Using Karaoke to enhance reading while listening: impact on word memorization and eye movements

Emilie Gerbier, Gérard Bailly, Marie-Line Bosse

USING KARAOKE TO ENHANCE READING WHILE LISTENING: IMPACT ON WORD MEMORIZATION AND EYE MOVEMENTS

Emilie Gerbier[1], Gérard Bailly[1], Marie-Line Bosse[2]
[1] Univ. Grenoble Alpes/CNRS, GIPSA-Lab, F-38000 Grenoble, France
[2] Univ. Grenoble Alpes/CNRS, LPNC, F-38000 Grenoble, France
[1]firstname.lastname@gipsa-lab.fr, [2]marie-line.bosse@ujf-grenoble.fr

Abstract

This article reports the use of a karaoke technique to drive the visual attention span (VAS) of subjects reading a text while listening to the text spelled aloud by a reading tutor. We tested the impact of computer-assisted synchronous reading (S+) that emphasizes words when they are uttered, vs. non-synchronous reading (S-) in a *reading while listening* (RWL) task. Thirty-five 6th grade pupils read 12 stories, each involving one pseudoword presented four times, and each displayed in either condition. They were then unexpectedly tested on their memory for the orthography and for their acquired semantic knowledge of the pseudowords. Although no benefit was observed in the orthographic task, the synchronous condition significantly boosted the semantic memory by 10 points compared to the non-synchronous one (28% vs. 17% correct). We also provide some preliminary analysis on the gaze data collected during reading, suggesting differences between both conditions in terms of first fixation duration, fixation position on the word and onset delay relative to the corresponding speech onset.

Index Terms: reading while listening; eye tracking; visual attention span; memory; incidental learning; self-teaching hypothesis; audio-visual synchronization

1 Introduction

Multisensory presentation is known to improve memorization of words [1]: Shams & Seitz [2] have notably shown that hearing and seeing birds at learning stage improve their later identification on pictures even when no sounds are present at test. Similarly, aural-written verification during reading while listening (RWL) could help L1 and L2 learners to develop auditory discrimination skills, refine word recognition and gain awareness of form–meaning relationships. The basic idea is that incidental learning of words is favored by the joint activation of both orthographic and phonological forms [3]. This is consistent with the self-teaching hypothesis [4] that states that the phonological ability enables the learner to autonomously acquire an orthographic lexicon.

Several authors and publishers recommend the intensive use of digital texts, audiobooks and talking books [5][6] to increase the reading proficiency of students. The present paper adds to the corpus of experimental works that address the relevance and effectiveness of the use of synchronized talking books, i.e., a karaoke-style highlighting of words as they are being read aloud by a pre-recorded professional reader. The so-called synchronous reading (S+) requires an aural pacer, i.e. a prior alignment of text and oral reading.

2 State of the art

Numerous studies have demonstrated effects of redundancy in bimodal word processing [7][8]. Lewandowski and Kobus [9] notably showed a significant gain in word recall when the word was presented concurrently in the auditory and visual channels. Several cognitive models have been proposed to account for bimodal processing, for instance the Dual Route Cascaded model (DRC) [10] or the Bimodal Interactive Activation Model (BIAM) [11].

Several studies compared RWL with reading only (RO) or listening only (LO) conditions. Montali and Lewandowski [12] showed that RWL led to better comprehension of text passages than RO in normal readers, and better than RO and LO in poor readers. Moreover students preferred RWL over the other modes of presentation. Chang [13] found that students showed a strong preference for RWL vs. LO mode and gained 10% in word retrieval and comprehension. She also demonstrated that RWL to audiobooks increased listening fluency and vocabulary size [14]. Shany and Biemiller [15] also showed that RWL resulted in twice the amount of reading as teacher-assisted RO and led to higher scores on listening comprehension measures. Rasinski [16] however reported no significant difference between RWL and repeated reading for reading fluency of third-grade students. Holmes and Allison [17] also reported that fifth-grade good readers seemed to be negatively affected by RWL while the average and poor readers were not. Torgesen et al [18] also showed that length matters: the benefit of RWL over RO disappeared when learning-disabled adolescents studied chapter-length material over a week.

3 Synchronous reading

While text-to-speech systems can intrinsically provide synchronous reading [17] [18], several systems have already been proposed to synchronize audiobooks with the source text. The FAME system [21] can modulate

the granularity of the visual highlighting accompanying the audio narration of the text. The Talking Books Project [22] was an in depth study conducted in 10 infant classrooms with the aim of assessing the benefit of using electronic books in reading education. The software displayed text and images of a story book and introduced `read-aloud` features, whereby a child could select a sentence to be vocalized and watch the words being highlighted as they are vocalized (text and audio are synchronized). The children's comprehension, assessed via graded story re-telling, and their word decoding ability were tested before and after the use of the software and were contrasted with the results obtained under traditional teaching methods. The results concluded that there was a significant improvement in both comprehension and decoding when electronic books were used. The SWANS authoring system [23] has been used to produce prototype learning activities in various languages. Littleton et al. [24] showed that phonological awareness affects boys' use of talking books. Röber et al. [25] also showed that audiobooks combining narratives with game elements favors interactive behaviors.

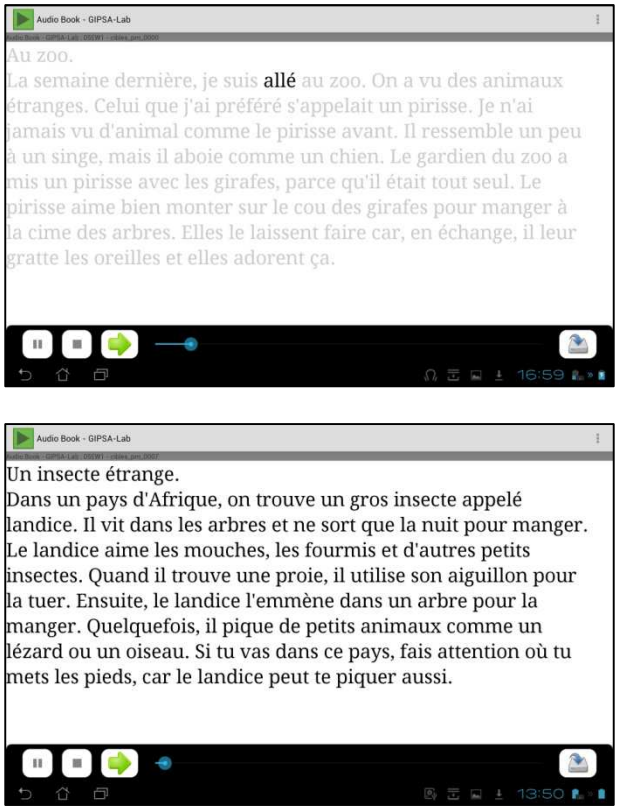

Figure 1. Screenshots of the ASUS screen for two RWL conditions used: Synchronous (S+; top) vs. non-sychronous (S-;bottom). Pseudowords are respectively "pirisse" and "claintond".

3.1 The GIPSA-Lab synchronous audiobook reader

Bailly & Barbour [26] have developed a first web-based application for karaoke-style RWL. An Android® application was further developed for the ASUS Transformer Pad, that offers a complete control of the visual presentation and highlighting of the text in synchrony with an audio file (see Figure 1). A xml file lists synchronization points at four different levels: the phone, the syllable, the word or the breath group. Each level can be differently highlighted using a combination of specific character/background colors. Sets of characters associated with pauses (e.g., space or punctuation) can also be highlighted by changing their background color. In this case, the highlighting fades away until the end of the pause. Each level has mandatory fields that allows interactive reading:

- Phone: duration, number of associated characters in the word, liaison feature (optional)
- Syllable: number of constitutive phones, accent level (optional)
- Word: orthography, part-of-speech (POS) tag
- Breath group: duration of the prephonatory pause

Each audiobook is organized in chapters. The audiovisual presentation of each chapter is described by a unique audio and xml file. The xml file also offers various parameter settings such as character size and style, default character, background colors, scrolling placement, line spacing, and margins.

Note that an average audiovisual asynchrony can be set by varying the duration of the initial pause.

```
SOMMES Aux 1 s o^ m _ _ z$
PEUT Vrb 1 p x _ t$
ABOYER Inf 0 a b w&a j e _
ADRIATIQUE Adq 0 a d r i&j a t i k _ _
ANNEXE Nom 0 a n _ e^ k&s _
BANJO Nom 0 b a~ _ d&z^ o
EPFL Npr 0 x^ p&e e^&f e^&l
```

Figure 2. Excerpt of the French aligned lexicon with four fields: the orthography, the POS tag, the feature indicating a latent liaison and the phonetic alignment. Latent liaisons are postfixed with a $. Multiple phones associated with the same character are linked with a "&". Note that this artifact allows letter spelling (see last line). Experimental design

3.2 Aligning resources

The mandatory fields of the four levels of the xml file can be obtained by automatically processing and aligning text files with the corresponding sound track. One of the key features of this processing is the character-to-phone correspondence. This correspondence is performed by a data-driven phonetizer [27] trained on an aligned lexicon of 200'000 French entries: the phoneme-to-grapheme alignment was performed semi-automatically using automatic refinements that include the use of silent or double phonemes (see Figure 2).

4 Experimental design

S+ and S- RWL were compared in French 6th grade pupils in a typical incidental self-teaching experiment.

Pupils silently read short stories in French whose topics revolved around a thing or an animal, named by a pseudoword (e.g., *landice*) that was presented four times in the text. They read the texts on the screen and could hear a narrator spell the story aloud. Half of the texts were presented in S+ and the other half in S- (see Figure 1). Eye movements were monitored during reading. The pupils were unexpectedly tested on their memory for the orthography of the pseudowords and for the semantic category of the pseudowords.

4.1 Audiovisual presentation

A first series of experiments on adult subjects had been performed to fine-tune the design of the experiment. Two of the main conclusions of these preliminary results were that:

- a strict synchrony between audio and visual presentations is uncomfortable: the word under focus should precede speech production by 300ms on average. This average delay may be text- and reader-dependent and deserves further research. We have however kept this average value for our experiment with the pupils.
- a strong highlighting combining change of background (e.g., yellow highlighting) and character color bothers adult readers. We thus decided to keep the background unchanged and to trigger focus on the current word only by switching its color from light gray to black (see top of Figure 2), black on white being the default contrast for S-. The focus word has thus the same appearance in S+ and in S-.

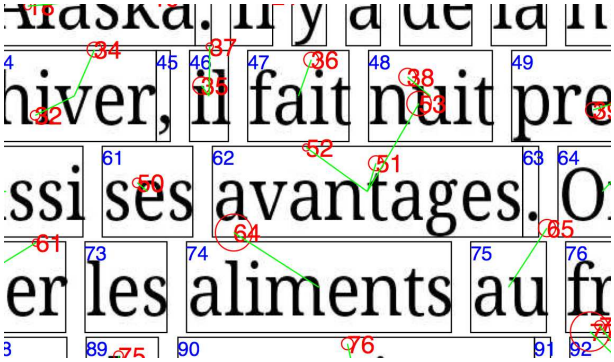

Figure 3. Example of associations between fixations (red circles) and bounding boxes of words (blue numbers), shown here by green lines.

4.2 Gaze tracking

In order to collect gazing strategies and study the impact of RWL conditions on the visual attention span, we monitored the pupils' gaze patterns by displaying the ASUS screen on a monitor that embeds a SMI® RED250 eye tracker. A very accurate synchronization between gaze data recording and audiovisual presentation is performed by inserting audio triggers in one of the audio channel of the stereo wave file played by the ASUS tablet. A home-made electronic circuit converts these triggers into TTL (transistor-transistor logic) signals that are fed into an appropriate pin of the parallel port of the PC driving the READ250 eye tracker. This signal triggers the monitoring of gaze data and guaranties perfect synchrony between gaze data and audiovisual stimuli.

Gaze fixations are associated with the bounding boxes of displayed words (note that the paragraphs and the character size have been designed to avoid scrolling) using Dynamic Time Warping (DTW) in order to compensate for calibration and tracking issues: local distances are proportional to the distance between the current fixation point and the considered bounding box (see Figure 3).

4.3 Method and procedure

Participants: Thirty-five 6th grade pupils were individually tested in their middle school in the area of Grenoble, France. They aged from 10;6 to 12;8 years old (19 girls) and belonged to two different classes. Their parents' consents were obtained. Nine pupils were assigned in each of the counterbalancing patterns W, Y, and Z, and eight pupils in the pattern X (see below).

Text material: 12 French stories, already used in previous unpublished experiments [28], were used. They were 82- to 100- words long and appeared on the screen all at once over 8 to 10 lines. Each story revolved around a fictional object or creature named with a pseudoword (PW; e.g., *landice*), that appeared four times in the text (neither in the title nor in the last 4 words of the text). Any given story always included the same PW. Twelve two-syllables PW were used that were 5- to 9- letters long and that were controlled for trigram frequency [29]**.** Each syllable of each PW could be written down using two homophonic graphemes, enabling us to create three alternatives to the target orthography of the PW (see Table 1 in the appendix). For instance, the target PW *pirisse* could be transformed to *pirice* or *pyrisse* (one grapheme changed), or *pyrice* (both graphemes changed). The alternative graphemes were on average as frequent in French as the target grapheme. The target orthography of the PW was chosen to avoid the most expected transcription using an inverse orthographic phonetizer. The alternative PW were similar to the target PW with respect to trigram frequencies in French. The 12 PW were divided into two sets (A and B) of 6 PW. The target graphemes used in subset A and B were different. For example, subset B included *l<u>an</u>di<u>c</u>e*, and subset A contained *piri<u>ss</u>e* and *m<u>en</u>dint*. Note that, in French, *an* and *en* are alternative graphemes for /ã/, and *ce* and *sse* alternatives for a final /s/.

Audio material: the texts were read aloud by a native male speaker at a moderate speaking rate (5.59 syllables/s) in a narrative mode.

Experimental design: During the reading phase, two blocks of six texts were displayed, each either in the S+ or in the S- condition. In order to compensate for potential serial and/or item effects on subsequent memory performance, the condition (S+ vs. S-) and the pseudoword subset (A vs. B) associated with each block were counterbalanced across subjects, creating 4 different counterbalancing patterns:

- W: set A as S+, then set B as S-
- X: set B as S+, then set A as S-
- Y: set A as S-, then set B as S+
- Z: set B as S-, then set A as S+.

Each subject was attributed a counterbalancing pattern and, in each counterbalancing pattern, the presentation order of the 6 pseudowords of a block was randomly determined.

Procedure: The pupils were informed that 12 texts (children short stories) would appear on the computer screen, one after the other, while those texts would be spelled aloud by a narrator in the headset in the same time. They were instructed to try to move their eyes over the text according to the pace that the narrator would adopt, and also to pay attention to the stories. Before each block of 6 texts, the experimenter described the settings of the forthcoming block (colors of text and words) and presented a very short, basic text as an example of those settings.
During the reading phase, the experimenter could monitor the pupils' eye movements on-line and check that the instruction of following the narrator pace was followed.
Immediately after the reading phase, pupils were tested on an orthography test in which they had to choose the correctly spelled pseudoword (i.e., how it was written in the stories) among the three other phonologically identical alternatives (e.g., *pirisse*, *pirice*, *pyrisse*, *pyrice*).
Next, they were presented with each pseudoword aurally and had to choose the right category among a list of 12 categories (e.g., cake, animal of the zoo, insect).
The experimenter then asked the participants about their preference for either reading mode, on a 5-point scale.

5 Results

5.1 Lexical orthography and semantic memory

The orthographic choice task was performed similarly for the S+ (Mean=2.23 correct response out of 6; Standard Deviation=1.35) and the S- items (M=2.0; SD=1.03). The responses were however above chance level (1.5 out of 6).
The semantic task was performed better for the items in the S+ (M=1.69 correct response out of 6; SD=1.21) than in the S- condition (M=1.06; SD=0.97; Wilcoxon test, W=367.5, p=0.015). The responses were also above chance level (0.5 out of 6).

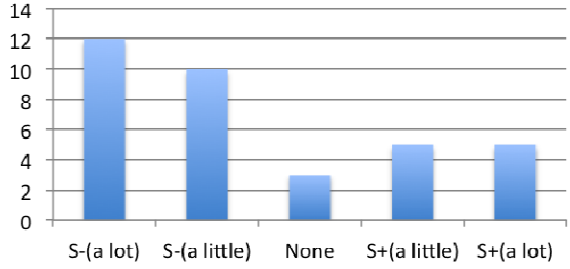

Figure 4. Number of children as a function of their RWL preference. S-: non-synchronous; S+: synchronous.

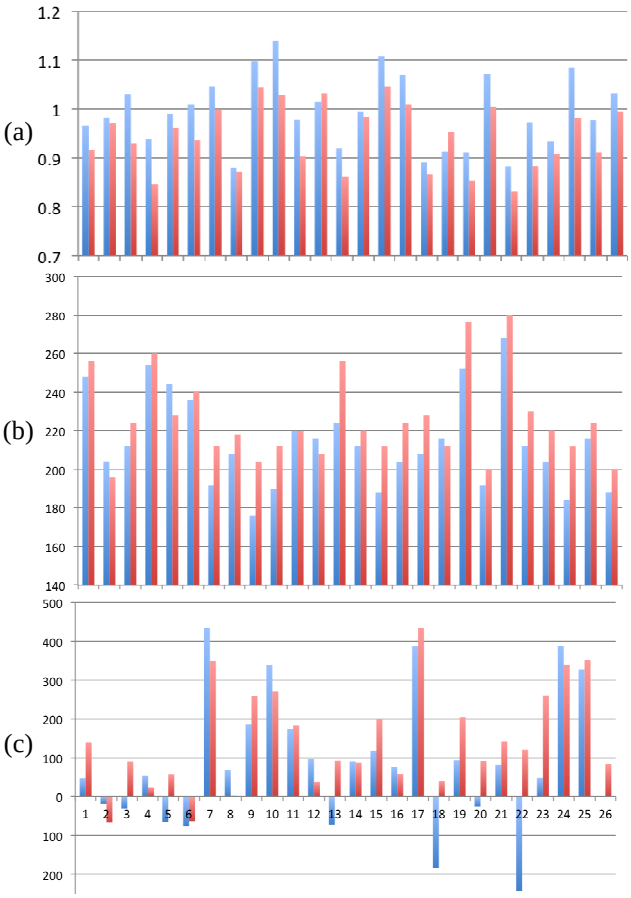

Figure 5. For each of the 26 subjects and LWR condition (S+ in red; S- in blue): (a) mean number of fixations per word; (b) mean duration (ms) of the first fixation on all fixated word and (c) relative onset of first fixation (ms) on all fixated word relative to its corresponding enunciation onset: because of the 300 ms asynchrony, a 0 onset means that the word is fixated 300 ms before it is spelled by the narrator.

5.2 Subjective preference

A majority of children (22) expressed a preference for the S- condition; most of them said that the S+ condition was uncomfortable or that it "went too fast". However, ten children preferred the S+ condition, saying that it

helped them follow the text. See Figure 4 for more details.

5.3 Gaze data

The gaze data from nine children were discarded due to an eye-gaze calibration issue or the fact that they deliberately did not follow the narrator's pace with their eyes (sometimes anticipating several lines ahead of the narrator). The new sample included 16 girls and 10 boys, from 11;4 to 12;8 years old. There were 6 pupils in counterbalancing pattern W, 5 in X, and 7 in Y and Z.

Considering all words (see Figure 5**).** Words were fixated more often in S- than in S+. Fixated words were fixated longer in S+ than in S-; Despite the asynchrony that we set, in which the words were highlighted 300 ms before they were spelled, the fixations were mostly anticipated relative to their actual highlighting in S+. Fixations also tended to occur earlier in S+ than in S- (in about 70% of subjects). This is true when all fixations or only the first fixations on a given word are considered.

In sum, the S+ condition did constrain the pupils' natural eye movements and forced them to fix the highlighted words less often but for a longer time than they would naturally do in a RWL task.

Considering only the pseudowords (see Figure 6. Figure 6**).** First fixations on pseudowords were longer in the S+ than in the S- condition (as it is the case for all words)**.** From the first occurrence of a pseudoword in the text to its fourth occurrence, first fixations tended to be shorter and more premature relative to the audio narration. Moreover, the fixation location on the word tended to move towards the right, from about 30% to 40% within the word bounding box.

6 Conclusions

Synchronous reading (S+) has interesting effects on the RWL task: (a) the significant improvement in correct rate of semantic recall compared with S- reading suggests that S+ has a positive impact on attention despite its sometimes uncomfortable practice; (b) Synchronous reading does change middle school pupils gaze behavior, since fixations are less numerous and last longer.

These preliminary results are quite encouraging for educational applications. If the benefit of S+ reading over classical RWL in recall and/or comprehension tasks is replicated in further studies, then its implementation in real-world reading situations can be promising. In particular, it could be used to help children learn to read more efficiently, especially those with reading difficulty, by helping them fixing the currently spelled words. More research is needed to understand the cognitive changes associated with S+ compared with S- RWL.

Several improvements can be still made to improve synchronous reading. First, more research is needed to investigate the most confortable duration of the audio-visual asynchrony, that was set to 300 ms here. Another question is the most relevant size of the unit to be highlighted: depending on the reading level and the educational aims, highlighting syllables, words, or phrases may be differently efficient.

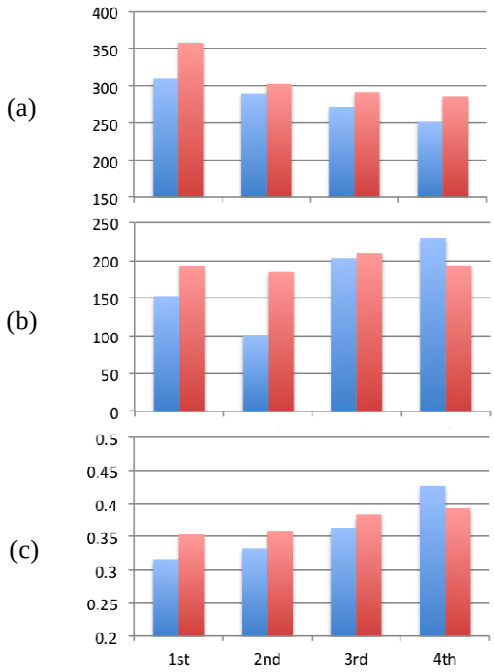

Figure 6. Mean measures for first fixations on any given PW, as a function of its occurrence in the text and RWL condition (S+ in red; S- in blue): (a) Mean duration (ms), (b) Mean anticipation (ms) (c) Mean position on the PW (proportion).

7 Acknowledgments

This work was supported by the ANR-12-BSH2-0013 ORTHOLEARN project. We warmly thank Julien Minet for the electronic circuit design.

8 Appendix

Table 1. List of pseudo-words. Alternatives 1and 2 differs from target PW by one grapheme while Alternative 3 -- the most regular orthography as predicted by our inverse orthographic phonetizer -- differs by two graphemes.

Target PW	Alternative 1	Alternative 2	Alternative 3
claintond	cleintond	claintont	cleintont
teinart	tainart	teinard	tainard
jaulu	jolu	jaullu	jollu
fortie	phortie	fortit	phortit
pirisse	pyrisse	pirice	pyrice
mendint	mandint	mendin	mandin
phatin	fatin	phatint	fatint
lyonit	lionit	lyonie	lionie
veingard	vaingard	veingart	vaingart
naigont	neigont	naigond	neigond
solloi	saulloi	soloi	sauloi
landice	lendice	landisse	lendisse

9 References

[1] L. Shams, Y. Kamitani, and S. Shimojo, "What you see is what you hear," *Nature*, vol. 408, p. 788, 2000.

[2] L. Shams and A. R. Seitz, "Benefits of multisensory learning," *Trends Cogn. Sci.*, vol. 12, no. 11, pp. 411–417, 2008.

[3] E. Maloney, E. F. Risko, S. O'Malley, and D. Besner, "Tracking the transition from sublexical to lexical processing: On the creation of orthographic and phonological lexical representations," *Q. J. Exp. Psychol.*, vol. 62, no. 5, pp. 858–867, 2009.

[4] Share, D.L., "Phonological recoding and self-teaching: Sine qua non of reading acquisition," *Cognition*, vol. 55, pp. 151–218, 1995.

[5] A. Thoermer and L. Williams, "Using digital texts to promote fluent reading," *Read. Teach.*, vol. 65, no. 7, pp. 441–445, 2012.

[6] G. Underwood and J. D. Underwood, "Children's interactions and learning outcomes with interactive talking books," *Comput. Educ.*, vol. 30, no. 1–2, pp. 95–102, 1998.

[7] S. A. Bird and J. N. Williams, "The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling," *Appl. Psycholinguist.*, vol. 23, no. 04, pp. 509–533, 2002.

[8] M. Yates, "Phonological neighbors speed visual word processing: evidence from multiple tasks.," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 31, no. 6, p. 1385, 2005.

[9] L. J. Lewandowski and D. A. Kobus, "The effects of redundancy in bimodal word processing," *Hum. Perform.*, vol. 6, no. 3, pp. 229–239, 1993.

[10] M. Coltheart, K. Rastle, C. Perry, R. Langdon, and J. Ziegler, "DRC: a dual route cascaded model of visual word recognition and reading aloud," *Psychol. Rev.*, vol. 108, no. 1, pp. 204–256, 2001.

[11] K. Diependaele, J. C. Ziegler, and J. Grainger, "Fast phonology and the bimodal interactive activation model," *Eur. J. Cogn. Psychol.*, vol. 22, no. 5, pp. 764–778, 2010.

[12] Montali, J. and Lewandowski, L., "Bimodal reading: benefits of a talking computer for average and less skilled readers," *J. Learn. Disabil.*, vol. 29, no. 3, pp. 271–279, 1996.

[13] A. C.-S. Chang, "Gains to L2 listeners from reading while listening vs. listening only in comprehending short stories," *System*, vol. 37, no. 4, pp. 652–663, Dec. 2009.

[14] C. Chang, "The effect of reading while listening to audiobooks: Listening fluency and vocabulary gain," *Asian J. Engl. Lang. Teach.*, vol. 21, pp. 43–64, 2011.

[15] M. T. Shany and A. Biemiller, "Assisted reading practice: Effects on performance for poor readers in grades 3 and 4," *Read. Res. Q.*, pp. 382–395, 1995.

[16] T. V. Rasinski, "Effects of repeated reading and listening-while-reading on reading fluency," *J. Educ. Res.*, vol. 83, no. 3, pp. 147–151, 1990.

[17] B. C. Holmes and R. W. Allison, "The effect of four modes of reading on children's comprehension," *Lit. Res. Instr.*, vol. 25, no. 1, pp. 9–20, 1986.

[18] J. K. Torgesen, W. E. Dahlem, and J. Greenstein, "Using verbatim text recordings to enhance reading comprehension in learning disabled adolescents," *Learn. Disabil. Focus*, vol. 3, no. 1, pp. 30–38, 1987.

[19] B. Pisha and P. Coyne, "Jumping off the page: Content area curriculum for the Internet age," *Read. Online*, vol. 5, no. 4, 2001.

[20] L. Hecker, L. Burns, and J. Elkind, "Benefits of assistive reading software for students with attention disorders," *Ann. Dyslexia*, vol. 52, pp. 243–272, 2002.

[21] C. Duarte and L. Carriço, "Developing an adaptive digital talking book player with FAME," *J. Digit. Inf.*, vol. 8, no. 3, 2007.

[22] Medwell, J., "The talking books project: some further insights into the use of talking books to develop reading," *Reading*, vol. 32, no. 1, pp. 3–8, 1998.

[23] A. Stenton, "Can simultaneous reading and listening improve speech perception and production? An examination of recent feedback on the SWANS authoring system," *Procedia - Soc. Behav. Sci.*, vol. 34, pp. 219–225, 2012.

[24] K. Littleton, C. Wood, and P. Chera, "Interactions with talking books: Phonological awareness affects boys' use of talking books," *J. Comput. Assist. Learn.*, vol. 22, no. 5, pp. 382–390, 2006.

[25] N. Röber, C. Huber, K. Hartmann, M. Feustel, and M. Masuch, "Interactive audiobooks: combining narratives with game elements," in *Technologies for Interactive Digital Storytelling and Entertainment*, Darmstadt, Germany, 2006, pp. 358–369.

[26] G. Bailly and W. Barbour, "Synchronous reading: learning French orthography by audiovisual training," in *Interspeech*, Florence, Italy, 2011, pp. 1153–1156.

[27] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter-to-sound rules," in *ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 77–80.

[28] Chaves, Nathalie, "Rôle du traitement visuel simultané dans l'acquisition des connaissances orthographiques lexicales," PhD Thesis, Université Toulouse II Le Mirail, Toulouse, France, 2012.

[29] R. Peereman, B. Lété, and L. Sprenger-Charolles, "Manulex-infra: Distributional characteristics of grapheme—phoneme mappings, and infralexical and lexical units in child-directed written material," *Behav. Res. Methods*, vol. 39, no. 3, pp. 579–589, 2007.