# Tag Thunder: Web Page Skimming in Non Visual Environment Using Concurrent Speech

*Jean-Marc Lecarpentier, Elena Manishina, Fabrice Maurel*
*Stéphane Ferrari, Emmanuel Giguet, Gael Dias, Maxence Busson*

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
{firtsname.lastname}@unicaen.fr,

## Abstract

Skimming and scanning are two strategies generally used for *speed reading*. Skimming allows a reader to get a first glance of a document; scanning is the process of searching for a specific piece of information in a document. While both techniques are available in visual reading mode, it is rather difficult to use them in non visual environments. In this paper, we introduce the concept of *tag thunder*, which provides speed reading non-visual techniques similar to skimming and scanning. A tag thunder is the oral transposition of the tag cloud concept. Tag cloud key terms are presented using typographic effects which reflect their relevance and number of occurrences. Within a tag thunder, the relevance of a given key term is translated into specific speech effects and its position on the page is reflected in the position of the corresponding sound on a 2D stereo space. All key terms of a tag thunder are output according to a concurrent speech strategy, which exploits the *cocktail party effect*.

In this paper, we present our implementation of the tag thunder concept. The results of the evaluation campaign show that tag thunders present a viable non-visual alternative to visual speed reading strategies.

**Index Terms**: non-visual web navigation, human-computer interaction, text-to-speech synthesis, key term extraction

## 1. Introduction

Most users share a similar mental process when accessing informative content of web pages. They get a first glance of the page content (skimming), followed by a quick search for specific information (scanning). Then the reader spots different areas of interest and seeks for specific information in identified areas using a zoom-in zoom-out strategy.

Although several factors may influence whether skimming and scanning are successful, such document properties as layout, logical structure and typographic effects play an important role in the perception process. However, this information is usually not available to users in non-visual environments [1]. Figure 1 illustrates how a web page is perceived in visual and non-visual environments using a screen reader.


Figure 1: Perception of the same web page in visual and non-visual environments.

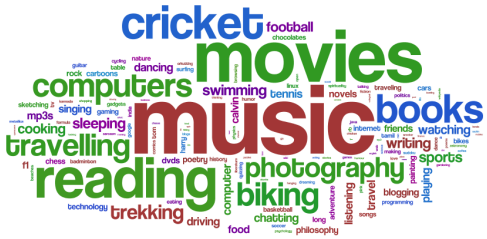


Figure 2: Example of a tag cloud.

To solve this problem, a number of non-visual replacement strategies [2, 3] have been proposed by screen readers such as faster speech rate depending on the textual block size, shortcuts which allow to jump from heading to heading, reading the beginning and the end of a paragraph, etc. However, these solutions are far from providing the reading capabilities of the visual mode [4].

This paper focuses on developing a strategy for a fast access to web page content in non-visual situations, that takes into account page layout and typographic clues. In particular, we transpose the visual concept of *tag cloud* to its audio version, called *tag thunder*.

In order to apply this concept to skimming and scanning, let us first consider a web page as a set of blocks. Figure 3 illustrates the result of a page segmentation. Each



*SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies*, pages 1–8,
San Francisco, USA, 13 September, 2016.

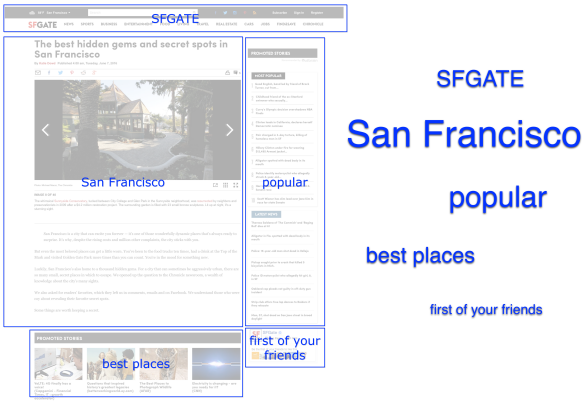


Figure 3: Segmentation of a page into zones and extracting key terms

zone is represented by key terms extracted from this zone which are combined into a tag cloud along with spatial and typographic effects that reflect the importance and relevance of each specific term, as shown in Figure 2[1]. Similarly, a tag thunder adds spatial and audio effects to key terms.

Tag thunders use concurrent speech strategy in order to represent the dense visual stimulus embodied by tag clouds. This strategy is based on the *Cocktail Party Effect*: users may identify key terms pronounced simultaneously or focus their attention on key terms that interest them among all the others.

This paper is structured as follows. In Section 2 we provide an outline of the related work in the area of non-visual content access strategies. In Section 3 we introduce our implementation of the tag thunder concept, specifically the three main steps in tag thunder creation: web page segmentation, key word extraction and vocal synthesis. Sections 4 presents the evaluation campaign we organized in order to assess the performance of our tag thunder implementation, as well as the potential of the tag thunder concept in general. We conclude this paper with a discussion and some directions for future work.

## 2. Related work

This Section presents some strategies developed in the field of assistive technologies, which facilitate the access to web content in non-visual environments.

Existing solutions for non-visual web page browsing often use Text-To-Speech (TTS) and Braille mode. Text-To-Speech has been used to convey document structure to users in non-visual situations via the content vocalization [5, 6]. To increase TTS efficiency, [7] proposed an oral transposition model based on layout reformulation strategies. These strategies combine a model of the written document [8], used to develop discursive forms from structured texts, and a prosodic model [9] used to reduce this new set of sentences in a more speech-adapted way. This approach brings a significant improvement in memorizing and understanding TTS output for strongly structured documents. But according to [10], the cognitive load is still hard to handle in comparison to visual reading.

Some early studies proposed to use summarization techniques to provide visually impaired people (VIP) with web pages skimming strategies [11]. However, a linearization step destroys the page layout which is at the core of the perception/action loop.

In the Accessibility through Simplification & Summarization project [12] (AcceSS), the content perceived as less important is removed from pages, thus modifying the page layout. A navigation page is then built, named *guide dog page*, which serves as a summary. Experiments show positive results when this method is combined with a JAWS screen reader. One of the limitations of this method is the incapacity of the pattern matching algorithm to correctly identify page sections. Furthermore, no simplification is proposed at textual level, providing no solutions to quickly browse large textual content.

SeEbrowser (Semantically Enhanced Browser) is a VIP-adapted audio web browser [13]. Manual semantic annotations are used to build ontologies modeling hierarchical relationships between elements within web pages. As the web page is loaded, the user may ask for Browser Shortcuts (BSs), go through them and interact using keyboard and audio feedback. Experiments show that this alleviates the information overload. However, the scanning strategy is still very long since users tend to listen to all BSs before choosing a relevant one.

Hearsay [14] is a non-visual multi-modal web browser which has been developed at Stony Brook University (New York, USA) since 2004. It supports different input modes: voice, keyboard and tactile interfaces. Possible output modalities are audio, screen and Braille. The browser provides many features: a segmentation module which analyzes web page structure and layout, a system of annotations which enables the addition of alternative text for pictures and other content blocks, algorithms detecting the changes between visited web pages, a context analyzer which detects the main content and identifies relevant information using hyperlinks. Experiments show a significant gain of time in finding the main content of a web page. In addition the system avoids repeating static content such as menus. Globally, most of these features made valuable contributions to improving user experience. Nevertheless, despite the positive results, two aspects still require improvement in comparison to visual reading. First, the page structure overview is not complete because it focuses on main content; the el-

[1]Image by Anand S, https://flic.kr/p/5BFE3V, CC-BY-2.0

ements are presented sequentially, making their browsing long. Thus, this method does not provide real skimming and scanning reading modes.

More and more work is now done using tactile strategies [15, 16, 17]. [18] incorporate patterns into web pages, thus enabling some elements and their relationships to be felt by running fingers over them. Such transformed documents are then given to VIPs using special paper with heat-sensitive ink. Putting the paper on a touch screen makes it possible to interact with it and obtain the oral transposition of a chosen web page section. Limitations come from the need to use a special paper with a heater. A similar concept is based on vibrotactile perception [19]. A special device captures contrast variations on the screen as fingers *browse* the content on a tablet. These variations are transformed into vibrations felt in a glove device worn on the other hand.

In recent years, some work has been carried out using Text-To-Speech tools within concurrent speech paradigm, exploiting the fact that human ears may concentrate on a specific audio source among many others [20]. The *Cocktail Party Effect* is a perfect example: even when many people are speaking simultaneously, we may concentrate our attention on one specific voice [21]. Variations in spatial location [22], as well as speech parameters (synchrony [23], frequencies [24]) may influence the perception of different voices. Using concurrent speech proved to accelerate blind people's scanning for relevant information [25, 26].

To resume this section, two main approaches (content summarization and concurrent speech synthesis) represent two interesting scanning strategies. However, they are not sufficient in providing real skimming abilities. The tag thunder concept combines both strategies: using segmentation and extraction techniques to give a summary of the page content and using concurrent speech synthesis to provide a quick overview.

## 3. Architecture

This Section presents our implementation of the tag thunder concept. It comprises three modules: web page segmentation, key term extraction and key term vocalization using concurrent speech synthesis.

### 3.1. Page segmentation

There exist numerous approaches to webpage segmentation [27, 28, 29, 30, 31]. We opted for the K-means++ algorithm [32, 33] [2]. The choice for unsupervised clustering algorithm was dictated, among other things, by the lack of unified web page layout, and robustness of K-Means algorithm in similar tasks [34]. It groups visible HTML elements into a desired number of zones based on their Euclidian distance. To optimize convergence and efficiency, each HTML element is enhanced with its computed styles based on underlying CSS and Javascript code [35]. Elements that are not part of the visual layout are ignored.

For the purpose of our experiment, the enhanced HTML is clustered into 5 zones. This choice of the number of zones was made with the objective to avoid a working memory overload, in accordance with the Miller's Law [36].

[2] http://scikit-learn.org/stable/modules/clustering.html#k-means

### 3.2. Key terms extraction and weighting

Each zone is represented by its key terms in the tag thunder. In our current implementation, key terms are n-grams of different lengths with a maximum order of 6.

For each n-gram, we compute $tf - idf$ [37] (term frequency – inverse document frequency). $Tf$ is the frequency of a given term in a zone. The $idf$ is computed using a corpus $C$ containing 953 551 articles of the "Le Monde" newspaper dating from 1987 to 2006. Similar to [38], our solution couples $tf - idf$ metric with additional parameters. We use Formula (1) to compute the final score for each key term.

$$Score = tf(term, zone) \cdot idf(term, C) \cdot \sum_{i=1}^{n} \sigma(c_i) \quad (1)$$

where $tf(term, zone)$ is the frequency of the term within its zone, $idf(term, C)$ is the number of documents in our corpus $C$ containing the term. $\sigma(c_i)$ is the weight for a characteristic $c_i$ such as font weight, size, variant, style, etc. $\sigma$ values were assigned empirically and reflect the visual perception of a given element. $\sigma(c_i)$ values range from $0.5$ to $5$.

For the purpose of our experiment, each zone is represented by one key term only.

### 3.3. Concurrent speech vocalization

This module generates the audio signal from a given key term and its zone properties. Based on [21, 23, 24], specific voice, volume, prosody, pitch, speech rate and synchronization characteristics are combined to build an audio track for a given key term.

Our synthesis module uses the Kali TTS [39] tool, developed at the University of Caen Normandie by the CRISCO laboratory. Kali supports speech rate acceleration without loss in intelligibility and sound quality, which is a very important feature in non-visual web browsing.

To vocalize the key terms, we use several cocktail party effect metaphors. Thus, we consider each zone as a discussion group in a cocktail party. Each metaphor provides rules which assign repetition frequency (Figure 4), volume (Figure 5) and a spot in a 2D audio space (Figure 6) to each key term. Vocalization of all the key terms

with their specific parameters produces the final tag thunder.

### 3.3.1. Repetition frequency

**Metaphor 1**: the larger the group talking about a topic, the more often related terms emerge.
**Rule 1**: vocalized key terms are played in a loop. Zone size influences repetition frequency within the loop.

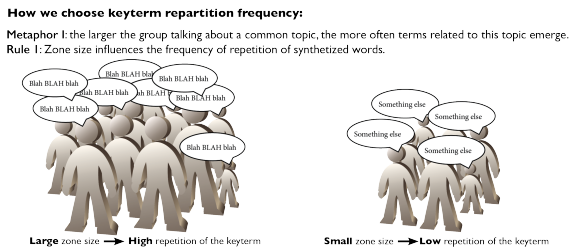


Figure 4: Repetition frequency metaphor

For each key term, the silence between two repetitions in the loop is proportional to the relative size of its zone. The larger the zone, the shorter the silence. In our experiment, silence duration has been empirically set between $0.5$ second and $5$ seconds.

### 3.3.2. Volume

**Metaphor 2.a (distinctiveness)**: the more a voice in a group stands out, the easier it is to detect its source.
**Metaphor 2.b (relevance)**: the more the words are repeated in a group, the more relevant they are.
**Rule 2**: volume is determined by zone contrast and key terms frequency in the zone.

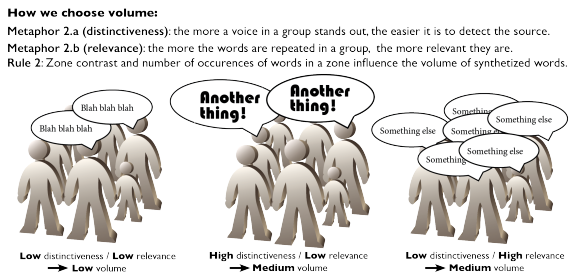


Figure 5: Volume metaphor

For each zone, contrast is computed based on the difference between the background color and the text. Volume is set within a $[min, max]$ interval, using the average of normalized contrast value and key term frequency. In our experiment, TTS constraints and perceptive tests led to setting the values for $min$ and $max$ to $4$ and $8$ points respectively, with each point representing 2 amplitude tones.

### 3.3.3. Spatialization

**Metaphor 3**: sound spatialization helps to physically place and distinguish several discussion groups.
**Rule 3**: zone coordinates influence the type of output voice and 2D spatialization of vocalized key terms.

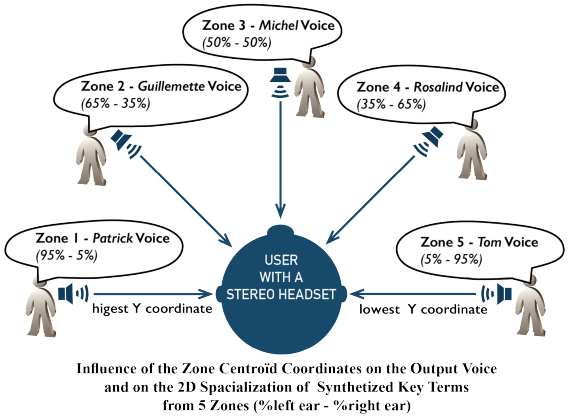


Figure 6: Sound spatialization metaphor

Voices are equally distributed in the 2D stereo space depending on the zone's centroid coordinates. In our experiment, sounds originate from $5$ sources (i.e. $5$ corresponding zones), as illustrated in Figure 6 .

## 4. Evaluation

We conducted an experiment in order to test the viability of the tag thunder concept and the quality of our implementation. In this Section, we present the experimental setting and the results.

### 4.1. Experimental setting

Our goal is to evaluate the system's capacity to provide fast skimming reading strategies. Here we present the results of the first experiment with sighted participants. The goal of this experiment is two-fold: to evaluate the relevance of the extracted key terms and to test the efficiency of tag thunder concept as a skimming strategy.

The experiment unfolds as follows. A participant sees a tag cloud followed by a web page, 15 seconds each. The page may or may not be the corresponding web page. The participant is asked whether the tag cloud corresponds to the displayed page.Possible answers are: definitely yes, probably yes, probably no, definitely no. Another participant is presented with the same data, but in the form of a tag thunder instead of the tag cloud and is asked to answer the same question. The experiment modalities were as follows:


(a) Tag Cloud and Tag Thunder output


(b) Webpage with a question form

Figure 7: User evaluation: web-based interface

- 18 sighted participants, each with 16 different stimuli (8 tag clouds - 8 tag thunders);
- 24 web pages from various web sites were used to generate a tag cloud and a tag thunder for each page;
- 24 other web pages were selected to create stimuli where the page and tags do not match;

Each couple (web page, tag set) was shown to 3 different participants; each participant evaluated an equal number of correct (matching) and incorrect couples.

Participants took the test autonomously, with a supervisor close by. The evaluation interface is shown in Figure 7.

## 4.2. Results

We present the evaluation results for Tag Clouds (TC) and Tag Thunders (TT) separately, as well as the combined overall results. We also separate the analysis of the correct (matching) and incorrect pages.

Figure 8 shows the dispersion of the total of 288 answers. It seems more difficult for participants to definitely validate a correct page than to definitely reject an incorrect page.

### *4.2.1. Agreement*

We split the analysis of agreement statistics into three interpretations: *4-var* with four different answers; *3-var* where 'probably yes' and 'probably no' are combined into 'not sure'; and *2-var* where the answers 'definitely yes' and 'probably yes' are combined into 'yes' and 'probably no' and 'definitely no' into 'no'.

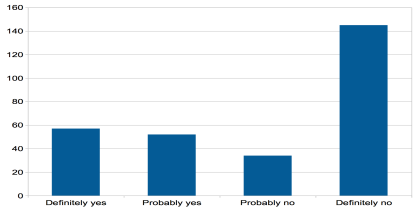


Figure 8: Dispersion of the 288 answers

| | Web page | 4-var | 3-var | 2-var |
|---|---|---|---|---|
| **TC** | Correct | 12.8 | 20.8 | 70.8 |
| | Incorrect | 75.0 | 75.0 | 91.7 |
| **TT** | Correct | 20.8 | 33.3 | 75.0 |
| | Incorrect | 66.6 | 66.6 | 95.8 |
| **all** | | 43.8 | 48.96 | 83.3 |

Table 1: Percentage of stimuli with the same answer

Table 1 presents the agreement statistics. The *2-var* interpretation shows a very high agreement rate when the incorrect page was shown, for both TT and TC. The *3-var* interpretation shows differences only for the correct pages, thus indicating that hesitations concerned correct pages only. This might mean that key terms were not always well suited to represent their zones in case of correct web pages, which created hesitation between 'probably yes' and 'yes' TTs tend to have a better agreement than TCs. Our hypothesis is that, in our experiment, textual key terms in a TC were displayed with fewer typographic effects whereas key terms in TTs had a full set of audio (or 'typophonic') effects described above. In general, the modality of the stimuli (TT vs TC) does not seem to influence the agreement rate between users.

### *4.2.2. Precision and Recall*

Precision and recall are computed on the *2-var* interpretation. Table 2 presents the results. There is a significant difference in the perception of TCs or TTs between cases where the page was the correct or incorrect one. For the correct pages, the precision is very high, which means that the participants manage to associate a given page to a TT/TC. On the contrary, the recall is somewhat lower: as discussed before, users find it difficult to validate a correct page. This suggests that a number of correct pages were labeled as incorrect, which in turn might suggest the insufficiency of the TT/TC representation in these cases. This is especially apparent for tag thunders: 31% of correct pages were labeled as incorrect. Again, these results suggest that the extraction module needs further improvement.

Overall, participants found the exercise difficult but made few mistakes. In general, the results of TTs are

| Format | TagCloud | | TagThunder | |
|---|---|---|---|---|
| **Web page** | **Correct** | **Incor.** | **Correct** | **Incor.** |
| **Precision** | **0.96** | 0.81 | **0.98** | 0.76 |
| **Recall** | 0.78 | 0.97 | 0.69 | 0.98 |
| **F-score** | 0.86 | 0.89 | 0.81 | 0.86 |
| **Accuracy** | 0.875 | | 0.84 | |

Table 2: 2-var results: precision, recall and F-score

comparable in the overall accuracy with the results of the TCs. We can conclude that the tag thunder concept is valid and that certain limitations originate from the internal implementation of each module. We discuss these limitations in the following Section.

## 5. Discussion and future work

The first objective of this work was to implement the concept of tag thunders. Evaluation results demonstrate the viability of this concept. However, each module requires separate thorough evaluation.

### 5.1. Page segmentation

According to evaluation results, most errors come from pages where the number of distinct informational sections is larger than the default number of expected zones (5 in this experiment). In this case, the obtained zones contain multiple sections of content handling distinct subjects. Selected key terms therefore do not fully represent that zone as a whole, rather one of the zone sections.

In the future work, we consider two potential improvements of our segmentation module. The first one mixes DOM based and image based approaches to page segmentation. The second one uses the Gestalt theory [40] to simulate the similarity, proximity and complexity principles.

### 5.2. Key term extraction

One of the main issues with key terms extraction was the maximum size of the n-gram order, which we fixed to 6. As a result we do not always obtain coherent phrases: abrupt endings, missing beginnings, etc. At the same time augmenting n-gram order would lead to longer key terms which might affect the user's ability to comprehend and retain the information contained in these n-grams.

As already mentioned, another issue was the complex multi-section structure of certain zones which does not allow to extract one key term that would represent the zone. One possible solution is to extract several short key terms, one per zone section, and join them into one compound key term. Some zones, like menus and footers, usually contain list items, making it difficult to extract one key term per zone. A solution is, again, to produce a key term which would either contain several elements (several menu items) or a meta key term, for example 'navigation menu', which would summarize the content.

Finally, several issues are related to the corpus used to compute $idf$. In this implementation, it was composed of news articles, produced between 1986 and 2006. One way to extend the coverage of the corpus is to acquire new vocabulary dynamically.

### 5.3. Vocalization

The evaluation results indicate that our audio representations in a form of tag thunders were comparable to their visual counterparts in clarity and intelligibility (accuracy values of 0.875 vs. 0.84). However, some users indicated a somewhat artificial sound of the generated tag thunders. More experiments with different sound settings and spatialization modes are in process. Binaural recording techniques may be used to render spatial variations in tag thunders with simple stereo headsets. Since the Kali TTS is not compatible with markup languages such as VoiceXML and SSML, our solution needs to integrate a compatible TTS so that we can use industry standards.

More experiments using different prosodic strategies will need to be made in order to determine which combination of sound effects give a user the best representation of the typography and page layout.

## 6. Conclusion

In this article, we proposed a strategy to facilitate skimming of web pages in non-visual environments. Our solution, which we call tag thunder, involves several processing steps: segmentation of a web page into zones, extraction of key terms from each zone and finally, vocalization of the key terms in a tag thunder. Evaluation results show that participants were able to measure the correspondence between a tag thunder and a web page.

The next step is to find the best compromise between the number of zones and key terms and the perceptive capacity of users. We intend to evaluate our concept with VIPs and use their feedback to direct our future work.

Our final objective is to integrate human computer interaction into our system, specifically for in-page navigation: once a zone is selected, we want to be able to 'navigate' to and explore that zone. In that case, headsets with sensors may enable interactions with movements of the head. Combining our approach with vibro-tactile devices would lead to multi-modal systems which facilitate access to web content in non-visual situations.

## 7. Acknowledgments

This research work was funded by the 'Region Normandie' with the CPER NUMNIE project.

## 8. Website

Tag thunder generator: `https://tagthunder.greyc.fr/demo/`
Experiment (French version): `https://tagthunder.greyc.fr/demotest`

## 9. References


[1] G. Dias and B. Conde, "Accessing the web on handheld devices for visually impaired people," in *Advances in Intelligent Web Mastering*, ser. Advances in Soft Computing, K. Wegrzyn-Wolska and P. Szczepaniak, Eds., 2007, vol. 43, pp. 80–86.

[2] Y. Borodin, J. P. Bigham, G. Dausch, and I. Ramakrishnan, "More than meets the eye: A survey of screen-reader browsing strategies," in *International Cross Disciplinary Conference on Web Accessibility (W4A)*, 2010, pp. 1–10.

[3] F. Ahmed, Y. Borodin, A. Soviak, M. Islam, I. Ramakrishnan, and T. Hedgpeth, "Accessible skimming: Faster screen reading of web pages," in *25th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2012, pp. 367–378.

[4] J. P. Bigham, A. C. Cavender, J. T. Brudvik, J. O. Wobbrock, and R. E. Lander, "Webinsitu: A comparative analysis of blind and sighted browsing behavior," in *9th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, 2007, pp. 51–58.

[5] S. Goose and C. Möller, "A 3d audio only interactive web browser: using spatialization to convey hypermedia document structure," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. ACM, 1999, pp. 363–371.

[6] L. Sorin, J. Lemarié, N. Aussenac-Gilles, M. Mojahid, and B. Oriola, "Communicating text structure to blind people with text-to-speech," in *Computers Helping People with Special Needs*. Springer, 2014, pp. 61–68.

[7] F. Maurel, N. Vigouroux, M. Raynal, and B. Oriola, "Contribution of the transmodality concept to improve web accessibility," *Assistive Technology Research Series*, vol. 12, pp. 186–193, 2003.

[8] J. Virbel, "The contribution of linguistic knowledge to the interpretation of text structures," in *Structured documents*. Cambridge University Press, 1989, pp. 161–180.

[9] M. Rossi, *L'intonation: le système du français: description et modélisation*. Editions Ophrys, 1999.

[10] F. Maurel, M. Mojahid, N. Vigouroux, and J. Virbel, "Documents numériques et transmodalité," *Document numérique*, vol. 9, no. 1, pp. 25–42, 2006.

[11] F. Ahmed, Y. Borodin, Y. Puzis, and I. Ramakrishnan, "Why read if you can skim: towards enabling faster screen reading," in *International Cross-Disciplinary Conference on Web Accessibility - W4A2012, Article No. 39*, 2012.

[12] B. Parmanto, R. Ferrydiansyah, A. Saptono, L. Song, I. W. Sugiantara, and S. Hackett, "Access: accessibility through simplification & summarization," in *Proceedings of the 2005 international cross-disciplinary workshop on web accessibility (W4A)*. ACM, 2005, pp. 18–25.

[13] S. Michail and K. Christos, "Adaptive browsing shortcuts: Personalising the user interface of a specialised voice web browser for blind people," in *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*. IEEE, 2007, pp. 818–825.

[14] Y. Borodin, F. Ahmed, M. A. Islam, Y. Puzis, V. Melnyk, S. Feng, I. Ramakrishnan, and G. Dausch, "Hearsay: a new generation context-driven multi-modal assistive web browser," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 1233–1236.

[15] M. Ziat, O. Gapenne, J. Stewart, and C. Lenay, "Haptic recognition of shapes at different scales: A comparison of two methods of interaction," *Interacting with Computers*, vol. 19, no. 1, pp. 121–132, 2007.

[16] N. A. Giudice, H. P. Palani, E. Brenner, and K. M. Kramer, "Learning non-visual graphical information using a touch-based vibro-audio interface," in *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 2012, pp. 103–110.

[17] A. A. Ahmed, M. A. Yasin, and S. F. Babiker, "Tactile web navigator device for blind and visually impaired people," in *Applied Electrical Engineering and Computing Technologies (AEECT), 2011 IEEE Jordan Conference on*. IEEE, 2011, pp. 1–5.

[18] Y. B. Issa, M. Mojahid, B. Oriola, and N. Vigouroux, "Analysis and evaluation of the accessibility to visual information in web pages," in *Computers Helping People with Special Needs*. Springer, 2010, pp. 437–443.

[19] W. Safi, F. Maurel, J.-M. Routoure, P. Beust, and G. Dias, "Blind browsing on hand-held devices: Touching the web... to understand it better," in *Data Visualization Workshop (DataWiz 2014) associated to 25th ACM Conference on Hypertext and Social Media (HYPERTEXT 2014)*, 2014.

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

[21] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[22] D. S. Brungart and B. D. Simpson, "Optimizing the spatial configuration of a seven-talker speech display," *ACM Transactions on Applied Perception (TAP)*, vol. 2, no. 4, pp. 430–436, 2005.

[23] M. Turgeon, A. S. Bregman, and B. Roberts, "Rhythmic masking release: effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping." *Journal of Experimental Psychology: Human Perception and Performance, 31(5)*, p. 939, 2005.

[24] C. J. Darwin, D. S. Brungart, and B. D. Simpson, "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers." *The Journal of the Acoustical Society of America, 114*, p. 2913, 2003.

[25] J. Guerreiro and D. Gonçalves, "Text-to-speeches: evaluating the perception of concurrent speech by blind people," in *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. ACM, 2014, pp. 169–176.

[26] ——, "Faster text-to-speeches: Enhancing blind people's information scanning with faster concurrent speech," in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, 2015, pp. 3–11.

[27] A. Sanoja and S. Gançarski, "Block-o-matic: A web page segmentation framework," in *Multimedia Computing and Systems (ICMCS), 2014 International Conference on*. IEEE, 2014, pp. 595–600.

[28] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "Vips: A vision-based page segmentation algorithm," Microsoft technical report, MSR-TR-2003-79, Tech. Rep., 2003.

[29] J. Cao, B. Mao, and J. Luo, "A segmentation method for web page analysis using shrinking and dividing," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 25, no. 2, pp. 93–104, 2010.

[30] N. F. S. R. G. Adda, "Pré-segmentation de pages web et sélection de documents pertinents en questions-réponses," *TALN-RÉCITAL 2013*, p. 479, 2013.

[31] X. Liu, H. Lin, and Y. Tian, "Segmenting webpage with gomory-hu tree based clustering," *Journal of Software*, vol. 6, no. 12, pp. 2421–2425, 2011.

[32] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] S. Tatiraju and A. Mehta, "Image segmentation using k-means clustering, em and normalized cuts."

[35] E. Giguet and N. Lucas, "The book structure extraction competition with the resurgence software at caen university," in *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer, 2009, pp. 170–178.

[36] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review*, vol. 63, no. 2, p. 81, 1956.

[37] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[38] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical Automatic Keyphrase Extraction," *NRC/ERB-1057*, 1999.

[39] M. Morel and A. Lacheret-Dujour, "Kali, synthèse vocale à partir du texte : de la conception à la mise en oeuvre," *Traitement Automatique des Langues 42*, pp. 193–221, 2001.

[40] W. Köhler, *Gestalt Psychology.[British Ed.* G. Bells and sons, 1930.