Eye Got It: A System for Automatic Calculation of the Eye-Voice Span

Mohamed El Baha, Olivier Augereau, Sofiya Kobylyanskaya, Ioana Vasilescu, Laurence Devillers

Eye Got It: a System for Automatic Calculation of the Eye-Voice Span*

Mohamed El Baha[1], Olivier Augereau[3][0000−0002−9661−3762], Sofiya Kobylyanskaya[2], Ioana Vasilescu[2], and Laurence Devillers[2]

[1] IMT Atlantique, Brest, France
mohamed.el-baha@imt-atlantique.net
[2] Université Paris-Saclay, CNRS LISN, Paris, France
sofiya.kobylyanskaya@limsi.fr
[3] Lab-STICC CNRS UMR 6285, ENIB, Brest, France
augereau@enib.fr

Abstract. Over the past decade, eye movement has been widely looked into for describing and analyzing several cognitive processes and especially for human-document interaction, such as estimating reading ability and document understanding. Most of the existing applications have been done for silent reading but we propose to explore reading aloud interaction through a powerful measurement named the "eye-voice span" which measures the distance between the eyes and the voice. In this paper we present an open-source platform named "Eye got it" and the underlying algorithms that can be used for processing eye-tracking and voice data in order to compute automatically the eye-voice span.

Keywords: Eye Tracking · Eye-voice span · Human-document interaction · Voice analysis

1 Introduction

Studying the way people are interacting with documents can give insightful information about the documents and the readers. For example, it can be used for detecting if a document is hard to understand (for all readers) or if a reader is struggling to read the document and needs help. In the same way it can be used to detect if a document is interesting [3] or has some emotional content [12] or if a reader is interested or feels something while reading the document.

Eye movement research related to reading starts to attract attention thanks to the development of more affordable eye-trackers, which are able to obtain accurate measurements. Eye-tracking is the process of measuring either the point of gaze (the position where we look) or the movement of the eyes. The tracking is done by an eye tracker that measures the position and movement of the eyes. A great deal of research studies has been done with eye trackers. The goal is generally to estimate the gaze position when doing an activity, such as driving

* This work is supported by ANR (ref ANR-20-IADJ-0007), DFG and JST

a car, shopping, or for marketing and industrial applications, etc. In our case, we are considering education applications and more specifically we focus on L2 (second language) learning paradigms. We aim thus to provide valuable feedback about L2 English learning by adults, such as English for non-native speakers. It has been shown that the pattern of the eye movement can be used to predict the level of English of a reader [2] and even predict the reader's TOEIC score after reading some texts [1].

Most of this research has been done for silent reading but some researchers also looked into reading aloud. By simultaneously recording the eye movements and the voice while reading aloud, we can see how these two measurements are related to each other and how they vary with time which can provide more information about the reader and help, for example, to detect dyslexia [7]. According to Laubrock and Kliegl, when we read a text aloud, our eyes are generally looking ahead from our voice [10]. This distance is what we call the eye-voice span (EVS). The EVS has been a center of various research since 1920 where Buswell found patterns that describe eye movement during oral reading similar to silent reading such as forward and backward saccades, fixations and word skipping [5]. According to several studies, the EVS can be an accurate indicator of the reader's reading skills and text understanding. According to Buswell, a skilled reader will tend to maintain a significant average span between the eye and the voice, while a novice reader will tend to keep the eye and voice very close together, in many cases not moving the eye from a word until the voice has pronounced it [6]. Silva et al. [17] demonstrate in their experience that the word familiarity and the word length have a strong effect on the eye-voice span. Laubrock and Kiegl showed that the EVS is constantly regulated [10] thought the reading time according to cognitive, oculomotor, and articulatory demands and that the EVS can be used to predict regression, fixations, and saccades which are related to the reading skill.

The EVS is commonly measured either with a time reference such as millisecond (it is then called the temporal EVS), or with a spatial reference such as the number of letters or words (in which case it will be called the spatial EVS). In his study, Buswell found out an average spatial distance of 15 letters for college students [5] reading a text in their native language. In more recent study, Inhoff et al. [9] reported a average temporal EVS of 500ms for standard reader and, De Luca et al. showed that dyslexic readers have an average spatial EVS of 8.4 letters whereas standard readers have an average spatial EVS of 13.8 letters [7].

The present work is part of a large-scale research project aiming at building an experimental paradigm dedicated to the learning of a second language using multimodal information such as the eye movement, the voice, and facial expressions. The part about facial expression will not be described in this paper. For this purpose and to help researchers to collect data, we built a system named "Eye Got it". It is open source and available on GitHub[4]. Our system integrates several algorithms for eye movement and voice processing and analysis, and computes the EVS automatically. In this paper, we will present the architecture of

[4] `https://github.com/oaugereau/Eye_Got_It`

this system, the underlying algorithms and describe preliminary data in terms of voice/eye movements recordings.

In the following section, we will start by defining the "Eye Got it" system and present the process of computing the EVS step by step. In the next section we will present an experiment that we set up in order to test and validate the system. Finally, we will conclude by explaining the limitations of the program and presenting possible improvements.

2 Eye Got it

We built an easy-to-use system called "Eye got it" that will allow researchers to record the voice and eye movement of people reading aloud texts in a second language with different levels. Our platform is used for displaying a text, recording eye tracking and voice data and displaying multiple choice questions to assess the user's understanding.

After recording the data, the same system processes eye tracking and audio features and calculates the EVS automatically. The EVS will then be integrated into the analysis permitting to evaluate the reader's level of understanding and language skills. The overview of the system is described in Fig. 1.

2.1 Eye tracking

Our system is compatible with stationary eye trackers such as Tobii Pro Nano [5]. These eye trackers are fixed with mounting plates on a computer screen and give us as a raw output the coordinates (in pixels) on the screen of the eye gaze through time. The first step consists in obtaining the words' positions of the screen and in processing the raw eye-tracking data by computing the fixations and saccades from the raw eye gaze data. When we are reading, the eye movement is not a smooth movement but a sequence of fixations (when our eyes maintain a position on a single location) and saccades which are quick movements of the eyes between fixations.

For detecting fixations and saccades, we implemented two existing algorithms into our system: a) the Buscher et al. algorithm [4], and b) the Nystrom and Holmqvist algorithm [13]. Buscher et al. algorithm detects fixations when neighboring gazes are closed to each other (based on the two-dimensional location on the screen) whereas Nystrom and Holmqvist algorithm detect saccades based on the angular rotation speed of the eyes. The output of the fixation-saccade processing is a list of fixations. Each fixation has a starting time, a duration and a center. Between a pair of fixations are the saccades. The result of the output of such an algorithm computed by our system is displayed in the Fig. 2.

It should be noted that the quality of the eye-tracking recordings depends on several factors such as the movements of the recorded person and especially the head movements, the brightness of the room, the presence of other infrared

[5] `https://www.tobiipro.com/product-listing/nano/`

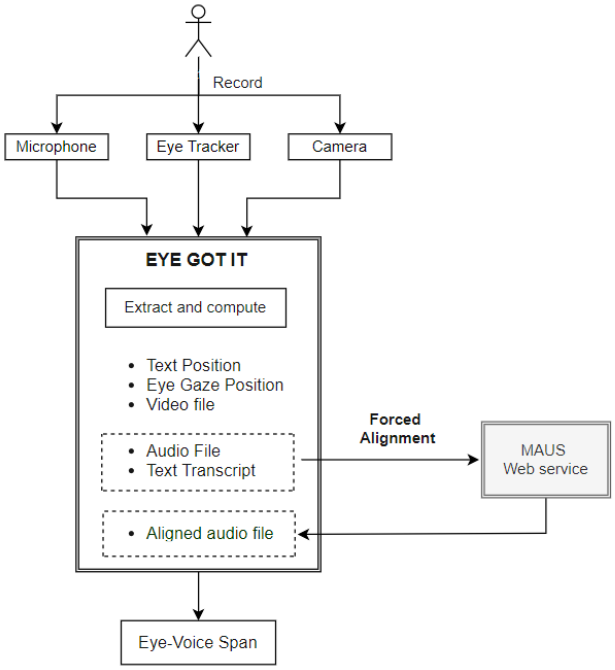

Fig. 1. The computing process of the EVS thought "Eye Got it" system. The system records the reader's voice with a microphone, the eye movement with an eye tracker and the facial expression with a camera. The position of the text and words on the screen is known by the system.

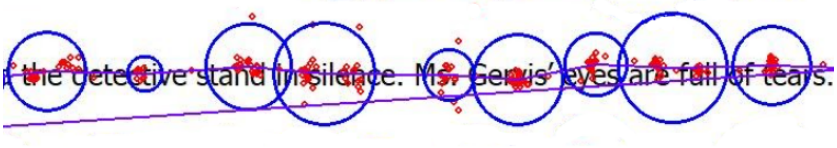

Fig. 2. Result of the Buscher et al. [4] fixation-saccade algorithm. The red dots are the eye gazes, i.e. the raw output of the eye tracker. The blue circles correspond to the fixations: the diameter of the circles is proportional to the duration of the fixation. The purple lines are the saccades. Long backward saccades are observed when the eyes of the reader jump from the end of one line to the next one.

lights than the eye trackers, the use of glasses or contact lenses, etc. For these reasons, it is important to frequently control the calibration of the eye-tracker and proceed to re-calibration if necessary. Even after careful calibration, it can happen to obtain an eye gaze recording that is inaccurate (such as the ones in

Fig. 3). If such inaccurate recordings are used to compute the EVS, the results will not be correct.

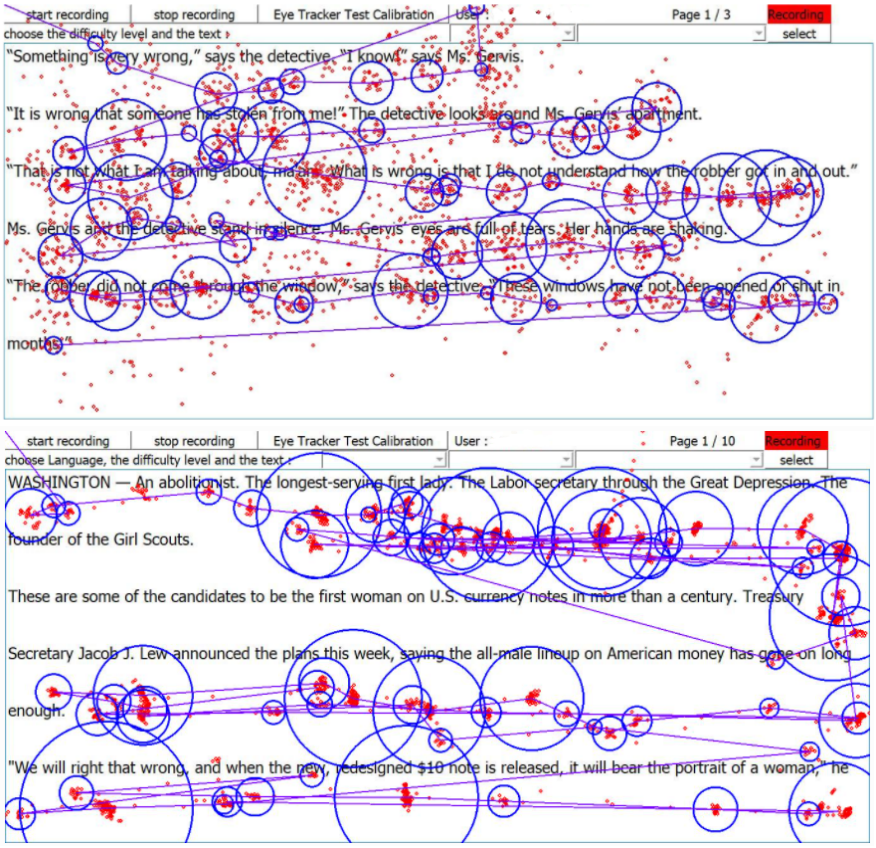

Fig. 3. Eye tracking recordings with low quality. On the top we can see that eye gazes are far from each other and this phenomenon is not possible (since the eyes cannot have such a movement), so we know that it is a problem from the eye tracker. On the bottom we can see another problem: the fixations seem natural this time, but they are not aligned with the text. This might be due to a calibration problem.

In order to avoid low quality recordings/processing errors, we set up three metrics to estimate the quality of a recording based on known eye tracking patterns from the literature:

– the percentage of fixations whose center is inside the bounding box of a word. When reading a text, the eyes are moving from word to word. Some fixations can be outside of the bounding box of a word, especially at the beginning and end of the recording or when the reader's eyes jump from one line to another but most of the time they should be aligned with the text;

– the percentage of words with at least one fixation. We know that there is not necessarily a fixation per word (for skilled readers, the eyes typically move about seven to nine letter spaces with each saccade [14]) but still, it is not possible to read a text without a significant percentage of words with a fixation;
– the percentage of eye gazes used to process a fixation. Except during the saccades, most of the eye gazes are near each other and will form the fixations. A high number of isolated eye gazes should not be obtained.

A threshold was experimentally defined for each metric. If the value of each metric is higher than the threshold, then the recording is validated for processing the EVS. The three thresholds can be found in Table 1.

Metrics	Thresholds
% Fixations in words	80%
% Words with at least a fixation	60%
% Eye gazes in a fixation	70%

Table 1. We introduce in our system three metrics to control the quality of the eye tracking recordings. The thresholds were selected experimentally and are dependent on the quality of the eye-tracker and on its calibration. If the values of the three metrics are higher than the thresholds then the recording is considered to be processed by our system.

2.2 Speech processing

Speech is used by many researchers as complementary information to eye-tracking features in different fields such as linguistics, psychology and cognitive sciences. For example, eye fixations combined with the analysis of the word stress used by the speaker can give salient information about word recognition strategy implemented by the speaker during a listening task [15].

The analysis of the correlation of the latency (the pause before starting to pronounce a word and the start of the fixation on the word that is to be read) with the eye gaze in object naming task allows to get information about linguistic planning: the fixation on a word is lasting until the phonological form of the word is recognized. This explains why the fixations on the longer words are longer than on shorter words [8]. In the object naming task, the analysis of eye fixations combined with speech helped to understand that two successive words can be processed in parallel by the speakers under the assumption that both words are known and easy to name by the speakers [8].

In order to compute the EVS, we need to know which word has been pronounced at what time. To do this we align the voice signal, recorded with a microphone and the read text. The Munich AUtomatic Segmentation system (MAUS) computes the phonetic labeling and segmentation of a speech signal

based on a given phonological pronunciation, i.e a forced alignment between a voice and a text [16]. MAUS is based on Hidden Markov Model and supports 21 languages. It provides an accurate annotation and will help us to save an important amount of time that could be spent on manual annotation. However, the presence of disfluencies (such as pauses and hesitations) especially in L2 production influenced by L1 specificity can worsen the performance of automatic speech recognition systems [18]. That is why manual annotation can be needed to adjust the frontiers of the words and phonemes detected automatically.

In spontaneous interactive speech, disfluencies reflect various speaker's intentions such as maintain their speaking turn, end the interaction, mention a new piece of information [19]. We suppose that the alignment of disfluencies with eye gaze during the reading aloud experiment can provide aid in formulating hypotheses about cognitive processing of the text such as an attempt to pronounce correctly a word or a combination of letters, to get the sense of the sentence by returning the eye to the beginning of the sentence or of the text, to understand the user's interface.

2.3 EVS computation

To compute the EVS, we will combine the two results that we have obtained on a single axis of time, namely the audio file aligned with the script and the position of the eyes, which will allow us to calculate this distance between the spoken word, and the gazed word. To get the EVS, a user will need to perform the four following steps with "Eye got it": (1) recording the eye movements and the voice while reading aloud, (2) processing the fixations and saccades, (3) aligning the voice with the text from the other and (4) computing the distance between the eye gaze and the pronounced word. We detail these three steps as follows.

1. The user starts a recording session and reads a text aloud displayed on the screen via the "Eye Got it" interface. During this time a microphone and an eye tracker start recording. Optionally a camera can be used for recording the facial movement but this will not be detailed further in this paper. The program then organizes all the collected data in the following format: the coordinates of each word and their corresponding bounding box ("Text position"), the voice recording ("Audio file"), the text file read during the session ("Text input") and the coordinates of the gaze indexed by time during the reading session ("Gaze position").
2. The fixations and the saccades are computed. If the center of a fixation is inside the bounding box of a word, the fixation is associated with that word. This way, we can deduce the specific moment when a word is actually being looked at by the user. The quality of the recording is computed with three metrics, if the quality is not high enough then the other steps will not be processed.
3. The voice is aligned with the text based on the MAUS web service. The recorded audio file and the text are given as an input and MAUS returns a file which contains audio snippets with words from the text read by the user.

4. Using the aligned audio file with the text to read, the system checks if a word has been looked at when the user pronounced the word. If this is the case, it will compute the distance between the word pronounced and the word looked at. This distance can be represented as a time duration and the number of words between the word pronounced and the word looked at.

3 Experiment

In this section we present a preliminary experiment that we set up for testing and validating the different parts of the system and the user experience of "Eye Got it".

3.1 Participants

Five French native participants volunteered to participate in the experiment (average age 25 years, one woman and four men). All had normal or corrected-to-normal vision and had English as a second language (L2 speakers). The participants were asked to read three texts aloud while their eye movements and voice were being recorded. They all signed a consent form and were free to stop the experiment at any time.

3.2 Apparatus and Material

The texts are displayed on a 22-inch computer screen with a resolution of 1280 x 960 pixels. The voice was recorded using an AKG Perception Wireless 45 Sports Set Band-A 500-865 MHz microphone, connected to the computer via a USB Jack cable. Video recording is done via the computer's internal web camera, but this data has not been used. The eye-tracking recording is done using a professional eye tracker, a Tobii Pro Nano with a sampling rate of 60 Hz, fixed underneath the screen. All the devices are controlled by the Eye-Got-It software on a standard PC. The texts displayed are from the corpus "English For Everyone"[6]. Three texts of different levels are displayed (one text per level): beginner (170 words), intermediate (260 words), and advanced (237 words). All texts are spread over three pages with a font size of 25 pts.

3.3 Procedure

Firstly, we calibrated the eye tracker and tested if all devices were correctly functioning. The calibration is done with Tobii "Eye Tracker Manager" and is checked after each recording via the recording interface. Secondly, each participant creates a session by entering their information (name, first name, sex, date of birth, and language level). After that, the participants are asked to read aloud the text that is displayed on the screen. They will go through the three texts of

[6] `https://englishforeveryone.org/`

different levels (beginner, intermediate, and advanced) that are displayed in a random order. Thirdly, after the recording session, all data collected during the session is saved and organized automatically by "Eye Got It". Before processing the EVS, the quality of the eye tracking recordings is estimated with the three proposed thresholds. If the quality seems to be not high enough, we advise to not compute the EVS. We observe that we sometimes found low quality recordings (as previously shown in Fig. 3), despite meticulous calibration and care during the recording. Finally, the system processes and plots the EVS for each recording of each participant.

3.4 Results

The format of the output is a graph such as the one displayed on Fig4, where the y-axis is the EVS defined by the number of words and in the x-axis is the total duration of the lecture (in seconds). In this example, the maximum value of the EVS is seven words, which might correspond to the easier part of the text (for this reader). When the EVS reaches zero, it means that the reader is pronouncing the same word that he or she is looking at. The variation of the EVS is correlated to the cognitive process of each word and thus the understanding of that word. The average of the EVS will reflect the language skill of the reader. In general, the preliminary results seem to be consistent with the tendencies described in [9] [4] [17], but further recordings will be needed to confirm this outcome.

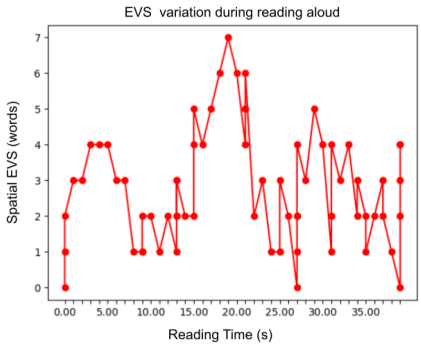

Fig. 4. An example of an automatically generated EVS graph of a participant while reading a text aloud. The variation of the EVS reflects the reader's reading skill.

4 Discussion

In this section, we comment on the preliminary results obtained during the first experiment setup conducted among 5 participants and analyze the two aspects

of the system: from the eye tracking and speech modality processing issues. The described results permit us to verify the correct functioning of the "Eye Got It" platform and will serve us as a baseline for the system's quality improvement as well for further data collection. We also present the limitations of the systems and their possible remediation.

4.1 Eye tracking

A pillar factor of an accurate EVS computation is eye movement recordings, and as explained in the sections above, eye-tracking is very sensitive and depends on many parameters. Nevertheless, we can obtain decent results by following good practices during the recording, such as controlling the brightness of the room, using multiple calibration points and asking the participants to restrain their body movement while recording. However, this will not guarantee the quality of the recorded eye gaze, and we still might end up with some recordings that are not usable. In this case, it is important to have a way to automatically estimate the quality of the recordings. Our proposition with three thresholds must be refined in the future. In the cases where all the fixations are shifted vertically, some algorithms have been proposed by some research to automatically correct the recordings [20,11]. For example, Lima Sanches et al. proposed a vertical error correction based on Dynamic Time Warping and match the text lines and the fixations lines [11].

4.2 Audio

From the audio perspective, the precision of the calculation of EVS depends on the quality of the audio provided to the system and on the reading strategies of the participants. We identified two main difficulties that need to be taken into account when calculating EVS.

The first one will happen if an audio file which does not correspond to the text to read is used in MAUS. As MAUS is based on a forced alignment, it cannot detect this kind of error and will still give us an output. But the output of MAUS will lead to a totally incorrect EVS. The Fig. 5 illustrates two following situations: one where the EVS is calculated with an audio file corresponding to the text input and one where the audio does not correspond to the text input. As we can see, the EVS where a different text was read than the expected one, the number of points is much lower which could indicate that there is a problem. Another parameter that could also influence MAUS alignment is the accent of the reader. In our experiment, the participants were French native speakers but their English pronunciation was correct enough to be aligned with MAUS. But if a reader has a very strong accent, the alignment might fail (in the same way as if a different audio file was used) and will also lead to an incorrect computation of the EVS.

The second possible difficulty is the use of disfluencies (such as pauses, hesitations, stuttering, repetitions, etc) by the speakers while reading aloud. As we discussed earlier, the disfluencies are crucial elements in verbal communication,

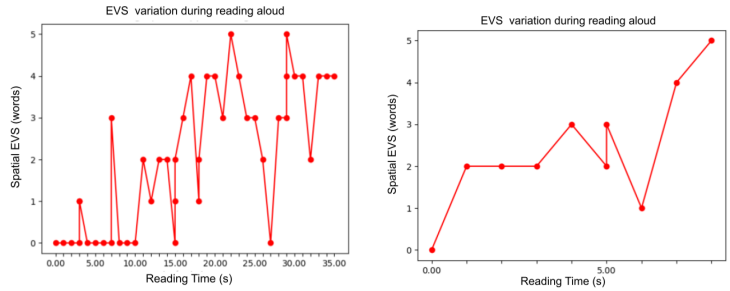

Fig. 5. Two EVS graphs: on the left the audio input was corresponding to the text input; on the right the audio file was not corresponding to the text input (i.e. the reader was not reading the displayed text). We can see that the discrepancy between the text input and the audio will be reflected by fewer points on the EVS graph.

so we cannot ignore them while analyzing the experiment, but they also can lead to errors in speech-text alignment and thus in computing the EVS. For instance, the Fig. 6 shows the difference between two EVS computed with and without stuttering when the same text is read by the same speaker.

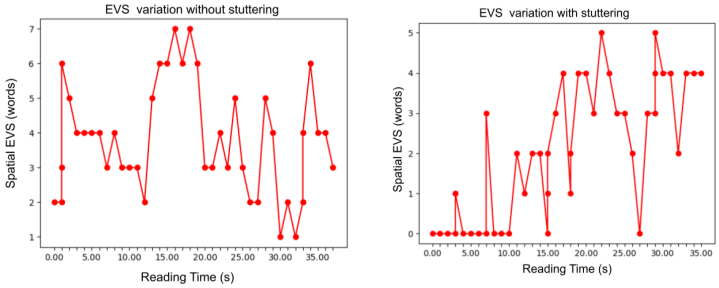

Fig. 6. The EVS graph of a reader reading the text with few disfluencies (left), and the EVS graph of same text but with stuttering. The stuttering tends to generate more zero values.

As we can observe, the EVS often reaches zero when the speaker is stuttering, which is not the case without it. One possible solution to this problem is to add disfluencies to the text to input and to correct MAUS output manually before the EVS calculation.

5 Conclusion

Current approaches to measure the EVS when reading aloud are not fully automatic and require manual intervention in several parts of the process. We

propose in this paper "Eye Got It", an automatic system for computing the eye-voice span, ready-to-use, free, and easy to manipulate. The accuracy of the results expected from this system is highly dependent on the quality of the input data. In particular, the quality of the recordings, whether it is the eye tracking recording or the audio recording. Thus, the proposed system finally provides an aggregation of several tools and methods to automate the computation of several features that helps to describe and analyze the skill of a second language reader. We also presented a preliminary experiment conducted on five participants to test and to validate that the system is working correctly.

Many points of improvement were raised, notably, the need for a metric to evaluate the accuracy of the audio alignment, since MAUS processes an alignment even if the audio and text input are not corresponding. Furthermore, MAUS is designed for native speakers, so readers with a strong accent cannot use the system or they might obtain an EVS that does not reflect their reading behavior. A possible improvement to solve this problem is to integrate an alignment model for non-native readers. We are able to classify native and non-native readers, via a convolutional neural network trained on spectrograms, the next step would be to train an alternative version of MAUS for the mother language of the reader in order to take into account their accent.

Acknowledgment

We would like to thanks the students from ENIB who participated to the development of the "Eye got it" system: Axel NOUGIER, Victor MENARD, Marine LE GALL, Yohan MAUPAS, Nicolas MENUT, Maelie MIGNON, Alexandre TROFIMOV and Asma NAIFAR.

References

1. Augereau, O., Fujiyoshi, H., Kise, K.: Towards an automated estimation of english skill via toeic score based on reading analysis. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 1285–1290. IEEE (2016)
2. Augereau, O., Kunze, K., Fujiyoshi, H., Kise, K.: Estimation of english skill with a mobile eye tracker. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct. pp. 1777–1781 (2016)
3. Buscher, G., Dengel, A., Biedert, R., Elst, L.V.: Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. ACM Transactions on Interactive Intelligent Systems (TiiS) **1**(2), 1–30 (2012)
4. Buscher, G., Dengel, A., van Elst, L.: Eye movements as implicit relevance feedback. In: CHI'08 extended abstracts on Human factors in computing systems, pp. 2991–2996 (2008)
5. Buswell, G.T.: An experimental study of the eye-voice span in reading. No. 17, University of Chicago (1920)
6. Buswell, G.T.: The relationship between eye-perception and voice-response in reading. Journal of Educational Psychology **12**(4), 217 (1921)

7. De Luca, M., Pontillo, M., Primativo, S., Spinelli, D., Zoccolotti, P.: The eye-voice lead during oral reading in developmental dyslexia. Frontiers in human neuroscience **7**, 696 (2013)
8. Huettig, F., Rommers, J., Meyer, A.S.: Using the visual world paradigm to study language processing: A review and critical evaluation. Acta psychologica **137**(2), 151–171 (2011)
9. Inhoff, A.W., Solomon, M., Radach, R., Seymour, B.A.: Temporal dynamics of the eye–voice span and eye movement control during oral reading. Journal of Cognitive Psychology **23**(5), 543–558 (2011)
10. Laubrock, J., Kliegl, R.: The eye-voice span during reading aloud. Frontiers in psychology **6**, 1432 (2015)
11. Lima Sanches, C., Augereau, O., Kise, K.: Vertical error correction of eye trackers in nonrestrictive reading condition. IPSJ Transactions on Computer Vision and Applications **8**(1), 1–7 (2016)
12. Matsubara, M., Augereau, O., Sanches, C.L., Kise, K.: Emotional arousal estimation while reading comics based on physiological signal analysis. In: Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding. pp. 1–4 (2016)
13. Nyström, M., Holmqvist, K.: An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. Behavior research methods **42**(1), 188–204 (2010)
14. Rayner, K., Chace, K.H., Slattery, T.J., Ashby, J.: Eye movements as reflections of comprehension processes in reading. Scientific studies of reading **10**(3), 241–255 (2006)
15. Reinisch, E., Jesse, A., McQueen, J.M.: Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. Quarterly Journal of Experimental Psychology **63**(4), 772–783 (2010)
16. Schiel, F.: A statistical model for predicting pronunciation. In: ICPhS (2015)
17. Silva, S., Reis, A., Casaca, L., Petersson, K.M., Faísca, L.: When the eyes no longer lead: familiarity and length effects on eye-voice span. Frontiers in psychology **7**, 1720 (2016)
18. Tomokiyo, L.M.: Linguistic properties of non-native speech. In: 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100). vol. 3, pp. 1335–1338. IEEE (2000)
19. Vasilescu, I., Rosset, S., Adda-Decker, M.: On the role of discourse markers in interactive spoken question answering systems. In: LREC (2010)
20. Yamaya, A., Topić, G., Martínez-Gómez, P., Aizawa, A.: Dynamic-programming–based method for fixation-to-word mapping. In: International Conference on Intelligent Decision Technologies. pp. 649–659. Springer (2017)