Concurrent Speech Synthesis to Improve Document First Glance for the Blind

Fabrice Maurel, Gaël Dias, Stéphane Ferrari, Judith-Jeyafreeda Andrew, Emmanuel Giguet

Concurrent Speech Synthesis to Improve Document First Glance for the Blind

Fabrice Maurel, Gaël Dias, Stéphane Ferrari, Judith-Jeyafreeda Andrew, Emmanuel Giguet
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
14000 Caen
France
{*fabrice.maurel, gael.dias, stephane.ferrari, judith-jeyafreeda.andrew, emmanuel.giguet*}*@unicaen.fr*

***Abstract*—Skimming and scanning are two well-known reading processes, which are combined to access the document content as quickly and efficiently as possible. While both are available in visual reading mode, it is rather difficult to use them in non visual environments because they mainly rely on typographical and layout properties. In this article, we introduce the concept of tag thunder as a way (1) to achieve the oral transposition of the web 2.0 concept of tag cloud and (2) to produce an innovative interactive stimulus to observe the emergence of self-adapted strategies for non-visual skimming of written texts. We first present our general and theoretical approach to the problem of both fast, global and non-visual access to web browsing; then we detail the progress of development and evaluation of the various components that make up our software architecture. We start from the hypothesis that the semantics of the visual architecture of web pages can be transposed into new sensory modalities thanks to three main steps (web page segmentation, keywords extraction and sound spatialization). We note the difficulty of simultaneously (1) evaluating a modular system as a whole at the end of the processing chain and (2) identifying at the level of each software brick the exact origin of its limits; despite this issue, the results of the first evaluation campaign seem promising.**

***Keywords*-Web Accessibility; Document Layout; Oral Transposition; Non Visual Skimming;**

I. Introduction

Devices promoting non-visual access to digital information have seen strong development over the past decade. Voice synthesis and Braille are the main technologies used by blind people. However, they can be impractical, intrusive or ineffective on mobile media where tactile interaction is essential. The purpose of this article is to describe an original approach that focuses on the layout contrasts to build oral stimuli that promote non-visual navigation in web pages.

A. Document Layout and Non Visual Accessibility

Figure 1 illustrates the difficulty of considering a web page without the visual modality. It is not necessary to perceive the details of the figure to convince yourself that while the left page allows to collect a large amount of information in a few seconds (general subject, category of site type, central elements vs. peripheral ones ...), the right one does not offer such possibilities; this is due to the radical modification at the visual structure level. Yet, it is the same page depending on whether it is produced by a visual web browser or intended for the input of a conventional screen reader used by blind people on desktop computers.

Figure 1. Sight vs. Blind Web Page Perception

Accessibility software solutions embedded in tablets and smartphones (VoiceOver, Talkback) change the web page visual structure less strongly, but the text is synthesized as it is flown over by the finger. This solution is interesting but daunting when the goal is to browse new documents, especially in the context of a web navigation: the blind person must perceive, interpret and relate the snippets of speech synthesis produced by its interaction with the physical, logical and thematic organization of the web page. For this, he moves his finger on almost the entire screen to proceed to a heavy and somewhat random learning phase. As a result, most users rarely do this and have a very "practical" practice of touch devices, confining themselves to web sites and interfaces they know perfectly. This lack of freedom and serendipity of web page browsing hampers the development of original and user-adapted high-level reading strategies. This is a constraint that we wish to lift, in the name of a "right to stroll for all". To reduce the digital divide, it is imperative to allow a non-visual reading that is both global and naturally interactive; and thus to increase the perceptual capacities of blind people to access the informational and organizational structure of web pages.

B. Document Layout and High Level Reading Strategies

The visual properties of the text allow sight people to develop non-sequential reading strategies. It is widely accepted by linguists and psycho-linguists that these properties act on many dimensions. (1) They play positively or negatively on the legibility of documents and therefore on their cognitive accessibility; (2) Like prosody, they convey a part of the semantics of the message, showing the limits of the sentence as a textual unit: they inscribe the text spatially into textual objects whose linguistic scope may be below or beyond the sentence; (3) By the quality of affordance they provide to the document, they exploit our natural perceptive tendencies to suggest coherent interpretative paths; (4) They generate new possibilities specific to individual reading intentions. In this, they support the emergence of creative strategies through experience or serendipity; (5) They promote the ability of the eye to quickly combine both local and global information gathering operations. It is this interaction and this dynamic, which underlies all the others, that we wish to preserve during the transposition of visual properties into new sensory modalities. Two processes, skimming and scanning, involved in the development of rapid reading strategies, are discussed in the following subsection.

C. Document Layout, Skimming and Scanning

Skimming and scanning are two cognitive reading processes, which are used to access the document content as quickly and efficiently as possible. Scanning refers to the process of searching for a specific piece of information, and skimming is the action of passing through a document in a first glance to get an overview of its content. These two capabilities for processing written text, more or less aware, can be repeated in different combinations until individual objectives are met. Layout and typography are crucial to the success and effectiveness of these processes. Our reflection focuses on the possibility of making them usable to the blind. Several studies aim to improve non-visual scanning by using the tactile modality as in [23]; in this article, we focus on the skimming process: **oral transposition of web pages visual structure to promote the development of blind browsing strategies based on non-visual skimming process**.

The paper is structured as follows. Section II describes the related works on which our solution is based. Section III details the software architecture from the web page to the production of oral stimuli. It also presents the progress made in evaluating each module of the system. Finally, IV concludes the paper with a discussion and outlines future works.

II. Related Works

Much research in psychology and human-machine interaction has focused on the substitution of one sensory modality by one or more others [20]. Part of this research works to improve the access to documents content, in particular graphs [7], diagrams [27] and tables [21]. Only a few, as [17] or [26], have specialized in taking into account page layout and typographical effects; this dimension is too often evacuated despite its linguistic and cognitive interest. As [16], some approach aims to go further in exploiting the visual structure of document. The authors give it a role of support for interaction; the visual structure is then a vector of fluidity: if layout leaps out from the text at the eyes, it should be able to leap out at fingers or ears.

A. Positioning on Non-Visual Web Access

At first, non-visual accessibility should respect both the self-determination capacity of the user, the principles of "design for more" and to be part of an enactive approach to encourage the emergence of new and user centred designs.

1) Auto-determination: An essential criticism of many proposals is the methodological approach adopted. Most of them focus on what blind people want to do rather than on how to get it done [1]; risk is a limitation of the user's ability to self-determine. Another tendency to "think instead" of the blind can be observed: content is simplified to make it easy to digest in tactile or oral modalities. For example, [19] or [25] are working to reduce cognitive load by searching the web page for "relevant" information: the idea is to eliminate disruptions caused by "peripheral" elements or to use summary techniques. The designer considers that improving accessibility must sacrifice a certain amount of content. In doing so, the researcher or engineer puts himself in the position of the one who knows and imposes on the user what is interesting for him. The *University of Montreal*'s DEFI Accessibility group provides a satisfactory definition of universal accessibility: it is "the character of a product, process, service, environment or information that, with a view to equity and an inclusive approach, enables everyone to carry out activities independently and obtain equivalent results". It is this consideration of the user's self-determination in an interactive system that we wish to respect as a matter of priority.

2) Design For More: A limitation also observed for other sensory substitution systems is the incompatibility between the new solution and tools already known and used by the blind [2]. Even in a 'design for all' approach, this can reduce the usability of proven technologies on a daily basis and thus limit satisfaction with the new one. We support an approach we could call "Design for more" that aims to add new functionalities in a non-destructive way. Better still, their combination can be a source of discovery by serendipity.

3) Enaction/Emergence: Our ambition is not to identify ourselves subjectively with a theoretical and formatted reader. Our approach is based on the notion of enaction. Enaction is a theory, close to situated cognition, that focuses on how organisms organize themselves in interaction with

the environment. In an enactive perspective [29], we would like to provide the real user with all the visual information in its complexity. It is up to us to inject the appropriate stimuli in the environment, to encourage interaction and the development of the perception/action loop. It is up to each user to learn how to appropriate them and to perpetuate their own interpretations. It is up to time and practice to allow emergence of non-visual reading strategies by the user's adaptation to this interactive environment.

B. Positioning on Non-Visual Skimming

A second point is the propensity of these works to be directly oriented by the reality of the user's experience, sometimes too changing and immanent to methodologically extract new stable knowledge from it. To minimize this risk and create new paradigms, we justify the use of metaphor in the knowledge development process. Although its role in a perspective other than scientific mediation may be debatable, we believe that when the force of metaphorical tension is well adjusted, in a context of discovery, it participates in the work of inferring new hypotheses. We applied this point of view to build our non-visual skimming solution.

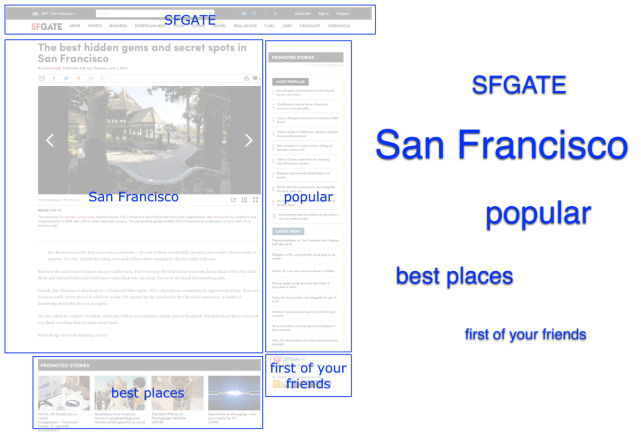


Figure 2. From Web Page to Tag cloud

Let's go back to the web page on the left of the Figure 1 given in the introduction. Let's continue with a segmentation into zones (see section III-1) and the extraction (see section III-2), for each area, of some keywords representative of the content (Figure 2, left). Finally, let's erase the other elements from the page; arrange the terms retained in the same spatial relationship as the area they represent and shape them graphically to make them all the more prominent as the area seems important. We obtain a visual stimulus frequently used in web 2.0: a tag cloud (Figure 2, right). The idea is to transpose this concept into the world of sound (see section III-3), transforming this cloud of words into a "thunder" of words (or tag thunder). The analogy underlying the development of this concept is an extension of the metaphor known to psychologists as the "cocktail party effect" [14].

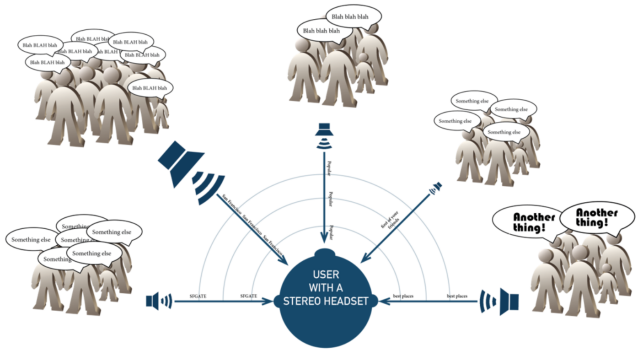


Figure 3. Cocktail Party Effect Metaphor

1) Cocktail Party Effect Metaphor: In psycho-acoustics, this metaphor denotes the possibility of focusing one's auditory attention on a verbal flow in the noisy atmosphere of a reception; whether it is directed towards sound sources outside the conversation or towards one's interlocutors. We will follow this metaphor thread by considering the relationship between the blind reader and the areas of the page being visited, in the same way as between a guest, located in the centre of a room, and the different discussion groups that have formed there: the exchanges are sequential within a group but competing between the different groups; the guest, as a newcomer, must take enough information from the sound environment to identify which discussion he or she wishes to get involved in. For example, in Figure 3, some terms emitted by each group of people are captured by the listener as repeated flows from separate parallel sources. The strength, density and speed of the flows depend on the characteristics of the issuing group; so the metaphor suggests a relationship between formal characteristics of the area and specific acoustic parameters, as proposed in section III-3.

2) Spatialized Concurrent speech: One of the solutions which has been mentioned in several research to improve the reading speed of blind people, is based on the notion of concurrent speech. In particular [8] shows experimentally the interest of this approach but also some perceptual limits. To go further, the originality of the work in progress is to propose to exploit 3D sound spatialization and binaural hearing techniques [9], to overcome these difficulties by optimizing the perceptual separation of sound sources. In the work presented here, spatialization is limited to a half-plane, perpendicular to the user, in which the 5 sound sources are placed in front of the listener on the left, right, centre and both diagonals.

The following section describes the software architecture tested to automatically produce a tag thunder from any URL.

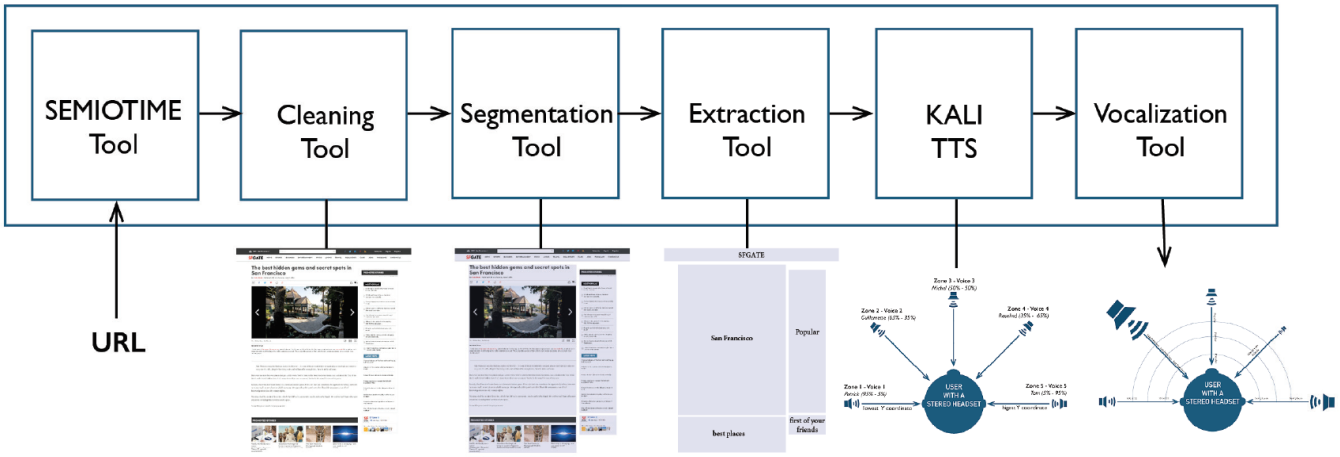


Figure 4. Modular Architecture: from URL to Tag Cloud and Tag Thunder

III. Software Architecture and Evaluations

To create a tag thunder from a URL, in addition to operations of centralizing information with visual impact (Dynamic JavaScript injection to obtain all the useful properties of style), our solution is based on three different main software layers. Web page segmentation, keywords extraction to create tag clouds and oral generation of tag thunders are each managed by a different layer, through a Firefox extension. Figure 4 illustrates the architecture and the different layers, detailed thereafter. Each following sub-subsection describes the theoretical principles, state of development and existing evaluations of the three main modules.

1) Web Page Segmentation (WPS): There are several approaches to segment a page into blocks. [24] proposes a method based on analyzing the Document Object Model (DOM) of a web page. [4] and [31] uses a visual approach by analyzing the page rendition in a browser to extract areas. Other techniques may be based on image processing [5], semantic structures or graph resolution [13]. Some, as [22], develops hybrid method. A critical review of these approaches can be found in [10]. However, they deal with tasks that imply constraints far from ours. We consider that non visual skimming requires three characteristics to be filled.

First, the number of zones has to be fixed in order to foster the emergence of regularities in the output and to comply with the maximum number of concurrent oral stimuli a human-being can cognitively distinguish. Within this context, [8], [15] have shown that the cognitive load can rise up to 5 different stimuli, thus limiting the maximum number of zones resulting from the WPS process. In the medium term, the stability of the WPS into exactly 5 zones will help the user to built associations between the position of a sound and a logical function often found in the same place for a given category of web pages (header, shopping cart, ...). As a consequence, it may enable the advent of new non visual reading strategies. *Second*, each zone should be associated to a unique sound source spatially located in accordance with its position in the web page. Thus, each zone should be a single compact block made of contiguous web elements, and the zones should not overlap. *Third*, segmentation must be complete, which means that no web page visible element should remain outside a given zone, as the objective is to reveal the overall visual structure of a document and not just parts of it.

We studied three different algorithms adapted to comply to these constraints: the classical k-means, the F-K-means (a variant of K-means, which introduces the notion of force between elements instead of the euclidean distance), and the Guided Expansion algorithm (GE), which follows a propagation strategy including alignment constraints. A manual evaluation of the algorithms has been performed by 3 experts measuring two clustering indicators: compactness and separateness. Compactness is defined at the cluster level and evaluates how many of the elements within a cluster belong to a same cluster in the ground truth. Separateness is defined at the web page level and evaluates how much the proposed segmentation guarantees the separability between clusters when compared to the expert ground truth segmentation. In particular, each expert must give a mark ranging from 0 (unacceptable), 1 (bad), 2 (passable), 3 (good) and 4 (perfect). Based on this protocol, the three algorithms have been tested on a total of 53 web pages from 3 domains: Tourism (23 web pages), E-Commerce (12 web pages) and News (18 web pages), that are part of a research project corpus[1]. Overall results are presented in table I.

It is clear that the guided expansion algorithm shows the best figures both in terms of compactness and separate-

[1]This dataset is freely available for research purposes.

		Compactness		Separateness		GlobalScore	
		Avg.	Std.Dev	Avg.	Std.Dev	Avg.	Std.Dev
K-ME.	Expert 1	2.42	1.16	1.15	0.64	0.30	0.12
	Expert 2	1.90	0.87	1.20	0.60	0.26	0.11
	Expert 3	3.10	0.74	0.70	0.80	0.29	0.15
F-K-ME.	Expert 1	2.43	1.46	0.62	0.57	0.23	0.09
	Expert 2	1.83	1.15	0.40	0.50	0.16	0.07
	Expert 3	3.05	1.22	0.30	0.50	0.21	0.095
GE	Expert 1	**2.89**	1.24	**1.62**	0.93	**0.42**	0.19
	Expert 2	**2.41**	0.81	**1.90**	0.90	**0.41**	0.16
	Expert 3	**3.40**	0.68	**1.50**	0.90	**0.44**	0.18

Table I
OVERALL RESULTS FOR K-MEANS (K-ME.), F-K-MEANS (F-K-ME.) AND GUIDED EXPANSION (GE).

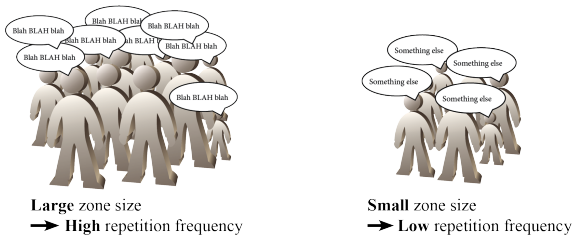
ness for the 3 human experts. However, while compactness receives average values between passable and good, separateness receives much lower values, between passable and bad. This finding is transverse to all three algorithms, clearly evidencing that finding coherent zones that match human expectations is a hard task, while building internally semantically coherent zones is easier. Also, figures show differences between K-means and F-K-means. In particular, both algorithms show similar compactness, but the F-K-means evidences worst results for separateness. This result can easily be explained as the F-K-means tends to create unbalanced clusters, that are either very small or rather big. This is confirmed by the higher standard deviation in terms of compactness for F-K-means than for K-means, signifying that F-K-means tends to create very compact clusters (but small) and "uncondensed" big ones, thus penalizing separateness.

However, human evaluation may be subject to bias as each expert evaluates the WPS process with his/her own subjectivity. As a consequence, we have proposed in [10] a quantitative evaluation that introduces different criteria of analysis. We showed that human and automatic evaluations are complementary to rank the algorithms according to several parameters (the number of inadequate separations of HTML elements, the number of overlaps between zones and the balance of created clusters), each parameter performing a specific complementary role for both compactness and separateness criteria. From both qualitative and quantitative evaluations, the Guided Extension algorithm seems to be the most efficient solution over all criteria.

2) keywords Extraction and Tag Clouds: Keyword extraction is achieved with an algorithm based on TF-IDF (Term Frequency-Inverse Document Frequency). The TF-IDF measure, used in Search Engines, finds context specific words since term frequencies are weighted by the inverse of the term frequency in a corpus. Hence, terms frequently found in the corpus have a lower score than words that frequently appear only in the computed text. Inspired from [30], our solution uses a TF-IDF measure coupled with the term's position within its block of text to calculate its score. This principle was improved by [11], using Ranking SVM instead of the raw position. The corpus used to compute the Inverse Document Frequency (IDF) is made of 953 551 articles of the newspaper "Le Monde", spanning 20 years from 1987 to 2006. Since the corpus' most recent data dates back to 2006, it introduced some silence and bias in the keyword extraction process. Although not yet evaluated, this technique seems to provide limited but satisfactory initial results when zones contain sufficient text. This difficulty will be addressed by working on the integration of meta-textual information resulting from the description of areas or images.

3) Tag Thunder Creation: Based on [3], [28], [6], specific voice, volume, prosody, speech rate and sound synchronization are combined to generate the audio signal from a given keyword and its zone properties. Our synthesis module uses the Kali TTS [18] tool, developed by the CRISCO lab. Kali supports speech rate acceleration without loss in intelligibility and sound quality, which is a very important feature in non-visual web browsing. We use several cocktail party effect metaphors to assign to the synthesized keywords a repetition frequency, volume and a location in the 2D audio space (3D solution is a work in progress). Vocalization of all the keywords with their specific parameters produces the final tag thunder.

Repetition frequency

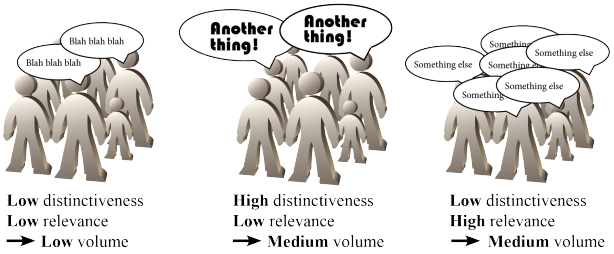


Metaphor 1 (size): the larger the group talking about a topic, the more often related terms emerge.
Rule 1: vocalized keywords are played in a loop. Zone size influences repetition frequency within the loop.

Figure 5. Repetition Frequency Metaphor

For each keywords, 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 (Figure 5).

Volume

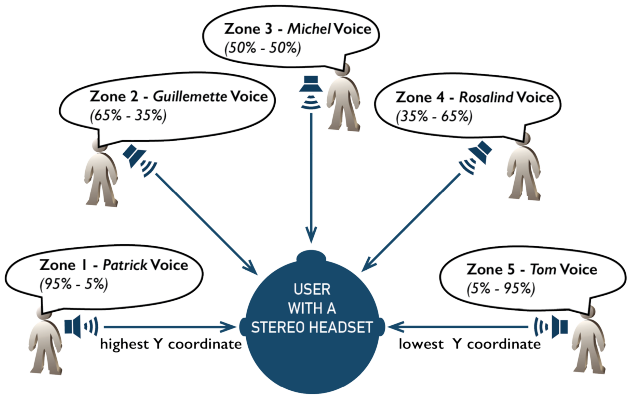


Metaphor 2.a (distinctiveness): the more a voice in a group stands out, the easier it is to detect its source.
Metaphor 2.b (relevance): the more the words are repeated in a group, the more relevant they are.
Rule 2: volume is determined by zone contrast and keywords frequency in the zone.

Figure 6. Volume Metaphor

For each zone, contrast is computed depending 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 keywords frequency (Figure 6).

Spatialization



Metaphor 3 (dimensions): 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 keywords.

Figure 7. 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 7.

Figure 8. Experiment Web Interface

Evaluation

An experiment was carried out according to the following protocol. 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/thunder 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 (Figure 8).

The experiment modalities were as follows:

- 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; there were as many correct (matching) couples as incorrect ones. Participants took the test autonomously, with a supervisor close by. Results, detailed in [15], show that participants found the exercise difficult but made few mistakes. In general, the results for tag thunder are comparable in the overall accuracy with the results of tag cloud. We concluded that the tag thunder concept is interesting but that certain limitations originate from the implementation of previous modules. It remains here all the difficulty of simultaneously (1) evaluating a modular system as a whole at the end of the processing chain and (2) identifying at the level of each software brick the exact origin of its limits [12].

IV. Conclusion and Future Works

In this article we have presented a general and theoretical approach to the problem of fast, global and non-visual access to web browsing. We started from the observation that the semantics of the visual architecture of web pages must be transposed into new sensory modalities. To allow blind users to exploit these new stimuli and increase their ability to develop high level reading strategies, we imposed specific constraints that led to develop the concept of tag thunder. We have detailed the progress of a first version of the various components that make up the software architecture. The main short-term developments concern (1) the improvement of web page segmentation and keywords extraction algorithms, (2) the improvement of the perceptual separation of sound sources by binaural techniques. (3) In an enactive approach, we can say that there is no perception without action; it is therefore a question of initiating a virtuous perception/action loop by adding interactive functionalities into the system. The work in progress is focused on this integration to manipulate the tag thunder and allows a navigation based on an easy alternation between global and local access to the web page; until the discovery of the desired textual information. Finally, we must built experiments to evaluate (4) with blind people the final skimming system and (5) how to combine this solution with scanning ones based on vibrotactile devices.

References

[1] V. Ashok, Y. Puzis, Y. Borodin, and I. V. Ramakrishnan, "Web screen reading automation assistance using semantic abstraction," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI 2017, Limassol, Cyprus, March 13-16, 2017*, G. A. Papadopoulos, T. Kuflik, F. Chen, C. Duarte, and W. Fu, Eds. ACM, 2017, pp. 407–418.

[2] S. F. Babiker, A. A. Ahmed, and M. A. A. Yasin, "Web navigation tool for visually impaired people," *IJITWE*, vol. 7, no. 1, pp. 31–45, 2012.

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

[4] 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.

[5] 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.

[6] 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.

[7] C. Goncu and K. Marriott, "Gravvitas: Generic multi-touch presentation of accessible graphics," in *Human-Computer Interaction - INTERACT 2011 - 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I*, ser. Lecture Notes in Computer Science, P. F. Campos, T. C. N. Graham, J. A. Jorge, N. J. Nunes, P. A. Palanque, and M. Winckler, Eds., vol. 6946. Springer, 2011, pp. 30–48.

[8] J. Guerreiro and D. Gonçalves, "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, ASSETS 2015, Lisbon, Portugal, October 26-28, 2015*, Y. Yesilada and J. P. Bigham, Eds. ACM, 2015, pp. 3–11.

[9] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 833–843, 2004.

[10] J. Jeyafreeda-Andrew, F. Maurel, G. Dias, S. Ferrari, and E. Giguet, "Web page segmentation for non visual skimming," *The 33rd Pacific Asia Conference on Language, Information and Computation*, to appear in 2019.

[11] X. Jiang, Y. Hu, and L. Hang, "A ranking approach to keyphrase extraction," *SIGIR'09*, 2009.

[12] G. Lejeune and L. Zhu, "A new proposal for evaluating web page cleaning tools," *Computación y Sistemas*, vol. 22, 12 2018.

[13] 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.

[14] I. Loddo and D. Martini, "The cocktail party effect. an inclusive vision of conversational interactions," *The Design Journal*, vol. 20, no. sup1, pp. S4076–S4086, 2017.

[15] E. Manishina, J.-M. Lecarpentier, F. Maurel, S. Ferrari, and M. Busson, "Tag thunder: Towards non-visual web page skimming," in *18th International ACM SIGACCESS Conference on Computers and Accessibility*, 2016, pp. 281–282.

[16] F. Maurel, G. Dias, J. Routoure, M. Vautier, P. Beust, M. Molina, and C. Sann, "Haptic perception of document structure for visually impaired people on handled devices," in *Proceedings of the 4th International Conference on Software Development for Enhancing Accessibility and Fighting Info-exclusion, DSAI 2012, Douro Region, Portugal, July 19-22, 2012*, ser. Procedia Computer Science, L. J. Hadjileontiadis, P. Martins, R. Todd, H. Paredes, J. Rodrigues, and J. Barroso, Eds., vol. 14. Elsevier, 2012, pp. 319–329.

[17] F. Maurel, M. Mojahid, N. Vigouroux, and J. Virbel, "Documents numériques et transmodalité. transposition automatique à l'oral des structures visuelles des textes," *Document Numérique*, vol. 9, no. 1, pp. 25–42, 2006.

[18] 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.

[19] B. Parmanto, R. Ferrydiansyah, A. Saptono, L. Song, I. W. Sugiantara, and S. Hackett, "Access: accessibility through simplification & summarization," in *Proceedings of the International Cross-Disciplinary Workshop on Web Accessibility, Chiba, Japan, May 10-14, 2005*, ser. ACM International Conference Proceeding Series, S. Harper, Y. Yesilada, and C. A. Goble, Eds., vol. 88. ACM, 2005, pp. 18–25.

[20] P. B.-y. Rita and S. W. Kercel, "Sensory substitution and the human-machine interface," *Trends in Cognitive Sciences*, vol. 7, no. 12, pp. 541–546, 2003.

[21] M. Rotard, S. Knödler, and T. Ertl, "A tactile web browser for the visually disabled," in *HYPERTEXT 2005, Proceedings of the 16th ACM Conference on Hypertext and Hypermedia, September 6-9, 2005, Salzburg, Austria*, S. Reich and M. Tzagarakis, Eds. ACM, 2005, pp. 15–22.

[22] W. Safi, F. Maurel, J.-M. Routoure, P. Beust, and G. Dias, "A hybrid segmentation of web pages for vibro-tactile access on touch-screen devices," in *3rd Workshop on Vision and Language (VL 2014) associated to 25th International Conference on Computational Linguistics (COLING 2014)*, 2014, pp. 95–102.

[23] W. Safi, F. Maurel, J. Routoure, P. Beust, M. Molina, C. Sann, and J. Guilbert, "Which ranges of intensities are more perceptible for non-visual vibro-tactile navigation on touch-screen devices," in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology, VRST 2017, Gothenburg, Sweden, November 8-10, 2017*, M. Fjeld, M. Fratarcangeli, D. Sjölie, O. G. Staadt, and J. Unger, Eds. ACM, 2017, pp. 81:1–81:2. [Online]. Available: https://doi.org/10.1145/3139131.3141222

[24] 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.

[25] J. Song, K. Choe, J. Jo, and J. Seo, "Soundglance: Briefing the glanceable cues of web pages for screen reader users," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019.*, R. L. Mandryk, S. A. Brewster, M. Hancock, G. Fitzpatrick, A. L. Cox, V. Kostakos, and M. Perry, Eds. ACM, 2019. [Online]. Available: https://doi.org/10.1145/3290607

[26] 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 - 14th International Conference, ICCHP 2014, Paris, France, July 9-11, 2014, Proceedings, Part I*, ser. Lecture Notes in Computer Science, K. Miesenberger, D. I. Fels, D. Archambault, P. Penáz, and W. L. Zagler, Eds., vol. 8547. Springer, 2014, pp. 61–68.

[27] M. Tixier, C. Lenay, G. L. Bihan, O. Gapenne, and D. Aubert, "Designing interactive content with blind users for a perceptual supplementation system," in *Seventh International Conference on Tangible, Embedded, and Embodied Interaction, TEI'13, Barcelona, Spain, February 10-13, 2013*, S. Jordà and N. Parés, Eds. ACM, 2013, pp. 229–236. [Online]. Available: https://doi.org/10.1145/2460625.2460663

[28] 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.

[29] Y. Visell, "Tactile sensory substitution: Models for enaction in HCI," *Interacting with Computers*, vol. 21, no. 1-2, pp. 38–53, 2009.

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

[31] J. Zeleny, R. Burget, and J. Zendulka, "Box clustering segmentation: A new method for vision-based web page preprocessing," *Information Processing & Management*, vol. 53, no. 3, pp. 735–750, 2017.