Analyzing Eye-Voice Coordination in Rapid Automatized Naming

Daniel Bone[1], *Chi-Chun Lee*[1], *Vikram Ramanarayanan*[1],
Shrikanth Narayanan[1], *Renske S. Hoedemaker*[2], *Peter C. Gordon*[2]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA
[2]Department of Psychology, University of North Carolina, Chapel Hill, NC
http://sail.usc.edu

Abstract

Rapid Automatized Naming (RAN) is a powerful tool for predicting future reading skill. A person's ability to quickly name symbols as they scan a table is related to higher-level reading proficiency in adults and is predictive of future literacy gains in children. However, noticeable differences are present in the strategies or patterns within groups having similar task completion times. Thus, a further stratification of RAN dynamics may lead to better characterization and later intervention to support reading skill acquisition. In this work, we analyze the dynamics of the eyes, voice, and the coordination between the two during performance. It is shown that fast performers are more similar to each other than to slow performers in their patterns, but not vice versa. Further insights are provided about the patterns of more proficient subjects. For instance, fast performers tended to exhibit smoother behavior contours, suggesting a more stable perception-production process.

Index Terms: Rapid Automatized Naming (RAN), reading strategies, eye-tracking, multi-modal, production-perception

1. Introduction

In the Rapid Automatized Naming (RAN) task, a participant is presented with a grid of 36 familiar stimuli (drawn from sets of six letters, numbers, colors, or objects) and must name them aloud as quickly and accurately as possible in order of appearance [1]. RAN performance is used to diagnose childrens reading disorders [2], predict their future literacy gains [3], and to characterize reading ability in adults [4, 5].

The predictive power of the RAN task likely results from its similarity to reading in terms of the demands associated with sequential processing of simultaneously presented stimuli, as there is little to no relationship between single item naming and reading skill [6, 7]. RAN performance depends heavily on sustained attention as shown by its relation to performance on other attention-demanding tasks [5]. In the RAN task, sustained attention is required in order to closely coordinate eye-movements, perceptual encoding, working memory, lexical processing and vocal execution, much like reading.

In order to control working memory load while optimizing overall speed, the encoding of each item in the array must be timed closely with its vocal response. Perceptual encoding of an item requires less time than its vocalization, allowing eye fixations to move ahead of the vocal sequence across items. More importantly, different stages of processing (encoding, lexical retrieval, vocalization) may take place in parallel across two or more sequential items, so that an efficient strategy involves the eyes leading the voice. However, this eye-voice discrepancy is limited by working memory and attentional demands resulting from simultaneously processing multiple items. As such, fast RAN performance depends on the control of eye movements in coordination with the voice stream in a way that does not overload working memory.

To date, the only metric of RAN performance is total completion time. However, this is a gross measure that does not detail events that led to the outcome. Specification of the precise skills supporting RAN performance may contribute to our understanding of causal factors of individual differences in reading achievement and ability, allowing for tailored intervention. Characterization of the patterns of eye-voice coordination that constitute optimal RAN performance requires joint analysis of both data streams.

Contributions of this work in the context of related prior work lie in three facets. First, little or no work attempts to model patterns of the RAN task, likely due to the complex nature of the perception-production dynamics. Although theoretical models of reading processes exist [8], they were not designed for RAN. A primary contribution of this novel work is to formulate the RAN dynamics such that signal processing can be employed. Second, other work has applied machine learning to eye-tracking data, for example to locate areas of interest [9] or to classify user intent [10]. In contrast, the current task does not use absolute duration features and has different task goals, and thus requires different methodology. Finally, this analysis complements other studies of human behavior [11, 12, 13]; like some such studies [14], we plan to examine cognitive processes of children with autism using collected RAN data, now that methods to establish normal patterns have begun to be explored.

2. Experimental Design

We investigate patterns in eye-tracking and voice signals of subjects performing Rapid Automatized Naming (RAN) subtasks, in an effort to understand how channel coordination effects completion time.

This work was supported by funds from NSF and NICHD R01 HD060440-06A2.

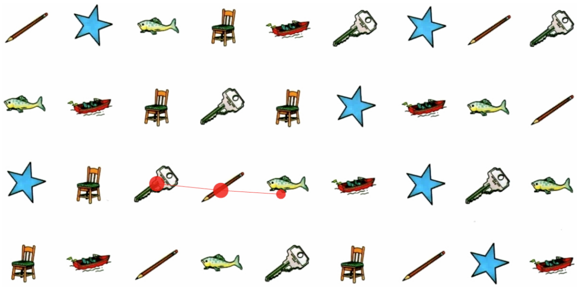

Figure 1: ‘Objects’ prompt and sample eye-tracking. Eye-tracking is visualized as red circles with diameter representative of fixation duration, and by connecting lines representing saccades between fixations.

2.1. Corpus

The UNC RAN Corpus consists of 30 performers in 4 RAN subtasks: Objects, Colors, Numbers, and Letters. Each subtask is performed twice. We focus our study on the 22 subjects for whom we have eye-tracking and vocal data for all 8 trials.

Each RAN task is comprised of 36 symbols, arranged in 4 rows and 9 columns as in Figure 1. Eye-tracking data is collected using an Eye Link 1000 by SR Research. The output of the eye-tracking is a set of fixation (x,y) positions and durations. The fixated symbols are then automatically determined from the (x,y) position with algorithms designed to lessen fixations on one row being mis-assigned to an adjacent row.

The subject reads aloud while parsing the list of symbols left-to-right and top-to-bottom. The start and end times of each vocalization are recorded automatically using forced aligment, and then manually corrected. In Figure 2, the fixated word and vocalized word are plotted over time. We can see that the eyes almost always lead the voice, but the coordination between perception and production is variable.

We remove the first word in the list because the performer is often in a very unstable state. In the rare case that fixation data are absent due to possible collection error, we use KNN imputation to complete the data.

2.2. Eye-Tracking and Voice Signals

The temporal patterning of both the visual and voice data streams, as well as the relationship between them (eye-voice lead) are modeled. The eye and voice signals in Figure 2 can be considered to represent the observable state sequence of a finite state machine. These state sequence observations can be approximately summarized by reformulating the signal such that each item, or symbol in order, is considered a point in time and the value corresponds to a duration for that item. We consider three rich feature contours: the total fixation duration on each item; the total pause+vocalization duration for each item; and finally eye-voice lead, the duration between when the eye first observes an item and when that item is spoken.

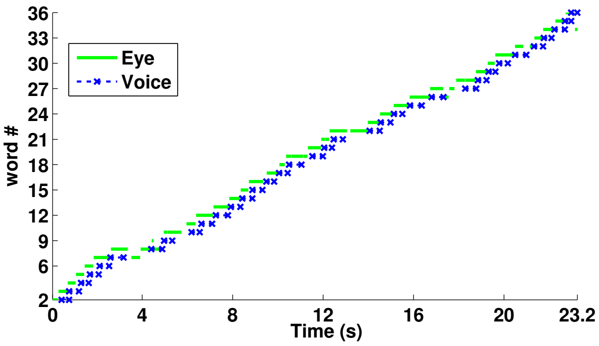

Figure 2: Example eye-voice trace with eye fixations in green and vocal productions in blue for each word.

It is critical that the descriptors of each RAN trial are decorrelated with the trial’s total completion time. Many of the absolute values of these state-sequence-based signals are initially correlated with that same completion time. For example, mean item pause+vocalization duration and mean total fixation duration on an item are, as expected, highly correlated with the total completion time of a list. Thus, the particular solution we employ is to normalize each feature contour to have unit magnitude. This is equivalent to seeking the percentage contribution of each item’s value towards creating the cumulative observation– e.g., the normalized pause+vocal duration feature contour represents the *percentage* of time spent pausing-before and speaking each symbol. One aspect of this approach is that regressions are not explicitly modeled; however, regressions are implicitly modeled since they will contribute to the recorded duration of the word from which the regression happens.

Normalized feature contours are shown in Figure 3. The greatest variability often occurs near a line change, except for the eye duration signal which is

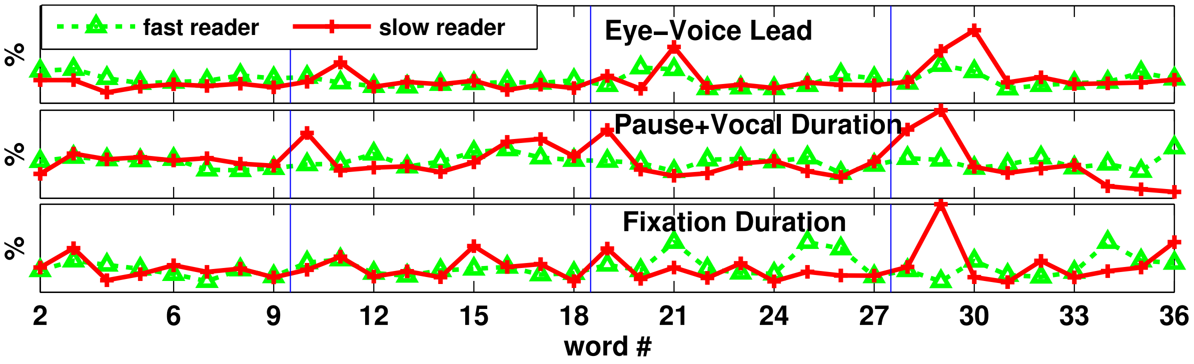

Figure 3: Signals from two subjects performing Numbers subtask. The vertical blue lines designate ends of rows.

less consistent in this trend. Faster performers tend to have smoother feature contours that vary less and more slowly.

3. Analysis of Eye and Voice Patterns

Fast speakers may exhibit behaviors that further stratify effective automatized naming skills. The following experiments investigate these potential patterns by quantifying interpretable trends in the eye and voice signals. In Section 3.1 we find support for the existence of common behaviors of fast performers; in Section 3.2 signal smoothness is quantified; and in Section 3.3 we investigate item-level behaviors.

3.1. Existence of Performance Patterns

Proficient performers may share certain qualities that can be observed in the dynamic patternings of the eye-voice lead, pause+vocal duration, and total fixation duration signals. Methods presented in this section can evaluate the potential similarities of performers with comparable speeds.

Exemplar-based templates can be generated per task using the fastest (or slowest) subjects that may show how the fastest (or slowest) subjects process each task. Templates are generated by taking the fastest (or slowest) performers feature contours and averaging– thus we are searching for a single common component in the patterns to support the utility of further analyses. As a measure of similarity, we use the L-1 norm between the template and the patterns of each remaining performer. These distances can then be compared to the completion time using Spearman's-rank correlation (Table 1).

Table 1: *Distance from exemplar-based templates as a measure of a subject's performance style, using Spearman's correlation. Bold and $*$ indicate significance at α=0.05 and α=0.10 levels, respectively.*

Category	Object		Color		Number		Letter	
Modeling Each Subtask Separately- Fast Performer Exemplars								
Eye-Voice	**.35**		**.50**		.30*		.15	
Pause+Vocal	**.42**		**.49**		.13		.10	
Fixation	.27*		0.05		.13		.14	
Modeling Each Subtask Separately- Slow Performer Exemplars								
Eye-Voice	.12		-.08		.11		.12	
Modeling Each Trial Separately- Fast Performer Exemplars								
Trial	1	2	1	2	1	2	1	2
Eye-Voice	**.47**	**.50**	.17	**.66**	.36	.03	.38	-.03

First, analysis is conducted on the signal patterns that occur without regard to the symbol-order, since both item-randomized trials are included in the same analysis. The four fastest trials are used to generate a template for each subtask. One interesting finding is that there are many significant and marginally-significant correlations that occur in the hypothesized direction; fast performers are more like other fast subjects than are slow subjects, even in these signals that which total duration information removed. Second, the eye-voice lead and pause+vocalization duration feature contours lead to higher correlations than the eye duration feature contour. Another observation is that the correlations are stronger for Objects and Colors subtasks. Objects and Colors subtasks are often analyzed separately from Numbers and Letters subtasks because the cognitive load is much lower for the latter (i.e., completion times are nearly half those of the former).

Next, we explore whether slow performers tend to display a common pattern and strategy or if there are varied, idiosyncratic processing difficulties among slow speakers. For the eye-voice feature contour– which provided the most significant results for fast-performer templates– slow subjects are not found to be more closely related to slow-performer exemplars, than fast-performers. This is tentative evidence that slow performers have varied factors contributing to reduced performance.

Lastly, we consider if incorporating the symbol-ID by examining trials individually will improve the correlations over subtask-level analyses. No consistent improvement is observed in the correlation values displayed in the last row of Table 1 when using the three fastest performers as exemplars. This may indicate that the actual symbol is not a significant factor, but more critical is the position of the item in the 2-D grid. For example, difficulty may arise near row boundaries.

3.2. Pattern Stability

Having established that some useful information is present in the dynamic patterns of these contours, it is desirable to further quantify such patterns. Speech from faster performers seems confident and their pace appears steady. In this sub-section, an informative measure that quantifies the potential lack of such temporal predictability in slower performers is examined.

Our method to measure the roughness (lack of smoothness) in a signal is to compute the standard deviation of the first-order difference (Δ) of the contour; the difference operator is a high-pass filter. We hypothesized that faster readers would show more stability, which should result in less variable, more stationary contours. In our case, the signals are not varying over actual time, but over item number (2-36).

Table 2: *Spearman's correlations between 'roughness' and total completion time. Bold indicates significance at α=0.05.*

Category	Object	Color	Number	Letter	All
Eye-Voice	**.59**	**.40**	**.36**	**.42**	**.44**
Pause+Vocal	**.49**	**.49**	.19	**.33**	**.35**
Fixation	**.31**	.07	.20	.27	**.21**

Results of correlating the standard deviation of the difference signal with the completion time (shown in Table 2) indicate a relationship between contour roughness and task speed. This is especially bourne out from the eye-voice lead signal measurements– the less predictable the item-to-item changes in eye-voice lead time, the slower a subject tends to be. One possible explanation is that slower performers have difficulty coordinating the internal flow of information between the visual and vocal channels. However, we performed one more test in which we computed a second feature, the standard deviation of the contours without the difference operator. We find that the two features are highly correlated (Spearman's ρ_{min}=0.87, ρ_{max}=0.98). This indicates that the variability we see in the signal is spurious, high-frequency variability and that the contours are very nonstationary. In essence, lack of a flat contour is equivalent to lack of a smooth contour, and both imply lack of steadiness. Beyond simply quantifying the mean rate of observable signals, we have found explanatory power in the rhythm of the observable signals.

3.3. Location-Dependence of Channel Coordination

As previously observed, item-position is a critical factor in determining how a subject controls the perception-production mechanisms. This effect is especially evident at the end and beginning of a row, where a slower subject may struggle because of the long saccade required to start at the next row as shown in Figure 3.

The relative percentage of time each of the three signals (eye-voice lead, pause+vocal duration, and total fixation duration) attribute to each item may indicate deviations in strategies employed by subjects of differing abilities (Figure 4). Spearman's rank-correlation between these percentages and completion time reveals a consistent trend near row boundaries for the eye-voice lead signal– shown in Table 3.

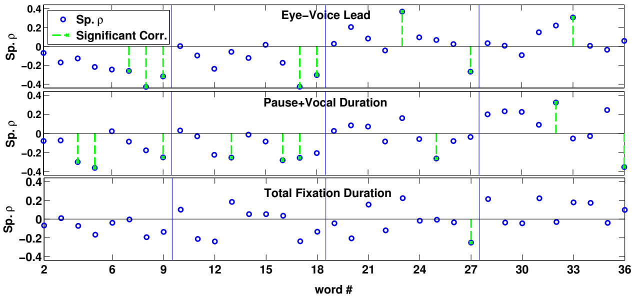

Figure 4: Correlations per-item for the Numbers task. Dotted lines indicate significance at α=0.10 level.

Table 3: *Eye-voice trends at end of rows. Bold and* $*$ *indicate significance at* α=0.05 *and* α=0.10 *levels, respectively.*

Category	Object	Color	Number	Letter
Row 1	-.22	-.27*	**-.32**	-.28*
Row 2	-.04	**-.38***	**-.31**	-.27*
Row 3	-.06	-.08	-.27*	.05
Row 4	-.14	**-.38**	.06	-.07

An interesting finding is that for faster performers, the amount of relative eye-voice lead for the last item in a row is higher than for slower performers. Eye-voice lead may increase for several reasons. For example, it may be larger if a person is storing more symbols in their decoded-symbol buffer which is only necessary if the person also decoded quickly; on the contrary, it is also larger if the person is pausing longer before saying a word. The eye-voice lead time tends to decrease for fast and slow speakers within a row, and often appears to reach a minimum at the end of the row. This may imply that subjects are approaching an equilibrium or peak speed. The last item in the row is the location most consistently observed for the trends under focus. It should be noted that look-ahead is typically 1-2 symbols.

Taking all factors into account, it is possible that the faster subjects are maintaining their buffer of visual information more effectively near the end of a line, and thus are better equipped to maintain pace after changing rows. Although these results are inconclusive, the evidence is consistent enough such that it may inspire further studies to understand this process more clearly.

4. Classification Task and Discussion

Measures relating to the common pattern between fast performers (3 features), to the smoothness (or lack thereof) of three contours (3 features), and to the relative row-end eye-voice lead (4 features) hold information about the total completion time for the subject, even though these measures have been normalized for total time. To further demonstrate that the created features are informative, we perform binary classification of reading speed as determined by the median speed per subtask. Since some trials are required for fast-performer templates, we classify the remaining 40 trials per subtask. Classification is performed using linear SVM [15]. The 10 features are further reduced by feature-selection with Fisher scoring [16]. Results are presented in Table 4.

Table 4: *Classification of performers grouped by speed.*

Category	Object	Color	Number	Letter	All
Accr.	60%	68%	65%	60%	63%

Classification accuracies of 60-68% are achieved. 63% classification accuracy on the 160 samples is significantly above chance ($p<0.01$). The accuracies are also consistent across tasks, even for the tasks with lower cognitive-load. The accuracy is as-expected based on the 0.4 correlations for many features, and still interesting given we are only considering RAN dynamics, which have not yet been studied.

5. Conclusion

In this work, we modeled the dynamics of the eyes, voice, and the interaction between the two during performance of a task that is strongly related to reading ability. We find fast performers are more similar to each other than to slow performers in their patterns, but not the other way around. With this supporting evidence, we sought to quantify the patterns of an efficient performer.

A major finding is that faster subjects show smoother patterns in all three signals than slower subjects. One possible explanation is that slower performers have difficulty coordinating the internal flow of information between the visual and vocal channels, and have more difficulty than faster performers in increasing their look-ahead near line-end in preparation for a line switch. Validation of the proposed measures was obtained through classification tasks.

Thus, beyond simply quantifying the mean rate of observable signals, we found explanatory power in the rhythm of the observable signals, as well as other descriptors. Such insights may lead to better stratification of optimal reading strategies, and thus to tailored intervention.

We also observed that some speakers employ chunking strategies; in particular, they perceive and produce 3-4 words at a time. In the future, the observed local "rhythm" differences will be explored. Additionally, behaviors from a large corpus of children with autism spectrum disorders (ASD) will be analyzed.

6. References

[1] M. B. Denckla and R. Rudel, "Rapid automatized naming of pictures objects, colors, letters and numbers by normal children.," *Cortex*, vol. 10, pp. 186–202, 1974.

[2] P.G. Bowers, "Tracing symbol naming speeds unique contribution to reading disabilities over time.," *Reading and Writing*, vol. 7, pp. 189–216, 1995.

[3] National Early Literacy Panel, *Developing Early Literacy: Report of the National Early Literacy Panel. Executive Summary.*, Washington, DC: National Institute for Literacy., 2008.

[4] H. L. Swanson, G. Trainin, D. M. Necochea, and D. D. Hammill, "Rapid naming, phonological awareness and reading: A meta-analysis of the correlation evidence.," *Review of Educational Research*, vol. 72, pp. 407–440, 2003.

[5] K. M. Arnell, R. Klein, M. F. Joanisse, M. A. Busseri, and R. Tannock, "Decomposing the relation between rapid automatized naming (ran) and reading ability.," *Canadian Journal of Experimental Psychology*, vol. 63, pp. 173–184, 2009.

[6] C. A. Perfetti, E. Finger, and T W. Hogaboam, "Sources of vocalization latency differences between skilled and less skilled young readers," *Journal of Educational Psychology*, vol. 70, pp. 730–739, 1978.

[7] K. E. Stanovich, "Relationships between word decoding speed, general name-retrieval ability, and reading progress in first-grade children," *Journal of Educational Psychology*, vol. 73, pp. 809–815, 1981.

[8] Alexander Pollatsek, Erik D. Reichle, and Keith Rayner, "Tests of the ez reader model: Exploring the interface between cognition and eye-movement control," *Cognitive Psychology*, vol. 52.1, pp. 1–56, 2006.

[9] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005)*, Cambridge, MA, 2006, pp. 547–554, MIT Press.

[10] Kai Puolamki and Samuel Kaski, Eds., *Proceedings of the NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling*, 2006.

[11] Shrikanth S. Narayanan and Panayiotis G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, 2012.

[12] Matthew P. Black, Athanasios Katsamanis, Brian R. Baucom, Chi-Chun Lee, Adam C. Lammert, Andrew Christensen, Panayiotis G. Georgiou, and Shrikanth S. Narayanan, "Toward Automating a Human Behavioral Coding System for Married Couples' Interactions Using Speech Acoustic Features," *Speech Communication*, 2011, In Press.

[13] C.C. Lee, A. Katsamanis, M.P. Black, B.R. Baucom, P.G. Georgiou, and S.S. Narayanan, "An Analysis of PCA-based Vocal Entrainment Measures in Married Couples' Affective Spoken Interactions," in *Proceedings of Interspeech*, 2011.

[14] Daniel Bone, Matthew P. Black, Chi-Chun Lee, Marian E. Williams, Pat Levitt, Sungbok Lee, and Shrikanth Narayanan, "Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist," in *Proceedings of Interspeech*, 2012.

[15] Chang, Chih-Chung and Lin, Chih-Jen, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[16] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, John Wiley and Sons, New York, 2 edition, 2001.