-Visual Web Interaction

Dhia-Eddine Merzougui[1], *Nilesh Tete*[1,2], *Fabrice Maurel*[1], *Gaël Dias*[1], *Mohammed Hasanuzzaman*[2,3], *Aurélien Bournonville*[1], *Edgar Madelaine*[1], *Thomas Berthelin Le Tellier*[1], *François Ledoyen*[1], *Laure Poutrain-Lejeune*[1], *François Rioult*[1], *Jérémie Pantin*[1]

[1]Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR6072, France
[2]Computer Science Department, Munster Technological University, ADAPT Centre, Ireland
[3]School of EEECS, Queen's University Belfast, Northern Ireland

dhia-eddine.merzougui@unicaen.fr, m.hasanuzzaman@qub.ac.uk

Abstract

Millions of visually impaired individuals face significant barriers to independent information access, limiting their autonomy and self-determination. Despite advancements in assistive technologies, challenges remain, particularly in handling web document structure and multi-channel information transposition. This paper addresses these issues by introducing the TagThunder experimental framework, which applies the "cocktail party effect" metaphor to transpose the morpho-dispositional semantics of web documents into an auditory tag cloud, enabling rapid non-visual skimming. Additionally, it explores interactive stimuli for structured information scanning through discrete and continuous guiding exploration strategies.

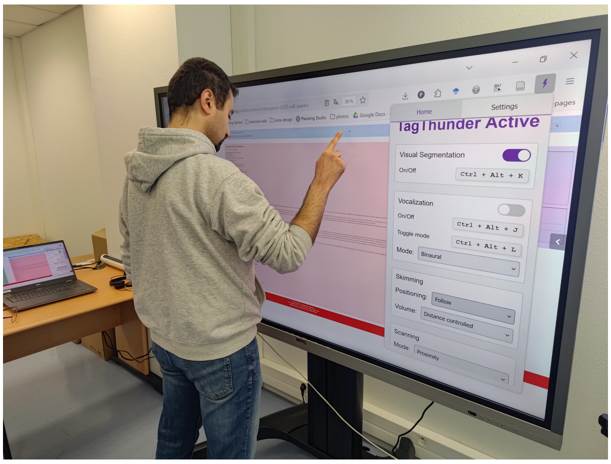
Index Terms: Visually impaired people, Morpho-dispositional semantics, Concurrent speech, Interaction strategies.

1. Introduction

Ensuring equitable access to information is essential for fostering autonomy and self-determination among visually impaired people (VIP). According to the International Agency for the Prevention of Blindness, over 43 million people worldwide are blind, with an additional 295 million experiencing moderate-to-severe visual impairments[1]. Despite digital accessibility advances, barriers remain in non-visual interaction, information retrieval, and multimodal content representation.

While visual web document access allows users to quickly grasp key information—such as topic, page type, and the distinction between central and peripheral elements—non-visual frameworks still fail to fully leverage the morpho-dispositional semantics of complex web documents. Most existing non-visual web search solutions minimally alter the visual structure and rely on synthesized speech as users hover over elements. However, this approach can be tedious, requiring users to interpret fragmented speech snippets and mentally reconstruct the physical and logical-thematic organization of a web page.

To bridge the digital divide, it is essential to support non-visual access that is more intuitive and interactive, enabling VIP to navigate and utilize the informational and structural information of web document in a more efficient way. To achieve this, we introduce the TagThunder experimental framework[2], which (1) transposes the morpho-dispositional semantics of web documents into an auditory tag cloud for rapid access and skimming, and (2) implements interactive strategies for non-visual navigation and information search, enhancing scanning capabilities. TagThunder utilizes web page processing algorithms as web services and a browser add-on that overlays an interactive layer, enabling seamless sensory exploration of web content through concurrent auditory feedback as illustrated in Figure 1.



Visual Segmentation
On/Off Ctrl + Alt + K

Vocalization
On/Off Ctrl + Alt + J
Toggle mode Ctrl + Alt + L
Mode: Binaural

Skimming
Positioning: Follow
Volume: Distance controlled

Scanning
Mode: Proximity

Figure 1: *The TagThunder experimental framework.*

2. TagThunder Architecture

Cocktail Party Effect Metaphor: The TagThunder experimental framework is built on an analogy extending the psychological concept of the "cocktail party effect". It denotes the ability to focus one's auditory attention on a verbal stream amidst the noisy environment of a gathering. We develop this metaphor by considering the relationship between a VIP and the different zones of a web page, akin to the relationship between a guest situated in a room with various discussion groups. As such, each zone is seen as a potential discussion group, where the topic being vocalized is the set of relevant key terms from the zone.

Processing Chain: TagThunder is designed as a pipeline, as shown in Figure 2. First, the web page undergoes preprocessing to generate an augmented HTML document containing visual, logical, and content information, which is then cleaned to remove non-visual elements. Second, the TDBU algorithm [1] segments the web page to identify meaningful zones. Third, YAKE! [2] extracts key terms from each zone based on lexical significance and local formatting. Finally, key terms are vocalized concurrently through the Kokoro text-to-speech model[3] and 3D spatialized with the Web Audio API[4] according to the zone typographical layout as proposed in [3].

[1] https://www.iapb.org/learn/vision-atlas/magnitude-and-projections/global/

[2] TagThunder is an experimental framework for research protocols and lacks full accessibility features necessary for VIP end users.

[3] https://huggingface.co/hexgrad/Kokoro-82M

[4] https://developer.mozilla.org/fr/docs/Web/API/Web_Audio_API

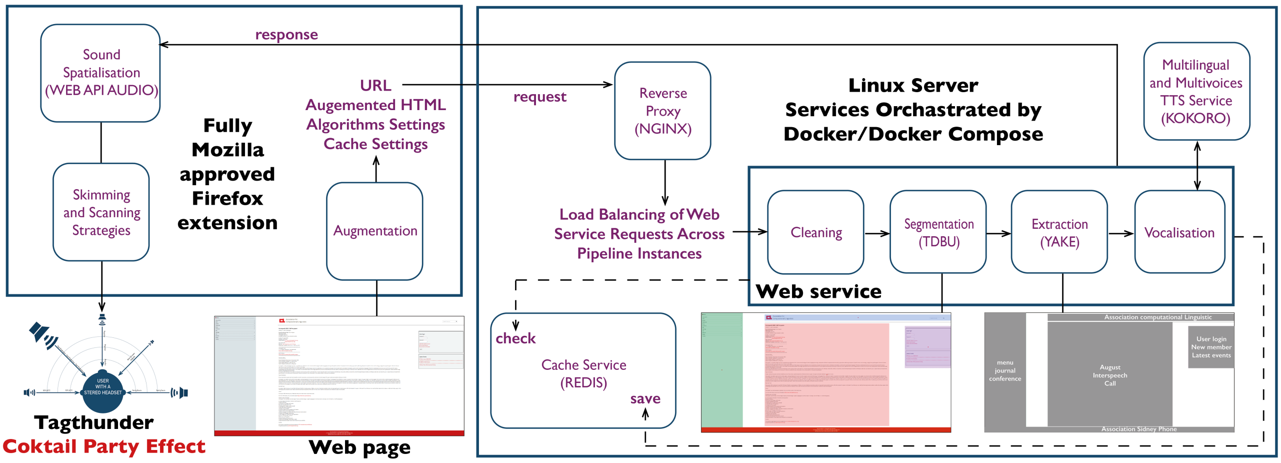


Figure 2: *The TagThunder architecture.*

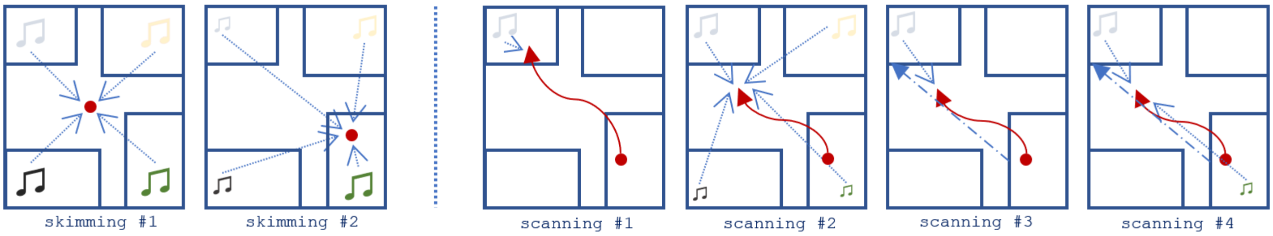


Figure 3: *Skimming and scanning strategies.*

3. Skimming and Scanning Strategies

Numerous studies in psychology and human-computer interaction have long explored sensory transposition. But few studies consider document visual structure as conveying morpho-dispositional semantics. The rare exceptions focus either on tactile [4] or oral [5, 6] transposition. The TagThunder experimental framework advances oral transposition by leveraging layout as an interactive tool, enabling VIP to employ both skimming and scanning reading strategies.

Skimming: Two skimming strategies are implemented, allowing non-visual users to access information at a glance. In skimming #1, the user's reference position is set at the center of the web page, and all key terms from different zones are vocalized concurrently without volume distinction. Skimming #2 lets users self-determine their reference point, with key terms vocalized concurrently and volume adjusted by distance to the reference point. Both strategies are illustrated in Figure 3.

Scanning: Four scanning strategies are implemented ranging from discrete to continuous navigation stimuli. Scanning strategy #1 allows the non-visual user to browse a web page while receiving the sequential vocalization of the key terms of the hovered zone as a unique stimulus (discrete strategy). With respect to continuous scanning, strategy #2 provides the user with all key terms from all zones constrained in volume by their distance to the successive reference points of the reading pathway. Scanning #3 simultaneously vocalizes the key terms of the zones that intersect the pathway direction—computed within a sequence of 5 distinct movements—constrained in volume by their distance to the current reference point. Finally, scanning strategy #4 concurrently vocalizes a fixed set of key terms composed of ones from the source and the target zones with respect to the reading pathway. The number of key terms from each zone is proportionally selected based the zone distance to the current reference point. The selection process relies on the most (dis)similar key terms between zones computed by the cosine similarity of phrase multilingual embeddings (e.g. XLM-RoBERTa), thus implementing an allocentric strategy. The Mozilla-approved Firefox add-on is available at https://dias.users.greyc.fr/tagthunder.zip for evaluation.

4. References

[1] W. Safi, "Non-visual vibro-tactile navigation of web pages on touch screen devices," PhD Thesis, University of Caen Normandie (France), 2019.

[2] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, pp. 257–289, 2020.

[3] F. Maurel, G. Dias, S. Ferrari, J.-J. Andrew, and E. Giguet, "Concurrent Speech Synthesis to Improve Document First Glance for the Blind," in *2nd International Workshop on Human-Document Interaction associated to ICDAR*, 2019.

[4] O. Moured, S. Alzalabny, T. Schwarz, B. Rapp, and R. Stiefelhagen, "Accessible document layout: An interface for 2d tactile displays," in *16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*, 2023, p. 265–271.

[5] J.-M. Lecarpentier, E. Manishina, F. M. S. Ferrari, E. Giguet, G. Dias, and M. Busson, "Tag thunder: Web page skimming in non visual environment using concurrent speech," in *7th Workshop on Speech and Language Processing for Assistive Technologies associated to INTERSPEECH*, 2016, pp. 1–8.

[6] J. Guerreiro, "Towards screen readers with concurrent speech: where to go next?" *ACM SIGACCESS Accessibility and Computing*, pp. 12–19, 2016.