Toward Exploiting EEG Input in a Reading Tutor

Jack Mostow, Kai-min Chang, and Jessica Nelson

Project LISTEN, School of Computer Science, RI-NSH 4103, 5000 Forbes Avenue,
Carnegie Mellon University, Pittsburgh, PA 15213, USA
mostow@cs.cmu.edu,
{kaimin.chang,jessica.nelson}@gmail.com

Abstract. A new type of sensor for students' mental states is a single-channel EEG headset simple enough to use in schools. Using its signal from adults and children reading text and isolated words, both aloud and silently, we train and test classifiers to tell easy from hard sentences, and to distinguish among easy words, hard words, pseudo-words, and unpronounceable strings. We also identify which EEG components appear sensitive to which lexical features. Better-than-chance performance shows promise for tutors to use EEG at school.

Keywords: EEG, reading tutor, power spectrum, frequency band, lexical feature.

1 Introduction

The ultimate automated tutor could peer directly into students' minds to identify their mental states (knowledge, thoughts, feelings, and so forth) and decide accordingly what and how to teach at each moment. The reality, of course, is that today's automated tutors attempt instead to infer students' mental states from a thin trickle of data, typically in the form of mouse clicks and keyboard input. Some ITS researchers (e.g. Anderson, Graesser, Picard, and Woolf, in too many papers to cite here) are exploring other types of data, such as speech, eye movements, posture, heart rate, skin conductance, and mouse pressure. This paper tests a complementary source of input from as close to the brain as non-invasively possible: electroencephalogram (EEG).

The EEG signal is a voltage signal that can be measured on the surface of the scalp, arising from large areas of coordinated neural activity. This neural activity varies as a function of development, mental state, and cognitive activity, and the EEG signal can measurably detect such variation. For example, rhythmic fluctuations in the EEG signal occur within several particular frequency bands, and the relative level of activity within each frequency band has been associated with brain states such as focused attentional processing, engagement, and frustration [1-3], which in turn are important for and predictive of learning [4].

The recent availability of simple, low-cost, portable EEG monitoring devices suddenly makes it feasible to take this technology from the lab into schools. The NeuroSky "MindSet," for example, is an audio headset equipped with a single-channel EEG sensor. It measures the voltage between an electrode that rests on the forehead and electrodes in contact with the ear. Unlike the multi-channel electrode nets worn in labs, the sensor requires no gel or saline for recording, and requires no

G. Biswas et al. (Eds.): AIED 2011, LNAI 6738, pp. 230–237, 2011.

expertise to wear. Even with the limitations of recording from only a single sensor and working with untrained users, the MindSet distinguished two fairly similar mental states (neutral and attentive) with 86% accuracy [5].

The ability to record longitudinal EEG data in authentic school settings is important for several reasons. First, we can analyze longer-term learning over intervals longer than a lab experiment, in contrast to short-term memory effects. Second, we can study data generated by children's *"in vivo"* behavior at school, rather than their more constrained behavior in unfamiliar lab settings under intense adult supervision. Third, we can get enough data over a long enough time from enough students to combat the notoriously noisy nature of EEG data with the statistical power of "big data," thereby enabling us to analyze the effects of different forms of instruction and practice on student learning and moment-to-moment engagement. Finally, longitudinal recording of EEG data on a school-based tutor offers the opportunity to make student-specific models actually useful, by obtaining enough data over time to train valid models, and applying them on enough occasions to pay off in better student learning.

To assess the feasibility of collecting useful information about cognitive processing and mental state using a portable EEG monitoring device, we conducted a pilot study in which participants wore a NeuroSky Mindset while using Project LISTEN's Reading Tutor [6]. The Reading Tutor displays text, listens to the student read aloud, and logs detailed longitudinal records of its multimodal tutorial dialogue to a database [7]. We linked this data to EEG data by user ID and timestamp.

We wanted to know if MindSet data can distinguish among mental states relevant to learning to read. More specifically:

1. Can EEG detect when reading is difficult? So we presented easy and hard text.
2. Can EEG detect lexical features? So we showed isolated words, varied by type.
3. What EEG components are sensitive, to what features? So we correlated them.

We used a within-subject design to compare the EEG signal during easy vs. difficult reading, at both the passage and single item level, during both oral and silent reading. Sections 2, 3, and 4 address questions 1-3; Section 5 concludes.

2 Can EEG Detect When Reading Is Difficult?

We implemented our experimental protocol in the Reading Tutor's homegrown language for scripting interactive activities. It displayed passage excerpts to read aloud, three easy and three hard, in alternating order. The "easy" passages were from texts classified by the Common Core Standards (www.corestandards.org) at the K-1 level. The "difficult" passages came from practice materials for the Graduate Record Exam (majortests.com/gre/reading_comprehension.php) and the ACE GED test (college.cengage.com:80/devenglish/resources/reading_ace/students). Each passage was followed by a multiple-choice cloze question (formed from the next sentence in the passage) to ensure that readers were reading for meaning. The protocol then repeated these tasks in a silent reading condition, using different text. Across the read-aloud and silent reading conditions, passages ranged from 62 to 83 words long.

10 adult readers participated in our lab, and 11 nine- and ten-year-olds at school. (A few other participants user-tested the protocol or had no EEG data.) We excluded

4 adults and 2 children due to missing or poor-quality data. We analyzed data for the remaining 6 adults and 9 children both separately and pooled across all 15 readers.

2.1 Training Procedure

We trained binary logistic regression classifiers to estimate the probability that a given sentence was easy (or hard), based on EEG data. We trained separate classifiers for each condition (oral and silent reading) and group (adults and children), and also classifiers for data pooled across both conditions and groups.

We trained and tested two types of classifiers for each classification task. We trained *reader-specific* classifiers on a single reader's data from all but one stimulus (passage or word), tested on the held-out stimulus, performed this procedure for each stimulus, and averaged the results to cross-validate accuracy within readers. For stimuli (e.g., passages) with multiple successive observations (e.g., sentences), cross-validating across stimuli avoids improperly exploiting statistical dependencies – such as temporal continuity – between observations of a reader on the same stimulus. We trained *reader-independent* classifiers on the data from all but one reader, tested on the held-out reader, performed this procedure for each reader, and averaged the resulting accuracies to cross-validate across readers.

As features for logistic regression we used the streams of values the MindSet logs:

1. The raw EEG signal, sampled at 512 Hz
2. A filtered version of the raw signal, also sampled at 512 Hz
3. Proprietary "attention" and "meditation" measures reported at 1 Hz
4. A power spectrum of 1Hz bands from 1-256Hz, reported at 8 Hz
5. An indicator of signal quality, reported at 1 Hz

We averaged measures 1-4 over the time interval of each stimulus, excluding the 15% of observations where measure 5 reported poor signal quality.

One problem in training classifiers is class size imbalance. We face this issue because we have more easy sentences than hard ones and more non-words than real words. A common solution is to resample the training data to obtain equal-size sets of training data. However, "random undersampling can potentially remove certain important examples, and random oversampling can lead to overfitting" [8]. To avoid bias due to class size imbalance, we employed three different resampling methods: random oversampling of the smaller class(es), with replacement; random undersampling of the larger class(es); and directed undersampling, in our case by truncating the larger class to the temporally earliest k examples. An adaptive tutor would use such temporal truncation to train user-specific models on each user's initial data. We show results for all three resampling methods.

We computed *classification accuracy* as the percentage of cases classified correctly; chance performance is one over the number of categories. To test whether a classifier was significantly better than chance, we first computed its overall accuracy for each reader, yielding a distribution of N accuracies, where N is the number of readers. Treating this distribution as a random value, we performed a one-tailed T-test of whether its mean exceeds chance performance for the classification task in question. Counting N readers rather than observations is conservative in that it accounts for statistical dependencies among observations from the same reader. Our significance criterion was $p < .05$, without correction for multiple comparisons.

2.2 Results

To find out if our data differed by population, grain size, or modality, we trained classifiers to distinguish between children vs. adults, words vs. sentences, and silent vs. oral reading. Children's and adults' data had no significant differences, but word and sentence reading differed sharply, as did silent and oral reading.

We trained classifiers to distinguish between easy and hard sentences read aloud, silently, or both, by adults, children, or both. Table 1 shows the results; values in **bold** here and later are significantly better than chance. Depending on the resampling method used, accuracy averaged from about 43% to 69% for reader-specific classifiers and 41% to 65% for reader-independent classifiers, respectively, suggesting that imperfect transfer across readers sometimes outweighs the advantage of training on more data; classification of fMRI brain images has a similar qualitative pattern [9]. Reader-specific classification of children's oral reading was especially good, which bodes well for detecting reading struggles in the Reading Tutor.

Table 1. Accuracy in classifying sentences from easy vs. hard text

		Reader-specific			Reader-independent		
	condition	**over-sample**	**under-sample**	**truncate**	**over-sample**	**under-sample**	**truncate**
adult	**oral**	0.49	0.56	0.53	**0.65**	0.54	0.41
	silent	0.44	0.43	0.56	**0.63**	0.54	0.54
	both	**0.53**	0.55	0.55	0.54	**0.56**	**0.54**
child	**oral**	0.62	0.62	**0.69**	0.59	**0.59**	**0.63**
	silent	0.47	0.46	0.45	0.50	0.52	0.48
	both	**0.64**	**0.59**	**0.65**	0.47	0.46	0.48
both	**oral**	0.57	**0.60**	**0.62**	0.52	0.52	0.53
	silent	0.49	0.57	0.50	0.53	**0.58**	0.50
	both	0.56	**0.61**	**0.60**	0.47	0.52	0.50

3 Can EEG Detect Lexical Features?

Besides text, our protocol displayed 10 words and 10 pseudo-words one at a time, ordered randomly, to read aloud. Words were all 2-syllable 7-letter words; half were easy and half were hard, to see if our data reflected difficulty in word reading; prior work [10] had found distinct EEG indicators of visual-spatial, orthographic, phonological, and semantic operations in reading. We included non-words to see if we could detect when readers saw unfamiliar words. The "easy" words had a Kucera-Francis (K-F) frequency of 30 or more (mean = 84) and an age of acquisition (AOA)

below 315 on a scale from 0-700 (mean = 254.4) [11]. The "hard" words had a K-F frequency below 10 (mean = 3.4) and an AOA above 450 (mean = 555.5). Pseudo-words were 3 letter pronounceable strings, chosen to vary in their number of orthographic neighbors (words that differ in spelling by only one letter), since EEG data (specifically, event related potentials) are sensitive to neighborhood size [12].

The isolated-item section also presented ten illegal 3-character strings to read silently, also with varying orthographic neighborhood sizes, also from the same study; the read-aloud condition omitted illegal strings because they are unpronounceable. We varied the orthographic neighborhood size of the pseudo-words and illegal strings from 0 neighbors to 22 neighbors, to enable (future) analysis of its effects.

We trained and evaluated classifiers just as described in Section 2.1, except that we trained multinomial logistic regression classifiers to estimate the probability that a word was easy, hard, a pseudo-word, or (in the silent condition) an illegal string. We evaluated their *rank accuracy* as the average percentile rank (normalized between 0 and 100) of the correct category if categories are ordered by the value of the regression formula; chance performance is 50%. Rank accuracy is a more sensitive criterion than classification accuracy for evaluating performance on multi-category tasks such as decoding mental states from brain data [9].

We expected it to be harder to distinguish among 3 or 4 kinds of isolated words and non-words than to tell easy from hard sentences, because reading an isolated word is so brief compared to reading a sentence. In addition, we had fewer samples of isolated words than sentences. Nevertheless, as Table 2 shows, rank accuracy averaged from about 45% to 58% for reader-specific classifiers, depending on the resampling method used, and about 39% to 59% for reader-independent classifiers.

Table 2. Rank accuracy (chance = 50%) in classifying words easy, hard, pseudo, or illegal

		Reader-specific			**Reader-independent**		
	condition	**over-sample**	**under-sample**	**truncate**	**over-sample**	**under-sample**	**truncate**
adult	**oral**	0.52	0.51	0.51	0.46	0.43	0.40
	silent	0.50	0.51	0.49	0.51	0.50	**0.59**
	both	0.51	0.53	0.49	**0.54**	**0.56**	**0.58**
child	**oral**	0.48	0.49	0.45	0.42	0.44	0.39
	silent	**0.58**	0.54	0.55	0.48	0.48	0.52
	both	0.49	0.46	0.45	0.42	0.44	0.39
both	**oral**	0.50	0.49	0.48	0.52	0.49	**0.58**
	silent	0.54	**0.56**	0.53	0.48	0.50	**0.54**
	both	0.49	0.49	0.46	0.50	0.51	**0.54**

4 What EEG Components Are Sensitive, to What Features?

To identify sensitive frequency bands, we fit 8 separate linear mixed effects models, one model for each combination of modality (oral vs. silent), item type (sentences vs. isolated words), and population (adults vs. children). A logit transform of the dependent variable predicts whether reading an item of that type in that modality is easy or hard for that population. As fixed factors we used the average value of each standard frequency band – Delta (1 to 3Hz), Theta (4 to 7 Hz), Alpha (8 to 11 Hz), Beta (12 to 29 Hz), Gamma (30 to 100 Hz), and Gamma+ (101 to 256 Hz) – averaged over the duration of the item. We included individual reader identity as a random factor to model the population of readers by allowing a separate intercept value for each reader. Linear mixed effects models are robust to missing data, so we included readers with partial data, for a total of up to 8 adults or 12 children in each model. We used the Wald Z statistic to test significance at the $p < .05$ level. Despite the small number of readers, we found statistically significant – but different – predictors for adult and child oral sentence reading: the beta band for adults and the gamma band for children.

Besides training classifiers to distinguish easy from hard reading, we performed a follow-up analysis to take advantage of between-sentence variance in lexical content. We took several lexical properties of words from the MRC Psycholinguistic Database [11] and computed their mean values for each sentence. The between-sentence variance of these per-sentence means provided a natural experiment on the EEG effects of lexical properties. We correlated these sentence-level values against the EEG power spectrum for each sentence. The within-sentence variance in lexical properties naturally diluted the correlations, as did EEG signal noise. Adjusting them to compensate for such variance would more accurately estimate presumably stronger true underlying correlations [13].

Table 3. Correlations of EEG power spectra to mean MRC lexical features of sentences: Concreteness CNC, imageability IMG, Mean Colorado Meaningfulness CMEAN, familiarity FAM, age of acquisition AOA, Brown verbal frequency BFRQ, Kucera and Francis written frequency KFRQ, Thorndike-Lorge frequency T-LFRQ, and # letters NLET

	Delta (1-3 Hz)	Theta (4-7 Hz)	Alpha (8-11 Hz)	Beta (12-29 Hz)	Gamma (30-100 Hz)	Gamma+ (101-256 Hz)
CNC						
IMG						
CMEAN		-0.08				-0.10
FAM						-0.08
AOA						
BFRQ	**-0.12**	**-0.13**			**-0.09**	**-0.12**
KFRQ			0.07			**0.10**
T-LFRQ						**0.09**
NLET	**0.11**	**0.13**	**0.10**	**0.10**	**0.14**	**0.16**

Table 3 shows the unadjusted correlations, with a row for each lexical feature and a column for each frequency band. It shows all correlations significant at $p < .05$ without correction for multiple comparisons, and in **bold** if significant using False Discovery Rate [14]. The table shows effects of word length (NLET) and verbal frequency (BFRQ) across multiple frequencies, and (with less confidence) effects of other features in other bands. Differences among features in which bands they correlate with would suggest that different frequency bands carry information about different word-level aspects of reading – information conceivably useful to an automated tutor.

5 Conclusions

We showed that the EEG data from a single electrode portable recording device can discriminate between reading easy and hard sentences reliably better than chance, across populations (adults and children) and modalities (oral and silent reading). We identified frequency bands sensitive to difficulty and to various lexical properties, which suggests that they can detect transient changes in cognitive task demands or specific attributes of lexical access.

Much work remains. We need to detect additional mental states. We need to improve classifier accuracy by collecting more data and by using more sophisticated training methods. Besides manipulating stimuli experimentally, we can label training data based on observable events in longitudinal data, such as improved performance.

Nevertheless, the statistically reliable relationship between reading difficulty and relatively impoverished EEG data illustrates its potential to detect mental states relevant to tutoring, such as comprehension, engagement, and learning. At the level of longitudinal data aggregated across students, such information could help generate and test hypotheses about learning, elucidate the interplay among emotion, cognition, and learning, and identify specific tutor behaviors to prefer. At the level of dynamic data about an individual student, the tutor could adapt to the student, either by responding immediately to a detected mental state, or by adapting more slowly to a cumulative student model updated over time. In summary, this pilot study gives hope that a school-deployable EEG device can capture tutorially relevant information.

Acknowledgments. This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A080157 and R305A080628 to Carnegie Mellon University. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or U.S. Department of Education. We thank the students, educators, and LISTENers who helped generate, collect, and analyze our data, Sarah Laszlo for stimuli, and the reviewers for helpful comments.

References

1. Marosi, E., Bazán, O., Yañez, G., Bernal, J., Fernández, T., Rodríguez, M., Silva, J., Reyes, A.: Narrow-band spectral measurements of EEG during emotional tasks. Int. J. Neurosci. 112(7), 871–891 (2002)
2. Lutsyuk, N., Éismont, E., Pavlenko, V.: Correlation of the characteristics of EEG potentials with the indices of attention in 12-to 13-year-old children. Neurophysiology 38(3), 209–216 (2006)

3. Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., Craven, P.L.: EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviat. Space Environ. Med. 78(5 Suppl), B231–B244 (2007)
4. Baker, R., D'Mello, S., Rodrigo, M.M., Graesser, A.: Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies 68(4), 223–241 (2010)
5. NeuroSky: NeuroSky's eSenseTM Meters and Detection of Mental State. Neurosky, Inc. (2009)
6. Mostow, J., Beck, J.: When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens. In: Schneider, B., McDonald, S.-K. (eds.) Scale-Up in Education, vol. 2, pp. 183–200. Rowman & Littlefield Publishers, Lanham, MD (2007)
7. Mostow, J., Beck, J.E.: Why, What, and How to Log? Lessons from LISTEN. In: Proceedings of the Second International Conference on Educational Data Mining, Córdoba, Spain, pp. 269–278 (2009)
8. Chawla, N.V., Japkowicz, N., Kolcz, A.: Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter 6(1), 1–6 (2004)
9. Mitchell, T., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M.A., Newman, S.D.: Learning to decode cognitive states from brain images. Machine Learning 57, 145–175 (2004)
10. Bizas, E., Simos, P.G., Stam, C.J., Arvanitis, S., Terzakis, D., Micheloyannis, S.: EEG Correlates of Cerebral Engagement in Reading Tasks. Brain Topography 12(2), 99–105 (1999)
11. Coltheart, M.: The MRC Psycholinguistic Database. Quarterly Journal of Experimental Psychology 33A, 497–505 (1981)
12. Laszlo, S., Federmeier, K.D.: The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. Psychophysiology (in press)
13. Behseta, S., Berdyyeva, T., Olson, C.R., Kass, R.E.: Bayesian Correction for Attenuation of Correlation in Multi-Trial Spike Count Data. Journal of Neurophysiology 101(4), 2186–2193 (2009)
14. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29(4), 1165–1188 (2001)