Towards Using EEG to Improve ASR Accuracy

Yun-Nung Chen, Kai-Min Chang, and Jack Mostow
Project LISTEN (http://www.cs.cmu.edu/∼listen)
School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213-3891, USA
{yvchen,kkchang,mostow}@cs.cmu.edu

Abstract

We report on a pilot experiment to improve the performance of an automatic speech recognizer (ASR) by using a single-channel EEG signal to classify the speaker's mental state as reading easy or hard text. We use a previously published method (Mostow et al., 2011) to train the EEG classifier. We use its probabilistic output to control weighted interpolation of separate language models for easy and difficult reading. The EEG-adapted ASR achieves higher accuracy than two baselines. We analyze how its performance depends on EEG classification accuracy. This pilot result is a step towards improving ASR more generally by using EEG to distinguish mental states.

1 Introduction

Humans use speech to communicate what's on their mind. However, until now, automatic speech recognizers (ASR) and dialogue systems have had no direct way to take into account what is going on in a speaker's mind. Some work has attempted to infer cognitive states from volume and speaking rate to adapt language modeling (Ward and Vega, 2009) or from query click logs (Hakkani-Tür et al., 2011) to detect domains. A new way to address this limitation is to infer mental states from electroencephalogram (EEG) signals.

EEG 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 EEG can measurably detect such variation.

Recently, a few companies have scaled back medical grade EEG technology to create portable EEG headsets that are commercially available and simple to use. The NeuroSky MindSet^{TM} (2009), 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 expertise to wear. Even with the limitations of recording from only a single sensor and working with untrained users, Furthermore, Mostow et al.(2011) used its output signal to distinguish easy from difficult reading, achieving above-chance accuracy. Here we build on that work by using the output of such classifiers to adapt language models for ASR and thereby improve recognition accuracy.

The most similar work is Jou and Schultz's (2008) use of electromyographic (EMG) signals generated by human articulatory muscles in producing speech. They showed that augmenting acoustic features with these EMG features can achieve rudimentary silent speech detection. Pasley et al. (2012) used electrocorticographic (ECoG) recordings from nonprimary auditory cortex in the human superior temporal gyrus to reconstruct acoustic information in speech sounds. Our work differs from these efforts in that we use a consumer-grade single-channel EEG sensor measuring frontal lobe activities, and that we use the detected mental state just to help improve ASR performance rather than to dictate or reconstruct speech, which are much harder tasks.

Section 2 describes how to use machine learning to distinguish mental states associated with easy and difficult readings. Section 3 describes how we use EEG classifier output to adapt ASR language models. Section 4 uses an oracle simulation to show how increasing EEG classifier accuracy will affect ASR accuracy. Section 5 concludes.

2 Mental State Classification Using EEG

We use training and testing data from Mostow et al.'s (2011) experiment, which presented text passages, one sentence at a time, to 10 adults and 11 nine- to ten-year-olds wearing a Neurosky Mindset^{TM} (2009). They read three easy and three difficult texts aloud, in alternating

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 382–385,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

order. The "easy" passages were from texts classified by the Common Core Standards[1] at the K-1 level. The "difficult" passages were from practice materials for the Graduate Record Exam[2] and the ACE GED test[3]. Across the reading conditions, passages ranged from 62 to 83 words long. Although instructed to read the text aloud, the readers (especially children) did not always read correctly or follow the displayed sentences.

Following Mostow et al. (2011), we trained binary logistic regression classifiers to estimate the probability that an EEG signal is associated with reading an easy (or difficult) sentence. As features for logistic regression we used the streams of values logged by the MindSet:

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

Head movement or system instability led to missing or poor-quality EEG data for some utterances, which we excluded in order to focus on utterances with clear acoustic and EEG signals. The features for each utterance consisted of measures 1-4, averaged over the utterance, excluding the 15% of observations where measure 5 reported poor signals. After filtering, the data includes 269 utterances from adults and 243 utterances from children, where 327 utterances are for the easy passages and 185 utterances are for the difficult passages. To balance the classes, we used the undersampling method for training.

We trained a reader-specific classifier on each reader's data from all but one text passage, tested it on each sentence in the held-out passage, performed this procedure for each passage, and averaged the results to cross-validate accuracy within readers. We computed classification accuracy as the percentage of utterances classified correctly. Classification accuracy for adults', children's, and total oral reading was 71.49%, 58.74%, and 65.45% respectively. A one-tailed t-test, with classification accuracy on an utterance as the random variable, showed that EEG classification was significantly better than chance.

3 Language Model Adaptation for ASR

Traditional ASR decodes a word sequence W^* from the acoustic model and language model as below:

$$\begin{aligned} W^* &= \mathrm{argmax}_W\, P(W \mid A) \\ &= \mathrm{argmax}_W\, \frac{P(A \mid W) \cdot P(W)}{P(A)} \end{aligned} \tag{1}$$

To incorporate EEG, we include mental state N as an additional observation in the decoding procedure:

$$\begin{aligned} W^* &= \mathrm{argmax}_W\, P(W \mid A, N) \\ &= \mathrm{argmax}_W\, \frac{P(A \mid W) \cdot P(W \mid N)}{P(A)} \end{aligned} \tag{2}$$

The six passages use a vocabulary of 430 distinct words. To evaluate the impact on ASR accuracy of using EEG to adapt language models, we needed acoustic models appropriate for the speakers. For adult speech, we used the US English HUB4 Acoustic Model from CMU Sphinx. For children's speech, we used Project LISTEN's acoustic models trained on children's oral reading.

We used separate trigram language models (with bigram and unigram backoff) for easy and difficult text – EasyLM, trained on the three easy passages, and DifficultLM, trained on the three difficult passages. Both language models used the same lexicon, consisting of the 430 words in all six target passages. All experiments used the same ASR parameter values.

As a gold standard, all utterances were manually transcribed by a native English speaker. To measure ASR performance, we computed Word Accuracy (WACC) as the number of words recognized correctly minus insertions divided by number of words in the reference transcripts for each reader, and averaged them.

Then we can adapt the language model to estimate $P(W \mid N)$ using mental state information. Using the EEG classifier described in Section 2, we adapted the language model separately for each utterance, using three types of language model adaptation: hard selection, soft selection, and combination with ASR output.

3.1 Hard Selection of Language Models

Given the probabilistic estimate that a given utterance was easy or difficult ($S_{\text{Easy}}(N)$ and $S_{\text{Difficult}}(N)$), hard selection simply picks EasyLM if the utterance was likelier to be easy, or DifficultLM otherwise:

$$\begin{aligned} P_{\text{Hard}}(W \mid N) &= I_C(N) \cdot P_{\text{Easy}}(W) \\ &+ (1 - I_C(N)) \cdot P_{\text{Diff}}(W). \end{aligned} \tag{3}$$

Here $I_C(N) = 1$ if $S_{\text{Easy}}(N) > S_{\text{Difficult}}(N)$, and $P_{\text{Easy}}(W)$ and $P_{\text{Diff}}(W)$ are the probability of word W in EasyLM and DifficultLM, respectively. For comparison, the Random Pick baseline randomly picks either EasyLM or DifficultLM:

[1] http://www.corestandards.org

[2] http://majortests.com/gre/reading_comprehension.php

[3] http://college.cengage.com:80/devenglish/resources/reading_ace/students

WACC		Adult			Child		
		Easy	Difficult	All	Easy	Difficult	All
(a)	Baseline 1: Random Pick	54.5	51.2	53.8	32.8	14.7	30.6
(b)	EEG-based: Hard Selection	**57.6**	49.4	52.7	**36.4**	**17.0**	**32.8**
(c)	Baseline 2: Equal Weight	63.2	59.9	56.5	37.3	19.5	33.4
(d)	EEG-based: Soft Selection w/o smoothing	57.2	48.8	52.4	35.8	17.2	32.5
(e)	EEG-based: Soft Selection w/ smoothing	**66.0**	**62.3**	**64.2**	**39.8**	**22.7**	**36.2**
(f)	Baseline 3: Weight from ASR ($\alpha = 0$)	63.8	60.6	61.5	39.2	20.0	35.0
(g)	Weight from ASR and EEG ($\alpha = 0.5$)	**64.5**	**63.4**	**63.5**	39.2	21.9	36.0

Table 1: ASR performance of proposed approaches using EEG-based classification of mental states.

$$\begin{aligned} P_{\text{Random}}(W) &= I_R \cdot P_{\text{Easy}}(W) \\ &+ (1 - I_{\text{Random}}) \cdot P_{\text{Diff}}(W). \end{aligned} \tag{4}$$

Here I_R is randomly set to 0 or 1.

3.2 Soft Selection of Language Models

Mental state classification based on EEG is imperfect, and using only the corresponding language model (EasyLM or DifficultLM) to decode the target utterance is liable to perform worse when the classifier is wrong. Thus, we use the classifier's probabilistic estimate that the utterance is easy (or difficult) as interpolation weights to linearly combine EasyLM and DifficultLM:

$$\begin{aligned} P_{\text{Soft}}(W \mid N) &= w_{\text{Easy}}(N) \cdot P_{\text{Easy}}(W) \\ &+ w_{\text{Diff}}(N) \cdot P_{\text{Diff}}(W). \end{aligned} \tag{5}$$

Here $w_{\text{Easy}}(N)$ and $w_{\text{Diff}}(N)$ are from classifier's output.

$$w_{\text{Easy}}(N) = S_{\text{Easy}}(N), w_{\text{Diff}}(N) = S_{\text{Diff}}(N) \tag{6}$$

Additionally, we can adjust the range of weights by smoothing the probability outputted by the EEG classifier:

$$\begin{aligned} w_{\text{Easy}}(N) &= \frac{\delta + S_{\text{Easy}}(N)}{2\delta + 1}, \\ w_{\text{Diff}}(N) &= \frac{\delta + S_{\text{Diff}}(N)}{2\delta + 1} \end{aligned} \tag{7}$$

Here $S_{\text{Easy}}(N)$ (or $S_{\text{Diff}}(N)$) is the classifier's probabilistic estimate that the sentence is easy (or difficult) and δ is the smoothing weight, which we set to 0.5. After smoothing the probabilities, $w_{\text{Easy}}(N)$ and $w_{\text{Diff}}(N)$ each lie within the interval $[0.25, 0.75]$, and $w_{\text{Easy}}(N) + w_{\text{Diff}}(N) = 1$. That is, Soft Selection with smoothing interpolates the two language models, but assigns a weight of at least 0.25 to each one to reduce the impact of EEG classifier errors. Notice that $\delta = 0$ is equivalent to EEG Soft Selection without smoothing.

For comparison, the Equal Weight baseline interpolates EasyLM and DifficultLM with equal weights:

$$P_{\text{Equal}}(W) = 0.5 \cdot P_{\text{Easy}}(W) + 0.5 \cdot P_{\text{Diff}}(W) \tag{8}$$

3.3 Combination with ASR Output

Given the ASR results from the Equal Weight baseline, we can derive $S'_{\text{Easy}}(N)$ as:

$$\begin{aligned} S'_{\text{Easy}}(N) &= \alpha \cdot S_{\text{Easy}}(N) \\ &+ (1 - \alpha) \cdot \frac{P_{\text{Easy}}(W_0)}{P_{\text{Easy}}(W_0) + P_{\text{Diff}}(W_0)} \end{aligned} \tag{9}$$

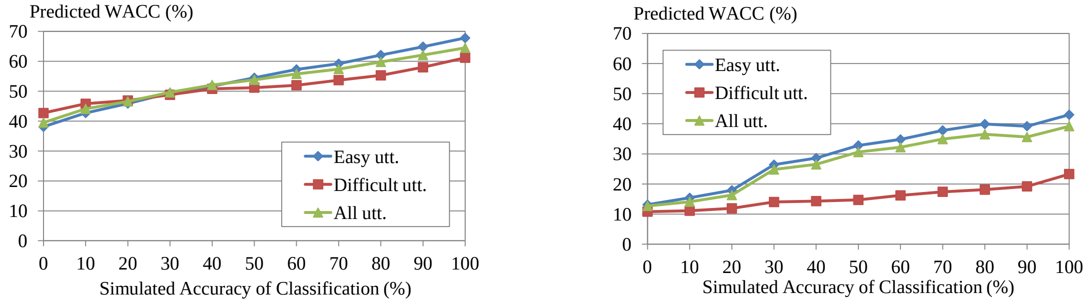
Here we can estimate $S'_{\text{Easy}}(N)$ based on the classifier's output and the probability of the recognized words W_0 in EasyLM. We can derive $S'_{\text{Diff}}(N)$ in the same way. Then we can use (5) and (7) to re-decode the utterances by using $S'_{\text{Easy}}(N)$ and $S'_{\text{Diff}}(N)$. Here α is a linear interpolation weight, where we set to 0.5 to give equal weights to ASR output and EEG. For comparison, the ASR baseline uses weights from only the ASR results, where $\alpha = 0$. Notice that the case of $\alpha = 1$ is equivalent to EEG Soft Selection with smoothing.

3.4 Results of Proposed Approaches

Table 1 shows the performance of our proposed approaches and the corresponding baselines as measured by WACC. According to one-tailed t-tests with word accuracy of an utterance as the random variable, the results in **boldface** are significantly better tgan their respective baselines ($p \leq 0.05$).

Hard Selection (row b) outperforms the Random Pick baseline (row a). Soft Selection without smoothing (row d) has similar performance as Hard Selection because the classifier often outputs probability estimates that are either 1 or 0. However, Soft Selection with smoothing (row e) outperforms the Equal Weight baseline (row c). The Weight from ASR baseline (row f) is better than the other baselines. Weight from ASR and EEG (row g) can further improve performance, but it's not better than Soft Selection with smoothing (row e) - evidence that EEG gives good estimation for choosing language models. In short, Table 1 shows that using EEG to choose between EasyLM and DifficultLM achieves higher ASR accuracy than the baselines that do not use EEG.

Comparing the first two baselines, the Equal Weight baseline (row c) outperforms the Random Pick baseline



(a) Adult

(b) Child

Figure 1: The simulated accuracy graphs plot the predicted ASR word accuracy against the level of EEG classification accuracy simulated by an oracle.

(row a) in every column, because the loss in ASR accuracy from picking the wrong language model outweighs the improvement from picking the right one. Similarly, EEG-based Soft Selection with smoothing (row e) outperforms EEG-based Hard Selection (row b) in every column because the interpolated language model is more robust to EEG classification error. The third base-line, Weight from ASR (row f) depends solely on ASR results to estimate weights; it performs better than other baselines, but not as well as EEG-based Soft Selection with smoothing (row e). That is, using EEG alone can weight the two language models better than ASR alone.

4 Oracle Simulation

To explore the relationship between EEG classifier accuracy and the effect of EEG-based adaptation on ASR accuracy, we simulate different classification accuracies and used Hard Selection to predict the resulting ASR accuracy by selecting between the ASR output from EasyLM and DifficultLM according to the simulated classifier accuracy. We use the resulting Word Accuracy to predict ASR performance at that level of EEG classifier accuracy.

Figure 1 plots predicted ASR WACC against simulated EEG classification accuracy. As expected, the predicted ASR accuracy increases as EEG classification accuracy increases, for both groups (adults and children) and both levels of difficulty (easy and difficult). However, Figure 1a and 1b shows that WACC was much lower for children than for adults, especially on difficult utterances, where even 100% simulated EEG classifier accuracy achieves barely 20% WACC. One explanation is that on difficult sentences, children produced reading mistakes and/ or off-task speech. In contrast, adults read better and stayed on task. Not only is predicted ASR accuracy higher on adults' reading, it improves substantially as simulated EEG classifier accuracy increases.

5 Conclusion

This paper shows that classifying EEG signals from an inexpensive single-channel device can help adapt language models to significantly improve ASR performance. An interpolated language model smoothed to compensate for classification errors yielded the best performance. ASR performance depended on the accuracy of mental state classification. Future work includes improving EEG classification accuracy, detecting other relevant mental states, such as emotion, and improving ASR by using word-level EEG classification. A neurologically-informed ASR may better capture what people intend to communicate, and augment acoustic input with non-verbal cues to ASR or dialogue systems.

Acknowledgements

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080628 to Carnegie Mellon University. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied of the Institute or the U.S. Department of Education. We thank the students, educators, and LISTENers who helped create our data, and the reviewers for their helpful comments.

References

Hakkani-Tür, D., Tur, G., Heck, L., and Shriberg, E. 2011. Bootstrapping domain detection using query click logs for new domains *Proceedings of InterSpeech*, 709-712.

Jou, S.-C. S. and Schultz, T.. 2008. Ears: Electromyograpical Automatic Recognition of Speech. *Proceedings of Biosignals*, 3-12.

Mostow, J., Chang, K.-M., and Nelson, J. 2011. Toward Exploiting EEG Input in a Reading Tutor. *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, 230-237.

NeuroSky 2009. NeuroSky's Sense^{TM} Meters and Detection of Mental State: Neurisky, Inc.

Pasley, B. N. and et al. 2012. Reconstructing speech from auditory cortex. *PLos Biology*, 10(1), 1-13.

Ward, N. G. and Vega, A. 2009. Towards the use of cognitive states in language modeling. *Proceedings of ASRU*, 323-326.