Towards an Automated Estimation of English Skill via TOEIC Score Based on Reading Analysis

Olivier Augereau, Hiroki Fujiyoshi and Koichi Kise
Osaka Prefecture University
1-1 Gakuencho, Nakaku, Sakai, Osaka 599-8531, Japan
augereau.o@gmail.com,
fujiyoshi@m.cs.osakafu-u.ac.jp,
kise@cs.osakafu-u.ac.jp

***Abstract*—Estimating automatically the degree of language skill by analyzing the eye movements is a promising way to help people from all over the world to learn a new language. In this study, we focus on the English skills of non-native speakers. Our aim is to provide an algorithm that can assess accurately and automatically the TOEIC score after reading English texts for few minutes. As a first step towards this direction, we propose an algorithm that can predict accurately this score after reading and answering some questions about the comprehension of few English texts. We use an eye tracker in order to record the eye gaze, i.e. the positions where the reader is looking at. Then we extract several features to characterize the behavior, and consequently the skill of the reader. We also add a feature based on the number of correct answers to the questions. By using a machine learning based on multivariate regression, the score is estimated user independently. A backward stepwise feature selection is used to select the relevant features and to optimize the estimation. As a main result, the TOEIC score is estimated with 21.7 points of mean absolute error for 21 subjects after reading and answering the questions of only 3 documents.**

I. Introduction

In Japan, the international English tests such as EIKEN, TOEFL or TOEIC are popular tests for assessing the language comprehension [2]. According to the report “Data & Analysis 2014” [17], the worldwide number of participants of TOEIC in 2014 was roughly 7 millions people, including 2.4 millions of Japanese participants. The “Report on Test Takers Worldwide” [16] shows that this standardized test has been taken in around 150 countries in the world.

This test consists of two parts: listening and reading comprehension. The listening comprehension lasts 45 minutes and the result is evaluated from 5 to 495 points. The reading comprehension lasts 75 minutes and the result is evaluated from 5 to 495 points. The final result ranges between 10 points to 990 points which indicates the level of understanding of the participant (the higher, the better). The correlation between listening and reading is around 0.85 and 0.89 [16], so estimating only the reading comprehension can give an accurate estimation of the TOEIC score.

Taking an examination such as TOEIC costs time and money: the participant needs to pay the inscription fee, go to a specific examination center, spend two hours for the test, and the result is obtained one or two months after the examination.

In order to help Japanese students, and people all around the world who would like to learn a new language, our goal is to build a system that can estimate in real time and automatically the reader’s skill simply by analyzing the eye movements while reading a text. Furthermore, recent research showed that it is possible to use an affordable eye tracker [11], webcam or even an unmodified tablet [18] to analyze the eye gaze, so our system could be used by a large number of people at home or school.

This research takes place in the context of the reading-life log [1]. It consists in analyzing the eye movements while reading. We want to quantify and qualify the reading activity and to extract mutual information between the readers and the documents during the daily life. Until now, some research have been done for counting the number of read words [8], classifying the type of read documents [10], analyzing the similarity of the documents [6], finding and annotating which parts of a document are difficult [13], analyzing the reader engagement [9], and in the present case, estimating the reader language skill.

In this paper, we propose the first step towards the automatic estimation of the TOEIC score based on eye tracking. We ask a subject to read a double page document containing a text on the left page and 4 MCQ (Multiple Choice Questions) related to the comprehension of the text on the right page. This is exactly the same principle as used in the Part7 of the TOEIC reading comprehension. We record the eye movement of the reader while reading the text and solving the questions. After reading few documents, we combine the eye gaze features with the number of correct answers to the questions and estimate the subject TOEIC score. The TOEIC score of each participant is known so the performance of our system is computed as the absolute error between the estimated score and the score from the ground truth. We will show that the performance of the system is getting more and more accurate when the number of read documents used to evaluate the skill of the subject increases. By using only 3 documents, the mean absolute error of the score estimation is 21.7 points (the TOEIC score is between 10 to 990 points) for 21 subjects.

The rest of the paper is organized as follows. In the next part, we present the related work about the analysis and estimation of reader’s skill based on eye movements. Then, we will explain the details of our algorithm to predict the TOEIC score. In the following section, the database, the subjects and the results about the experiment will be developed. At the end, we will discuss about the conclusion and future work.

II. Related Work

We start this section with some general points about eye tracking for reading. If we record our reading behavior with an eye tracker, we obtain a sequence of eye gazes representing the time and position where we look at. The sequence of eye gazes is then represented as a sequence of fixations (when the eyes stop on some specific places to extract information about the text) and saccades (a quick eye movement between two fixations) [14].

A. Analysis of the reading behavior and reader's skill

According to Rayner et al., the patterns of the eye gaze movements and the reading comprehension are correlated [15]. Their experiments show that, if the reader does not understand the text he reads, he might produce more fixations, spend more time and reread around the difficult words.

Kang realized a user study in order to compare the reading patterns of first language readers (L1) and second language (L2) readers [5]. He showed that even if the readers' attention distribution and understanding are similar between L1 and L2, L1 read approximatively twice faster than L2.

B. Language understanding estimation

Kunze et al. examined how to estimate the English skill of a reader [7]. They show that the average number of fixations and the standard deviation of the number of fixations is different depending on the reader skill. However, only 5 readers took the experiment and as the authors mention, the difference between the most skilled and least skilled reader is "not statistically significant". Furthermore, one of the readers does not fit to the assumption.

In 2014, Martinez-Gomez et al. proposed a system to recognize the understanding level and the Test of English (ToE) language skill (TOEIC or TOEFL score) of a reader [12]. They compare the result of their algorithm with the baseline regression (the output is the average value on the training data). The baseline regression performance is low: 85% of error for predicting the understanding and 86% of error for predicting the ToE. Unfortunately, the performances obtained by the authors are just slightly higher. In the best case they obtain: 82% of error for predicting the understanding and 78% of error for predicting the ToE. These numbers are not given by the authors but were estimated from the result figure.

In the same year, Copeland et al. proposed to estimate the reading comprehension based on eye gaze movements by using artificial neural networks [3]. The authors propose 4 different experiments where the reader have to answer 2 questions: one MCQ and one Fill-in-the-blank question. For each experiment, the subject has to read a text and the questions, but the displaying order of the text and questions is changed. The authors compare different classification algorithms and 2 Artificial Neural Network with 3 hidden layers and a different number of neurons. The misclassification rate of all classifiers are quite low, but the reason is because the dataset is heavily unbalanced (most answers are good). If we compare the result of the author with the baseline classifier (always classify to the largest class, in this case "correct answer") the result are the same or under the baseline.

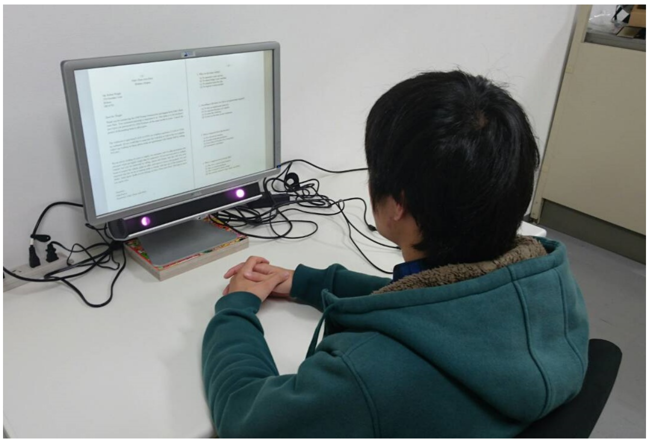


Fig. 1. Recording session with one subject reading one document. The eye tracker is fixed at the bottom of the screen. The document is displayed. It is constituted of 2 pages: the left page contains the text and the right page contains 4 MCQ with 4 possible answers for each question.

In 2015, Yoshimura et al. proposed a method to classify the TOEIC score in three classes: low, middle and high [19]. By computing 33 features related to blinks, saccades, fixations, etc. the authors succeeded to classify the English skill of 11 readers in 3 classes with an accuracy of 90.9%. In order to obtain this result, the readers needed to read 10 documents. This result is accurate but the authors shows that, if the number of documents is decreased or the number of classes increased, the performance drops drastically (by using 4 classes, the performance is only 54.5%).

As we can see, the current proposals for estimating the reader skill are promising but the results are still not enough convincing. Depending on the related work, the few number of readers and the unbalanced dataset make the task challenging. The proposal of Yoshimura et al. [19] seems the most promising but we want to propose a more discriminative method, i.e. not classifying the skill in 3 classes but directly estimating the score. Furthermore, our experiment consists in recording the reading behavior that also contains the solving behavior of the reader. We think the solving pattern will be useful to assess more accurately the reader skill.

III. Proposed Method

We propose a method for estimating the TOEIC score of a subject. The estimation is based on the reading behavior and the answers of the MCQ. Figure 1 illustrates the principle of the process. A document is presented on a screen. This document contains a text and 4 questions which is a typical sample of TOEIC test. The subject is free to read the document

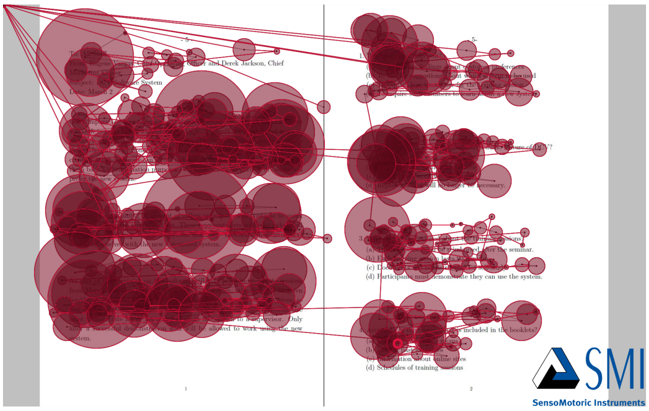


Fig. 2. Sequence of fixations and saccades recorded by the eye tracker. The circles represent the fixations and the lines represent the saccades. The diameter of a circle is proportional to the fixation time. When the reader blinks, the eye tracker cannot track the position of the eyes and output a gaze in the up-left corner.

in any order and to answers the questions when he wants. During this time, the eye gaze is recorded by an eye tracker. Then, based on the eye gaze recording and the answers to the questions, several features are computed and combined. After recording several subjects reading several documents, a backward stepwise feature selection is applied to find the best global features. Then, the TOEIC score of each subject is estimated based on a multivariate regression.

This principle is summarized in 4 steps:

1) Eye gaze feature extraction
2) MCQ feature extraction
3) Backward stepwise feature selection
4) Score estimation based on multivariate regression

Each step is detailed in the next subsections.

A. Eye gaze feature extraction

The eye tracker gives the information about when and where a user is looking at, while reading a text. An example of an output given by the eye tracker is displayed in Fig. 2. It consists of a sequence of eye gazes associated with timestamps representing the movement of the reader's eyes. The eye tracker also gives the information when the reader blinks. The fixations, saccades and blinks are extracted by the software BeGaze[1] from SMI company.

33 features are extracted from the reading behavior. These features are detailed in the table I, they are the same as the ones used by Yoshimura et al. [19]. These features represent mainly statistics about fixations, saccades and blinks such as blink duration, fixation dispersion, saccade amplitudes, etc.

TABLE I
FEATURES REPRESENTING THE READING BEHAVIOR. THESE FEATURES ARE COMPUTED FOR ONE SUBJECT READING ONE DOCUMENT.

Feature		Explanation
End_Time[ms]		elapsed time
Blink_Count		the number of blinks
Blink_Frequency[count/s]		the number of blinks per second
Blink_Duration	_Total	sum, average, maximum and minimum of the time of blinks
[ms]	_Average	
	_Maximum	
	_Minimum	
Fixation_Count		the number of fixation
Fixation_Frequency[count/s]		the number of fixation per second
Fixation_Duration	_Total	sum, average, maximum and minimum of duration of fixations
[ms]	_Average	
	_Maximum	
	_Minimum	
Fixation_Dispersion	_Total	sum, average, maximum and minimum dispersion of fixations
[px]	_Average	
	_Maximum	
	_Minimum	
Scanpath_Length[px]		sum of the distance saccade
Saccade_Count		the number of saccades
Saccade_Frequency[count/s]		the number of saccades per second
Saccade_Duration	_Total	sum, average, maximum and minimum of duration of saccades
[ms]	_Average	
	_Maximum	
	_Minimum	
Saccade_Amplitude	_Total	sum, average, maximum and minimum of the rotation angle of an eye ball during saccades
[°]	_Average	
	_Maximum	
	_Minimum	
Saccade_Velocity	_Total	sum, average, maximum and minimum of angular velocity of an eye ball during saccades
[°/s]	_Average	
	_Maximum	
	_Minimum	

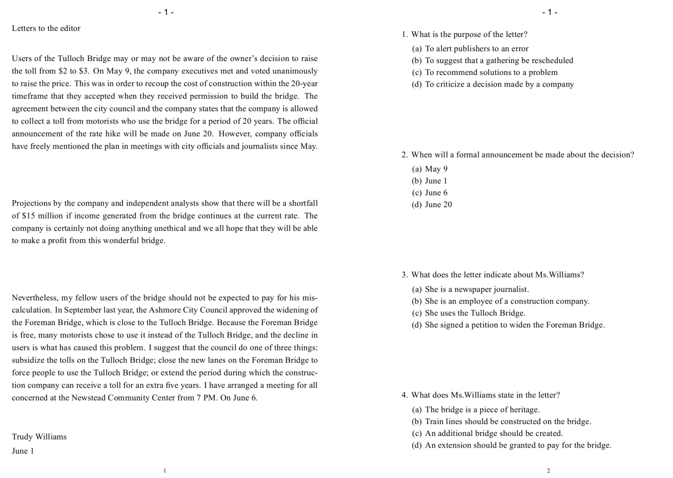
B. MCQ feature extraction

Each document contains 4 MCQ. For each MCQ, the reader needs to select one answer among four. The number of correct answers is used as one feature. The feature vector describing the reading behavior of one reader for one document is then constituted of 34 dimensions (33 about the eye gaze and one about the number of correct answers).

C. Backward stepwise feature selection

Among all proposed features, some are less relevant than others. In order to optimize the performance of the score estimation by using only the relevant features, we apply a backward stepwise selection [4]. This is a greedy algorithm consisting in computing the performance of the system by removing the worst feature at each round. First, all features are used and the performance of the system is computed based on the multivariate regression (detailed in the next section). Then, we compare this performance with the performance obtained by each combination of all features except one. The feature which produced the worst performance is removed. The process is repeated until the performance is not improved anymore. The remaining features are, by extent, the "best" features or the "selected" features.

[1]http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/begaze-analysis-software.html



- 1 -

Letters to the editor

Users of the Tulloch Bridge may or may not be aware of the owner's decision to raise the toll from $2 to $3. On May 9, the company executives met and voted unanimously to raise the price. This was in order to recoup the cost of construction within the 20-year timeframe that they accepted when they received permission to build the bridge. The agreement between the city council and the company states that the company is allowed to collect a toll from motorists who use the bridge for a period of 20 years. The official announcement of the rate hike will be made on June 20. However, company officials have freely mentioned the plan in meetings with city officials and journalists since May.

Projections by the company and independent analysts show that there will be a shortfall of $15 million if income generated from the bridge continues at the current rate. The company is certainly not doing anything unethical and we all hope that they will be able to make a profit from this wonderful bridge.

Nevertheless, my fellow users of the bridge should not be expected to pay for his miscalculation. In September last year, the Ashmore City Council approved the widening of the Foreman Bridge, which is close to the Tulloch Bridge. Because the Foreman Bridge is free, many motorists chose to use it instead of the Tulloch Bridge, and the decline in users is what has caused this problem. I suggest that the council do one of three things: subsidize the tolls on the Tulloch Bridge; close the new lanes on the Foreman Bridge to force people to use the Tulloch Bridge; or extend the period during which the construction company can receive a toll for an extra five years. I have arranged a meeting for all concerned at the Newstead Community Center from 7 PM. On June 6.

Trudy Williams

June 1

1

- 1 -

1. What is the purpose of the letter?
   (a) To alert publishers to an error
   (b) To suggest that a gathering be rescheduled
   (c) To recommend solutions to a problem
   (d) To criticize a decision made by a company

2. When will a formal announcement be made about the decision?
   (a) May 9
   (b) June 1
   (c) June 6
   (d) June 20

3. What does the letter indicate about Ms.Williams?
   (a) She is a newspaper journalist.
   (b) She is an employee of a construction company.
   (c) She uses the Tulloch Bridge.
   (d) She signed a petition to widen the Foreman Bridge.

4. What does Ms.Williams state in the letter?
   (a) The bridge is a piece of heritage.
   (b) Train lines should be constructed on the bridge.
   (c) An additional bridge should be created.
   (d) An extension should be granted to pay for the bridge.

2

Fig. 3. One of the 15 documents used for the experiment.

As we want our algorithm to be user independent, the feature selection is not applied for each subject independently but for all subjects.

D. Score estimation based on multivariate regression

The TOEIC score of each subject is known as a ground truth. The multivariate regression model represents this score as a function of 34 features:

$$y_i \quad = \quad \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \tag{1}$$

where y_i is the subject estimated TOEIC score of the subject i, β is a parameter vector, x the feature matrix, and ε_i the disturbance term.

Let S be the number of subjects and D the number of document, $R = S \times D$ is the number of recordings. Before applying the multivariate regression, each feature has R dimensions and is normalized in the range [-1,1]. Then, each recording vector composed of 34 features is scaled to the unit length (divided by the vector norm). For each reader, we compute the average value of each feature among the different documents.

The performance of the regression is then computed as the mean absolute error of the score estimation:

$$Error \quad = \quad \frac{1}{S} \Sigma_{i=1}^{S} |\dot{y_i} - y_i|, \tag{2}$$

where $\dot{y_i}$ is the subject real TOEIC score.

IV. Experiments

For the experiment, 21 subjects have been asked to read 15 documents. Each document is a sample of the reading section of the TOEIC test (part7). An example of document is displayed on the Fig. 3. This document is a double page. The left page contains a text and the right page contains 4 MCQ (with 4 possible answers) related to the text.

A. Procedure of recording

The reading behavior is recorded by using the stationary eye tracker SMI RED250 [2]. The sampling rate is 250 Hz, the gaze position accuracy is 0.4°and the spatial resolution is 0.03°.

No bite bar nor head fixation have been used, but the subject has been asked to refrain from doing large head movement during the recording. The answer of the questions were given verbally by the subject to the experimenter, also to limit the head movement of the subject.

Before reading a document, the calibration of the eye tracker is done by looking at 5 points. The experimenter check the accuracy of the eye tracking and proceed to the experience if it seems correct. During the experiment, the subject can read freely the text and questions in any order. He does not necessarily need to read all parts of the text.

Based on the features extracted from the eye movements and the number of correct answers, the TOEIC score of each subject is estimated. We compute the absolute difference between the real score and the estimated one. Then, by averaging the absolute error among all subjects, the performance of the system is obtained.

B. Information about the subjects

21 subjects participated to our experiments. The average age of the subject is 23 years old (from 21 to 25 year old with a standard deviation of 0.94). The subjects are mainly males, only 1 subject is a female. 8 subjects are undergraduate students and 13 master students. All subjects are Japanese. 15 subjects use contact lenses, 6 subjects do not need any correction (and no subject wear glasses).

The average of the TOEIC score of the subjects is 681.4 points. The maximum score of the subjects is 945 points, the minimum is 390 points and the standard deviation is 157.4 points.

The baseline regression performance is computed as the mean absolute error if all decisions were equal to the mean value (as a comparison, the baseline classification performance is obtain by always classifying to the largest class). In our case, the mean absolute error of the baseline regression is 137.2 points.

C. TOEIC score estimation

During the experiment, the subjects read and answer the questions of 15 documents. Doing the test with one document takes, in average 7 minutes (from 5 to 10 minutes). So, if we use 10 documents, around 70 minutes are needed, which is similar to the 75 minutes needed by the reading comprehension part of the TOEIC test. We want to reduce this time as much as possible while keeping a high accuracy of the estimation.

For estimating the score by reading N (with $N < 11$) documents, we took randomly 100 combinations of N documents among the 15 ones. In this way, the experiment is not dependent to a specific document for the estimation.

[2]http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/red250-red-500.html

For each combination of documents, we applied leave one out subject cross validation: one subject is used for testing and all other ones for learning (i.e. creating the multivariate regression model). We compute the performance of the system for this subject and then loop among all other subjects. The average performance for all subjects is calculated. Finally, we compute the average performance of the system for the 15 subjects and the 100 combinations of N documents.

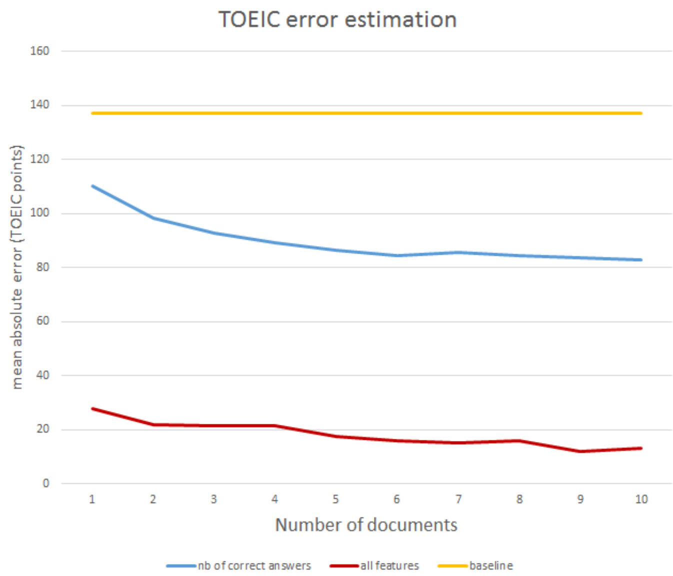


Fig. 4. Influence of the number of documents on the system performances. Using only the number of correct answers give a better estimation of the TOEIC score compare to the baseline. But using all the eye gaze features give a more accurate results.

We tested our algorithm by using 1 to 10 documents to estimate the TOEIC score. Figure 4 shows the details of the performance. By using only one document, the mean absolute error estimation is 27.7 points. By using 3 documents, the mean absolute error is reduced to 21.7 points. For 10 documents, the mean absolute error is 13.1 points. Compared to the mean absolute of the baseline regression (137.2 points), these estimations are quite accurate.

We can see that by using more documents, the estimation is getting more accurate. We remind that the TOEIC score is between 10 and 990, so our estimation by using only one document is accurate enough to estimate the skill of the reader.

These results are averages based on 100 combinations of N documents. It means that some combinations have higher accuracy, and some other lower accuracy. As we do not know yet how to evaluate what is a “good” document for estimating the TOEIC score, we prefer to try different combinations and to simply give the results as an average score.

For example, in the case of the estimation based on 10 documents, the best mean absolute error obtained was 0.43 points and the worst was 109 points (the average is 14.4 points and standard deviation 14.5 points). As the standard deviation is low, we can conclude that most of the combinations produced similar results which are accurate. Unfortunately few cases produce low performances, and we do not know why it happens with this specific combinations as the same documents are also used in other combinations and providing accurate results.

D. Feature selection

TABLE II
RANKING OF THE FEATURES FROM THE MOST IMPORTANT TO THE LEAST IMPORTANT. THE FEATURE SELECTION HAS BEEN APPLIED 100 TIMES WITH 10 DOCUMENTS SELECTED RANDOMLY. THE PERCENTAGE INDICATE HOW MANY TIMES THE FEATURE HAVE BEEN SELECTED.

Feature	Used time (%)
End Time [ms]	97
Saccade Velocity Total [°/s]	92
Saccade Count	89
Blink Count	89
Scanpath Length [px]	87
Saccade Velocity Maximum [°/s]	84
Fixation Dispersion Maximum [px]	83
Saccade Latency Average [ms]	83
Blink Duration Maximum [ms]	82
Fixation Count	80
Saccade Amplitude Total [°]	80
Saccade Amplitude Maximum [°]	80
Fixation Duration Total [ms]	78
Fixation Duration Maximum [ms]	78
Saccade Duration Total [ms]	77
Blink Duration Average [ms]	77
Saccade Duration Maximum [ms]	75
Fixation Duration Average [ms]	73
Blink Duration Total [ms]	73
Fixation Dispersion Total [px]	71
Blink Duration Minimum [ms]	70
Fixation Dispersion Average [px]	68
Saccade Velocity Average [°/s]	57
Fixation Dispersion Minimum [px]	53
Saccade Duration Average [ms]	41
Saccade Frequency [count/s]	39
Number of correct answers	36
Saccade Velocity Minimum [°/s]	36
Fixation Duration Minimum [ms]	34
Fixation Frequency [count/s]	33
Saccade Duration Minimum [ms]	30
Saccade Amplitude Minimum [°]	29
Saccade Amplitude Average [°]	28
Blink Frequency [count/s]	28

In the previous subsection we explained that for estimating the TOEIC score for N documents, we randomly selected 100 combinations of N documents among 15. Consequently, the feature selection have also been applied 100 times.

The table II shows how many times each feature have been selected by the backward stepwise feature selection for the estimation of the TOEIC score (in the case of using $N = 10$ documents). The table shows that half of the features are used 75% of the time, and almost all features are used at least 30% of the time. In average, 25 features are selected.

The feature about the number of correct answers to the MCQ is selected only 36% of time. This feature is not extracted from the reading behavior, so it is important to notice that the system could be improved to perform well without this feature. The most important features are: the time spend to read the document, the saccades velocity and count, and the blink count.

We notice that several features associated to the saccades, and especially the ones associated to the minimum values, tends to have a low rank. This minimum values are quite small for all the subjects and are not discriminative enough.

V. Conclusion and future work

Our aim is to develop an algorithm that can estimate automatically, quickly and accurately the language skill of a reader.

We proposed a method to estimate the TOEIC score based on the reading analysis and some questions. By applying a multivariate regression and backward stepwise feature selection, the score is estimated and the performance of the system computed. After reading 3 documents and answering the related questions, we showed that the TOEIC score can be estimated with a high accuracy as the error is 21.7 points, for 21 subjects.

This work could be useful for millions of people, and several aspects can be improved. The eye tracker used in our experiments is a professional eye tracker. We want to apply our method with an affordable eye tracker in order to make our technology usable for many more people. By example, an eye tracker such as Tobii Eye X is available for 119 euros [3]. We already started to use this eye tracker for analyzing the reading behavior [6], [11].

This research could also leads to new research about the importance of the documents for estimating the reading skill. Depending on the skill of the reader, different set of documents should be used. We can also see that, in a general way, some documents are more efficient than other documents to assess the reader's skill.

Acknowledgment

This work is supported in part by JST CREST and JSPS KAKENHI Grant Numbers 25240028 and 15K12172.

References

[1] Olivier Augereau, Koichi Kise, and Kensuke Hoshika. A proposal of a document image reading-life log based on document image retrieval and eyetracking. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 246–250. IEEE, 2015.

[2] Kiyomi Chujo and Kathryn Oghigian. How many words do you need to know to understand toeic, toefl & eiken? an examination of text coverage and high frequency vocabulary. *Journal of Asia TEFL*, 6(2):121–148, 2009.

[3] Leana Copeland, Tom Gedeon, and Sumudu Mendis. Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artificial Intelligence Research*, 3(3):p35, 2014.

[4] George H John, Ron Kohavi, Karl Pfleger, et al. Irrelevant features and the subset selection problem. In *Machine learning: proceedings of the eleventh international conference*, pages 121–129, 1994.

[5] Haijun Kang. Understanding online reading through the eyes of first and second language readers: An exploratory study. *Computers & Education*, 73:1–8, 2014.

[3] http://www.tobii.com/xperience/

[6] Riki Kudo, Olivier Augereau, Takuto Rou, and Koichi Kise. Reading similarity measure based on comparison of fixation sequences. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 1221–1226. ACM, 2015.

[7] Kai Kunze, Hitoshi Kawaichi, Kazuyo Yoshimura, and Koichi Kise. Towards inferring language expertise using eye tracking. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 217–222. ACM, 2013.

[8] Kai Kunze, Hitoshi Kawaichi, Kazuyo Yoshimura, and Koichi Kise. The wordometer–estimating the number of words read using document image retrieval and mobile eye tracking. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 25–29. IEEE, 2013.

[9] Kai Kunze, Susana Sanchez, Tilman Dingler, Olivier Augereau, Koichi Kise, Masahiko Inami, and Terada Tsutomu. The augmented narrative: toward estimating reader engagement. In *Proceedings of the 6th Augmented Human International Conference*, pages 163–164. ACM, 2015.

[10] Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise, and Andreas Bulling. I know what you are reading: recognition of document types using mobile eye tracking. In *Proceedings of the 17th annual international symposium on International symposium on wearable computers*, pages 113–116. ACM, 2013.

[11] Charles Lima Sanches, Olivier Augereau, and Koichi Kise. Eye gaze and text line matching for reading analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 1227–1233. ACM, 2015.

[12] Pascual Martínez-Gómez and Akiko Aizawa. Recognition of understanding level and language skill using measurements of reading behavior. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 95–104. ACM, 2014.

[13] Ayano Okoso, Kai Kunze, and Koichi Kise. Implicit gaze based annotations to support second language learning. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 143–146. ACM, 2014.

[14] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422, 1998.

[15] Keith Rayner, Kathryn H Chace, Timothy J Slattery, and Jane Ashby. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, 10(3):241–255, 2006.

[16] TOEIC. 2014 report on test takers worldwide: The toeic listening and reading test. http://www.ets.org/s/toeic/pdf/ww_data_report_unlweb.pdf. Accessed: 2016-04-04.

[17] TOEIC. Toeic program data & analysis 2014. http://www.toeic.or.jp/library/toeic_data/toeic_en/pdf/data/TOEIC_Program_DAA.pdf. Accessed: 2016-04-04.

[18] Erroll Wood and Andreas Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 207–210. ACM, 2014.

[19] Kazuyo Yoshimura, Kai Kunze, and Koichi Kise. The eye as the window of the language ability: Estimation of english skills by analyzing eye movement while reading documents. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 251–255, 2015.