EEG-assisted Modulation of Sound Sources in the Auditory Scene

Marzieh Haghighi[a,*], **Mohammad Moghadamfalahi**[a], **Murat Akcakaya**[b], and **Deniz Erdogmus**[a]

[a]Northeastern University, 360 Huntington Ave, Boston, MA 02115

[b]University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260

Abstract

Noninvasive EEG (electroencephalography) based auditory attention detection could be useful for improved hearing aids in the future. This work is a novel attempt to investigate the feasibility of online modulation of sound sources by probabilistic detection of auditory attention, using a noninvasive EEG-based brain computer interface. Proposed online system modulates the upcoming sound sources through gain adaptation which employs probabilistic decisions (soft decisions) from a classifier trained on offline calibration data. In this work, calibration EEG data were collected in sessions where the participants listened to two sound sources (one attended and one unattended). Cross-correlation coefficients between the EEG measurements and the attended and unattended sound source envelope (estimates) are used to show differences in sharpness and delays of neural responses for attended versus unattended sound source. Salient features to distinguish attended sources from the unattended ones in the correlation patterns have been identified, and later they have been used to train an auditory attention classifier. Using this classifier, we have shown high offline detection performance with single channel EEG measurements compared to the existing approaches in the literature which employ large number of channels. In addition, using the classifier trained offline in the calibration session, we have shown the performance of the online sound source modulation system. We observe that online sound source modulation system is able to keep the level of attended sound source higher than the unattended source.

1. Introduction

Approximately 35 million Americans (11.3% of the population) suffer from hearing loss; this number is increasing and is projected to reach 40 million by 2025 [1]. Within this population only 30% prefer using current generations of hearing aids that are available on the market. One of the most common complaints associated with hearing-aid use is

*Corresponding author: haghighi@ece.neu.edu (Marzieh Haghighi).

understating speech in the presence of noise and interferences. Effects of interfering sounds and masking on speech intelligibility and audibility has been widely studied [2], [3]. Specifically, it has been shown [3] that increase in SNR needed for the same level of speech understanding given a background noise for people with hearing loss can be as high as 30 dB more compared to people with normal hearing. Therefore, amplifying the target signal versus unwanted noises and interferences to facilitate hearing and increase speech intelligibility and listening comfort is one of the basic concepts exploited by hearing aids [3]. Identifying the signal versus noise is a main step required for the design of a hearing aid. It can be a difcult task in complicated auditory scenes like a cocktail party scenario, signal and interferences have acoustic features of speech and can instantly switch their roles based on the attention of the listener and can not be detected based on the predefined assumptions on signal and noise features. Our brain distinguishes the sources based on their spectral profile, harmonicity, spectral or spatial separation, temporal onsets and offsets, temporal modulation, and temporal separation[4],[5] and focus on one sound to analyse the auditory scene [6] in the so called cocktail party effect [7]. Existence of each cue can reduce informational and energetic masking of competing sources and help focusing our attention on the target source.

Brain/Body Computer Interface (BBCI) systems can be used to augment the current generations of hearing aids by discriminating among attended and unattended sound sources. They can be incorporated to provide external evidence based on top-down selective attention of listeners [8]. Attempts have been made to incorporate bottom-up attention evidences in the design of the hearing aids. Direction based hearing aids that detect attention direction from eye gaze and amplify sounds coming from that direction can be an example of bottom-up attention evidence incorporation [9]. Moreover, there are attempts to use electroencephalography (EEG)-based brain computer interfaces (BCIs) for the identification of attended sound sources. EEG has been extensively used in BCI designs due to its high temporal resolution, non-invasiveness, and portability. These characteristics, in addition to EEG devices being inexpensive and accessible, make EEG a practical choice for the design of a BCI that can be integrated into hearing aids to identify auditory attention. A crucial step in such an integration is to build an EEG-based BCI that employs auditory attention.

EEG-based BCIs that rely on external auditory stimulation have recently attracted attention from the research community. For example, auditory-evoked P300 BCI spelling system for locked-in patients is widely studied [10], [11], [12], [13], [14], [15]. It was shown that fundamental frequency, amplitude, pitch and direction of audio stimuli are distinctive features that can be processed and distinguished by the brain. Also, more recent studies using EEG measurements have shown that there is a cortical entrainment to the temporal envelope of the attended speech [16], [17], [18]. A study on the quality of cortical entrainment to auditory stimulus envelope by top-down cognitive attention has shown enhancement of obligatory auditory processing activity in top-down attention responses when competing auditory stimuli differ in space direction [19] and frequency [20].

Recently, EEG-based BCI has also been used in cocktail party problems for the classification of attended versus unattended sound sources [21], [22], [23]. In the identification of an attended sound source in a cocktail party problem, stimulus

reconstruction to estimate the envelope of the input speech stream from high density EEG measurements is the state-of-the-art practice [24], [21]. In the aforementioned model, envelope of the attended stimulus is reconstructed using spatio-temporal linear decoder applied on neural recordings. In one study that considered the identification of the attended sound source in a dichotic (different sounds playing in each ear) two speaker scenario, 60 seconds of high density EEG data recorded through 128 electrodes were used in the stimulus reconstruction. Two decoders using the attended and unattended speech were trained and it was shown that estimated sound source using the attended decoder has higher correlation with the attended speech compared to the estimated stimuli using unattended decoder with unattended speech [21]. Another study extended the approach in [21] to use almost one fourth of the high-density 96 channel EEG data in the identification of attended speaker in a scenario in which the speakers are simultaneously generating speech from different binaural directions [22]. Using 60 seconds of EEG data, they were able to replicate previous work [21] and show its robustness in a more natural sound presentation paradigm [22]. In a related study, authors have compared three types of features extracted from speech signal and EEG measurements to learn a linear classifier for the identification of the attended speaker using 20 seconds of data from high density 128 channels EEG recordings [23]. Moreover, a proof of concept hearing aid system that uses EEG assistance (with 60 seconds of EEG data) to decide on attended sound source by the hearing aid user was first demonstrated in [25]. In our previous related work, we have investigated the role of frequency and spatial features of audio stimuli signal in EEG activities in an auditory BCI system with the purpose of detecting the attended auditory source in a cocktail party setting. We reported high performance single channel classification of attended sound versus unattended one based on their frequency and direction using 60 seconds of EEG and stimuli data [26]. Even though these results are still far from satisfying the requirements for the EEG-based attention detector which can be incorporated in an online setting for a hearing aid application, they are motivating for further investigation in this area.

This paper presents two contributions to the literature on EEG-based auditory attention estimation:

- First, we show successful identification of attended speaker source in a diotic (both sounds playing in both ears) two speaker scenario using 20 seconds of EEG data recorded from 16 channels. The presented classifier outperforms EEG-based auditory attention detectors previously presented in the literature in terms of accuracy, with smaller number of EEG channels (sparse 16 versus dense 96 or more), and using time-series of shorter durations (around 20 seconds versus typically 60 seconds). In fact, using 20 seconds of EEG data from only one of the 16 EEG channels, we demonstrate high classification performance for the auditory attention detection.

- Second, we introduce the first online system that gives feedback on attention of the user in the form of attended to unattended source energy ratio amplification. The level of amplification of attended versus suppression of unattended source is assigned based on a probabilistic model defined over the classifier trained on the offline data including temporal dependency of the user's attention. The goal of

the online system is using the probabilistic information of the user's attention to enhance the concentration of the user on the target source in multi-speaker scenarios. The introduced framework for online system is a proof of concept for design perspective of an EEG-augmented hearing aid system. Finally, we show the introduced online system in average is able to keep the level of attended source higher despite statistical changes happening in online data compared to the offline data used for training the classifier.

2. System Overview

The diagram represented in Figure 1 summarizes the steps of the proposed BCI system. The proposed system gets the mixture of sounds from the environment as the input and modifies the gain of each specific sound. The output of this system is the input to the ear channel.

The decision on gain modification of each sound is made by the BCI module which consists of three submodules of gain controller, auditory attention inference system and hearing aid DSP system. Hearing aid DSP system estimates independent sound sources from the mixture of sounds in the environment and outputs the information to the gain controller and attention inference module. In this work, we assume that we have the estimated sources which are the outputs of the DSP system based on blind source separation.

Auditory attention inference system estimates the probability of attention on each specific sound source using EEG measurements and estimated sound sources. Gain controller system takes the estimated probabilities from the attention inference system to modify gains of each specific sound. The details of the attention inference system and gain adjustments are provided in the following sections.

2.1.   Online Gain Controller System

Lets assume that $S_n = (s_{1,n}, \ldots, s_{i,n}, \ldots, s_{M,n})$ is a matrix containing original sources that each $s_{i,n}$ is a column vector for i^{th} sound channel for n^{th} round of sending feedback. $\widehat{S}_n = (\widehat{s}_{1,n}, \ldots, \widehat{s}_{i,n}, \ldots, \widehat{s}_{M,n})$, would be the estimated source matrix after blind source separation, which we assume exists and its design is out of the scope of this paper. $\mathbf{w_n} = (w_{1,n}, \ldots, w_{i,n}, \ldots, w_{M,n})^\top$ is the vector of weights with $w_{i,n}$ being a scalar showing the gain of ith estimated sound source; and e_n is the EEG evidence vector for n^{th} round. $A_n = i$, indicates the attention of subject is on the ith sound source. Subject will start listening to all sounds with equal energy and then based on brain interface decisions for subject attention on each sound source, speech enhancement or automatic gain controller (AGC) module will assign appropriate weights to each sound source for $n + 1^{th}$ round of sending feedback according to the following equation:

$$\mathbf{w_{n+1}} = f\left(p\left(A_n \middle| \varepsilon_n\left(\widehat{S}_n, e_n\right)\right), \mathbf{w_{n-j}}\right) \; j = 1, 2, \ldots \quad (1)$$

Equation 1 states that the weights for the upcoming sound sources ($n + 1^{th}$ round) will be decided based on probability of attention given current EEG evidence (n^{th} round) and

previous weights that were used at the $n - 1^{th}$ round. The selection of optimal gain control policies (choosing the form of *f*) that considers other factors such as sound quality due to amplitude modulation, response time to changes versus robustness to outlier incidents influencing brain interface decisions, is anticipated to be a significant and important research area in itself, and we will explore alternative designs in future work.

2.2. Auditory Attention Inference System

This module calculates probability of attention given EEG evidence. It takes raw EEG measurements, (estimated) sound sources and weights to extract EEG features (evidence), as explained in Section 4. Then, using Bayes rule, the posterior probability distribution of attention over sources is expressed as the product of EEG evidence likelihood times the prior probability distribution over sources,

$$P\big(A_n = i \big| \varepsilon_n\big) \propto P\big(\varepsilon_n \big| A_n = i\big) P\big(A_n = i\big). \quad (2)$$

In our experiments, we start with a uniform prior over sources and then prior information will be updated based on the observed EEG evidence as explained in 5.2 as well.

3. Data Collection and Preprocessing

3.1. EEG Neurophysiological Data

Ten volunteers (5 male, 5 female), between the ages of 25 to 30 years, with no known history of hearing impairment or neurological problems participated in this study, which followed an IRB-approved protocol. EEG signals were recorded using a g.USBamp biosignal amplifier using active g.Buttery electrodes with cap application from g.Tec (Graz, Austria) at 256 Hz. Sixteen EEG channels (P1, PZ, P2, CPZ, CZ, C3, C4, T7, T8, FC3, FC4, F3, F4 and FZ according to International 10/20 System) were selected to capture auditory related brain activities over the scalp. Signals were filtered by built-in analog bandpass ([0.5, 60] Hz) and notch (60Hz) filters.

3.2. Experimental Design

Each participant completed one calibration and one online session of experiments. Both sessions included diotic (both sounds playing on both ears simultaneously) auditory stimulation while the EEG was recorded from the participants. Participants passively listened to the auditory stimuli through earphones.

Calibration Session—Total calibration session time was about 30 minutes. More specifically, a calibration session consisted of 60 trials of 20 seconds of diotic auditory stimuli with 4 seconds breaks between each trial. The diotic auditory stimuli are generated by one male and one female speaker. These speakers narrated a story (different story for different speakers chosen from audio books of literary novels) for 20 minutes. We consider every 20 seconds of this 20-minute-long diotic narration as a trial. During the calibration session, participants were asked to passively listen to 20-minute-long narration, and they were instructed to switch their attention from one speaker to another during different trials.

The instructions to switch attention from trial to trial are provided to the user on a computer screen using "f" and "m" symbols for female and male speakers, respectively.

Online Session—The online session is summarized in Figure 2. In the online session, similar to the calibration session one male and one female speaker narrated stories (different story for different speakers) for 20 minutes. The same speakers from the calibration session narrated the stories for the online session, but the stories used in the online session were different than the calibration session. We consider the 20-minute-long narration as 10 two-minute long sequences, each sequence containing 6 twenty-second trials. Before each sequence, participants were asked to attend to one of the speakers through instructions displayed on the computer screen. In each sequence, while the participants were listening to the narrated stories, weights that control the energy of each sound source were updated 6 times after every trial. The equal weight case is defined such that amplitude of each sound source was scaled to yield equal energy and each sequence started with an equal weight trial. There is a 0.5 second pause between 20-second-long trials within each sequence and the weights are updated within this 0.5 second period based on the attention evidence obtained from the EEG recorded from the participants and through the usage of automatic gain controller. Since, the participants were instructed to keep their attention on one of the speakers during each sequence, and during each sequence the weights are adjusted automatically in an online fashion to emphasize the attended sound source, we call this an online session. Silent portions of the story narration longer than 0.2 seconds were truncated to be 0.2 seconds, in order to reduce distraction of participants.

3.3. Data Pre-processing

EEG brain activity measurements were digitally filtered by a linear-phase bandpass filter ([1:5, 42]Hz). For each trial, t_x sec of EEG signal time-locked to the onset of each stimulus was extracted resulting in N samples. The acoustic envelope of speech stimulus signals were calculated using the Hilbert transform and filtered by a low pass filter (with 20Hz cut-off frequency). Then, t_x seconds of acoustic envelope signals following every stimulus and time locked to the stimulus onset were extracted. Optimizing t_x to get good performance with minimum time window is an important factor in the design of online auditory BCI systems. In this paper, we selected $t_x = 20$, based on the results of our previous work which are reported in [26]. The data length was selected based on the analysis we performed over the calibration data such that the length is chosen to optimize area under the receiver operating characteristics curve (AUC) of the intent inference engine with a constraint on the upper bound of the data length. More specifically, we analyzed the AUC as a function of the data length, and we chose the data length value when the changes in the AUC as the data length increased became more incremental for most of the participants.

4. Methods

4.1. Feature Extraction

Top down attention to an external sound source differentially modulates the neural activity to track the envelope of that sound source at different time lags [16], [17], [18]. Therefore, as discriminative features, we calculate the cross correlation (CC) between the extracted EEG

measurements and target and distractor acoustic envelopes at different time lags. $\boldsymbol{\tau_n} = [\tau_1,$ $\cdots, \tau_l, \cdots, \tau_L]^\top$ is the vector of discreet time lag delays in sample between EEG and acoustic envelop of played sounds. In our analysis, we consider $\tau \in [t_1, t_2] \times f_s$, with t_1 and t_2 as sampling times chosen as described below. For each channel, we calculate cross correlations between EEG and the male and female speakers' acoustic envelopes for the time lag sample values defined in τ. Assuming that τ is a $L \times 1$ vector, we concatenate the cross correlation values from male and female speakers into a single vector and hence each feature vector is $2L \times 1$ dimensional. As described in Section 3.3, N is number of EEG data and sound source samples used for CC calculation. We have examined the effect of reducing N on classification results in section 5.1.2.

Therefore, considering the defined notations, we calculate the correlation coefficient between EEG and sound sources at different time lag samples τ_l, denoted by $\rho_{e^{ch}, \hat{s}^i}[\tau_l]$:

$$\rho_{e^{ch}, \hat{s}^i}[\tau_l] = \frac{r_{e^{ch}, \hat{s}^i}[\tau_l]}{\sigma_{e^{ch}}, \sigma_{\hat{s}^i}} . \quad (3)$$

In (3), e^{ch} is EEG data recorded from channel ch, $\hat{s}^i$ is the envelope of ith estimated sound channel, τ_l is a time lag sample, and $r_{e^{ch}, \hat{s}^i}[\tau_l] = E\left[e^{ch}_{n+\tau_l:N}, \hat{s}^i_{n:N-\tau_l}\right]$ is the sample average between e^{ch} and $\hat{s}^i$. Therefore, $\rho_{e^{ch}, \hat{s}^i}[\tau_l]$ is a scalar representing the correlation coefficient between EEG in channel ch and i^{th} sound channel at time lag sample τ_l. So, $\rho_{\mathbf{e^{ch}}, \hat{\mathbf{s}}^i} = \left(\rho_{e^{ch}, \hat{s}^i}[\tau_1], \rho_{e^{ch}, \hat{s}^i}[\tau_2], ..., \rho_{e^{ch}, \hat{s}^i}[\tau_L]\right)$ is $1 \times L$ dimensional vector for L lags in τ range for channel ch and i^{th} sound channel. Feature vector will be formed by concatenation of correlation vectors for all $\hat{s}^i$'s. In our experiments which we have two sound sources this feature vector is specifically defined as $x^{ch} = \left[\rho_{\mathbf{e^{ch}}, \hat{\mathbf{s}}^1}, \rho_{\mathbf{e^{ch}}, \hat{\mathbf{s}}^2}\right]^\top$. x^{ch} is $2L \times 1$ vector for each channel and $x = (x^1, ..., x^{ch}, ..., x^{16})^\top$ is a $2L \times 16$ dimensional matrix which contains features for each trial.

4.2. Classification and Dimension Reduction

As explained in Section 3, the participants were asked to direct their auditory attention to a target speaker during data collection. The other speaker is the distractor. The labeled data collected in this manner is used in the analysis of discrimination between two speakers in a binary auditory attention classification problem. As explained in Section 4.1, for each trial we have **x** as the collection of $2L \times 1$ dimensional cross-correlation features for each channel. For analysis of data using all channels, we apply PCA first for dimensionality reduction to remove zero variance directions. Afterwards, feature vectors for each channel will be concatenated to form a single aggregated feature vector for further analysis. Then, we use Regularized Discriminant Analysis (RDA) [27] as the classifier in our analysis. RDA is a modification of Quadratic Discriminant Analysis (QDA). QDA assumes that data is

generated by two Gaussian distributions with unknown mean and covariances and requires the estimation of these means and covariances of the target and nontarget classes before the calculation of the likelihood ratio. However, since, L, the length of τ, as defined in Section 4.1, is usually large resulting in feature vectors with large dimensions even after the application of PCA, and the calibration sessions are short, the covariance estimates are rank deficient.

RDA eliminates the singularity of covariance matrices by introducing shrinkage and regularization steps. Assume each $\mathbf{x}_i \in \mathbb{R}^p$ is a $p \times 1$-dimensional feature vector for each trial and y_i is its binary label showing if the feature belongs to speaker 1 or 2, that is $y_i \in \{1, 2\}$. Then the maximum likelihood estimates of the class conditional mean and the covariance matrices are computed as follows:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{N} \mathbf{x}_i \delta(y_i, k), \tag{4}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_K} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \delta(y_i, k).$$

where $\delta(\cdot, \cdot)$ is the Kronecker-δ function, k represent a possible class label (here $k \in \{1, 2\}$, and N_k is the number of realizations in class k. Accordingly, the shrinkage and regularization of RDA is applied respectively as follows:

$$\widehat{\boldsymbol{\Sigma}}_k(\lambda) = \frac{(1-\lambda)N_k \boldsymbol{\Sigma}_k + (\lambda)\sum_{k=0}^{1} N_k \boldsymbol{\Sigma}_k}{(1-\lambda)N_k + (\lambda)\sum_{k=0}^{1} N_k}, \tag{5}$$

$$\widehat{\boldsymbol{\Sigma}}_k(\lambda, \gamma) = (1-\gamma)\widehat{\boldsymbol{\Sigma}}_k(\lambda) + (\gamma)\frac{1}{p} tr\left[\widehat{\boldsymbol{\Sigma}}_k(\lambda)\right]\mathbf{I}_p.$$

Here, λ, $\gamma \in [0, 1]$ are the shrinkage and regularization parameters, $tr[\cdot]$ is the trace operator and $\mathbf{I}_p$ is an identity matrix of size $p \times p$. In our system we optimize the values of λ and γ to obtain the maximum area under the receiver operating characteristics (ROC) curve (AUC) in a 5-fold cross validation framework. Finally, the RDA score for a trial with the EEG evidence vector $\mathbf{x}_i$, which is defined as:

$$s_{\text{RDA}}(\mathbf{x}_i) = \log\left(\frac{f_{\mathcal{N}}(\mathbf{x}_i; \boldsymbol{\mu}_2, \widehat{\boldsymbol{\Sigma}}_2(\lambda, \gamma))}{f_{\mathcal{N}}(\mathbf{x}_i; \boldsymbol{\mu}_1, \widehat{\boldsymbol{\Sigma}}_1(\lambda, \gamma))}\right), \tag{6}$$

where $f_{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the Gaussian probability density function with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Here s values are used to plot the ROC curves and to compute the AUC values. RDA can be considered as a nonlinear projection which maps EEG evidence to one dimensional score $\varepsilon = s_{\text{RDA}}(\mathbf{x})$.

Finally, the conditional probability density function of ε given the class label, i.e. $p(\varepsilon = \epsilon|A = i)$ needs to be estimated. We use kernel density estimation on the training data using a Gaussian kernel as

$$\hat{p}(\varepsilon = \epsilon \mid A = i) = \frac{1}{N_i} \sum_{A(v) = i} K_{h_i}(\epsilon - \epsilon(v)), \quad (7)$$

where $\epsilon(v)$ is the discriminant score corresponding to a sample v in the training data, that is to be calculated during cross validation, and $K_{h_k}(.)$ is the kernel function with bandwidth h_k.

For a Gaussian kernel, the bandwidth h_k is estimated using Silverman's rule of thumb $\left(\hat{h}_k = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} \cong 1.06\hat{\sigma}n^{-1/5}\right)$ for each class k [28]. This assumes the underlying density has the same average curvature as its variance-matching normal distribution [29].

5. Analysis and Results

As illustrated in our previous work, [26], features formed using the CC coefficient series $\rho_{\mathbf{e}\mathbf{ch}}, \hat{\mathbf{s}}^1, \rho_{\mathbf{e}\mathbf{ch}}, \hat{\mathbf{s}}^2$ as calculated in (3) show distinct patterns for attended vs unattended sound sources and these patterns are observed to be consistent across participants. For diotic presentation, the highest distinguishable absolute correlation between the sound sources and EEG is identified in the range of [0,400] ms. We accordingly extract features within this range of correlation delay, τ. In this range, we observe a negative correlation for both target and distractor speakers followed by an early positive correlation for the target stimulus and delayed and suppressed version of that positive correlation for the distractor stimulus. These results are quantitatively summarized in Table 1, more specifically this table reports the average temporal latency and the magnitude of the peak in cross correlation responses across all participants. Statistical significance of the difference between peak temporal latency of target and distractor has been tested using Mann-Whitney U-test ($p = .00012$).

In the rest of the analysis, we consider the correlation delay τ to be in the range of [0,400]ms to form the feature vectors.

5.1. Offline Data Analysis

5.1.1. Single channel classification analysis—Using the selected window of [0,400] ms as the most informative window for classification of target versus distractor responses, we first form the vector x^{ch} as shown in 4.1, we then use these features for each EEG channel independently to localize the selective attention responses using the classification scheme described in Section 4.2. As the results of our previous work suggested [26], we relocated electrodes to be more centered around the frontal cortex, see Section 3.1. Figure 3 shows the topographical map of classification performance in terms of area under the receiver operating characteristics curve (AUC) over the scalp, for all participants. Moreover, for each participant best channel AUC values are reported in Table 2. Figure 3 and Table 2 show that the classification accuracy varies across participants, but for each

participant channels located in central and frontal cortices have higher classification accuracy.

5.1.2. Classification performance versus trial length analysis—In this section we analyze the effect of trial length on classification performance. Specifically, using the calibration data, we consider different lengths (from 2 seconds to 20 seconds) of EEG and estimated sound sources to calculate the cross correlations and extract features accordingly to train our classifier to distinguish the attended sound source from the unattended one. Figure 4 shows the classification performance using all 16 channels. In this figure, different colors represent the performances of different participants. The blue curve is the average of performance over all 10 participants using different data lengths for classification. Dark and light shaded areas around the average line shows the 50 and 95 percent confidence interval calculated according to the bootstrap method, respectively. Figure 4(a) shows AUC performance while Figure 4(b) shows probability of correct decision (i.e., accuracy). Moreover, Figure 4(b) also compares our results with a related previous work that is presented in [23]. The performance reported for 128 channels in that previous work is illustrated as a green line in this figure. In this figure, we observe that using much smaller number of channels, our method outperforms the previous approach.

5.2. Online Controller Performance

Recall that as explained in Section 3.2, the online experiment includes listening to 10 two-minute sequences. During each sequence the participants were requested to focus their auditory attention to one of the speakers. Each sequence contains multiple trials and within each sequence we perform adaptive sound source weight estimation and update after every trial (20 seconds). More specifically, we calculate the EEG evidence as explained in Section 4.1. Using conditional probability density functions as described in section 4.2, we obtain the posterior estimate of the probability for each class being the intended source, which is proportional to class conditional likelihoods times prior knowledge on probability of attention. Then source weights for each source are adjusted as being proportional to the posterior probability of that class given EEG evidence.

$$w_{k,n+1}^{(i)} = p\big(A_k = i \big| \varepsilon_{k,n}, ..., \varepsilon_{k,1}\big) \quad (8)$$

$$= \frac{p\big(A_k = i \big| \varepsilon_{k,n}\big) p\big(A_k = i \big| \varepsilon_{k,n-1}, ..., \varepsilon_{k,1}\big)}{\sum_{j=1}^{2} p\big(A_k = j \big| \varepsilon_{k,n}\big) p\big(A_k = j \big| \varepsilon_{k,n-1}, ..., \varepsilon_{k,1}\big)} \quad , i = 1, 2 \quad (9)$$

$$= \frac{p\big(A_k = i \big| \varepsilon_{k,n}\big) \cdot w_{k,n}^{(i)}}{\sum_{j=1}^{2} p\big(A_k = j \big| \varepsilon_{k,n}\big) \cdot w_{k,n}^{(j)}} \quad , i = 1, 2. \quad (10)$$

In the equation above, k is the sequence index and n is the trial index. Each sequence contains 6 trials and during each sequence we assume that the user is focusing on the same sounds source. This equation assumes that the attention remains on the same source during the updates in each sequence. Also in this weight update equation above we initialize $p(A_k = i | \varepsilon_k, 0) = 0.5$. We trained the system using a calibration session and tested the learned model in an online session. Users attempted to amplify the designated target speech with their auditory attention using this brain interface in 10, two-minute-long trials. Figure 5 shows the average of decided weights (at every 20 seconds over 5 trials) for attended and unattended speech sources over the course of two minutes, for male and female narrators. Figure 5 (a) is showing the average of the estimated probabilities for each class at the end of each trial using its preceding 20 seconds of data, as stated in equation 8. Figure 5 (b) shows the average of employed weights instead of normalized probabilities. The difference between Figures 5 (a) and (b) is due to the limits imposed on weights ([0.25 to 0.75] which are shown with green constants). These limitations were imposed to ensure the audibility of both sources, to enable mistake correction in the event of algorithm/human errors, and to allow shifting attention if desired. Figure 6 illustrates two example sequences: one for a normal case in which there is no algorithm/human error (second row of the figure), and the first row of the figure demonstrates a case in which a participant is able to recover from a potential error in detecting the attended sound source. In this second case, the weight of the attended sound source was lower than the unattended one; however, since the system imposes a lower bound on the weights, the participant was able to recover and the weight of the attended sound source increased accordingly before the sequence ended.

5.3. Online Vs Offline data analysis

Since changes in energy and amplitude of competing sound sources will potentially change the statistics of the EEG measurements, analyzing how robust feature vectors are to these changes can help us understand the impacts of the weights of the sound sources on the attention and EEG models. Table 3 shows the generalization of the classifier trained on the calibration data and tested on the data collected during the online sessions when the EEG from all the channels were used. Specifically, the first and the second rows of the table present the AUC results when 5-fold cross validation is performed on the calibration and online session data, respectively. The results in the third row are obtained when the classifier is trained using the calibration data and tested on the online data. Therefore, the third row demonstrates the generalization of the trained auditory attention classifier from the calibration session to the online testing. Note here that the third row demonstrates the performance of the auditory attention classifier when it was used in online session where the sound source weights were adaptively updated. From this table, we observe that there is a decrease in the performance when the classifier is used in the online session compared to the calibration session. Even though the classification accuracy is acceptable when the classifier trained on the calibration data and tested on the online data as illustrated in row 3 of the table, a calibration session with varying weights on the sounds sources could potentially improve the classification accuracy further. This will be the focus of our future work.

6. Conclusion, Limitations and future work

This work is a novel attempt to investigate the feasibility of online classification of auditory attention using a non-invasive EEG-based brain interface. In a multi-speaker scenario, the brain interface presented in this manuscript utilizes an automatic gain control to adjust the amplitudes of attended and unattended sound sources with the goal of increasing signal-noise-ratio and improving listening and hearing comfort. Through an experimental study, we showed that the designed BCI together with the automatic gain control has the potential to improve the information rate by reducing the trial lengths and increasing the classification accuracies for shorter trial lengths compared to the performance results reported in the existing related works. Even though promising results were obtained with this proof of concept study, there are many opportunities to improve the performance of the system. For example, various different techniques could be investigated to optimize the automatic gain control scheme or the classification method with the purpose of enabling fast and accurate decision making in an online setting. This improvement is essential for the presented BCI to be a practical reality and potentially be a part of the future generations of hearing aids.

Acknowledgment

References

[1]. Kochkin S, Marketrak viii: 25-year trends in the hearing health market, Hearing Review 16 (11) (2009) 12–31.

[2]. Gelfand SA, Hearing: An introduction to psychological and physiological acoustics, CRC Press, 2016.

[3]. Chung K, Challenges and recent developments in hearing aids part i. speech understanding in noise, microphone technologies and noise reduction algorithms, Trends in Amplification 8 (3) (2004) 83–124. [PubMed: 15678225]

[4]. Yost WA, The cocktail party problem: Forty years later, Binaural and spatial hearing in real and virtual environments (1997) 329–347.

[5]. Yost WA, Sheft S, Auditory perception, in: Human psychophysics, Springer, 1993, pp. 193–236.

[6]. Bregman AS, Auditory scene analysis: The perceptual organization of sound, MIT press, 1994.

[7]. Cherry EC, Some experiments on the recognition of speech, with one and with two ears, The Journal of the acoustical society of America 25 (5) (1953) 975–979.

[8]. Ungstrup M, Kidmose P, Rank ML, Hearing aid with self fitting capabilities, uS Patent App. 14/167,256 (Jan. 29 2014).

[9]. Kidd G Jr, Favrot S, Desloge JG, Streeter TM, Mason CR, Design and preliminary testing of a visually guided hearing aid, The Journal of the Acoustical Society of America 133 (3) (2013) EL202–EL207. [PubMed: 23464129]

[10]. Schreuder M, Rost T, Tangermann M, Listen, you are writing! speeding up online spelling with a dynamic auditory bci, Frontiers in neuroscience 5.

[11]. Höhne J, Schreuder M, Blankertz B, Tangermann M, A novel 9-class auditory erp paradigm driving a predictive text entry system, Frontiers in neuroscience 5.

[12]. Kübler A, Furdea A, Halder S, Hammer EM, Nijboer F, Kotchoubey B, A brain–computer interface controlled auditory event-related potential (p300) spelling system for locked-in patients, Annals of the New York Academy of Sciences 1157 (1) (2009) 90–100. [PubMed: 19351359]

[13]. Halder S, Rea M, Andreoni R, Nijboer F, Hammer E, Kleih S, Birbaumer N, Kübler A, An auditory oddball brain–computer interface for binary choices, Clinical Neurophysiology 121 (4) (2010) 516–523. [PubMed: 20093075]

[14]. Furdea A, Halder S, Krusienski D, Bross D, Nijboer F, Birbaumer N, Kübler A, An auditory oddball (p300) spelling system for brain-computer interfaces, Psychophysiology 46 (3) (2009) 617–625. [PubMed: 19170946]

[15]. Kanoh S, Miyamoto K.-i., Yoshinobu T, A brain-computer interface (bci) system based on auditory stream segregation, in: Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, IEEE, 2008, pp. 642–645.

[16]. Ding N, Simon JZ, Cortical entrainment to continuous speech: functional roles and interpretations, Frontiers in human neuroscience 8.

[17]. Aiken SJ, Picton TW, Human cortical responses to the speech envelope, Ear and hearing 29 (2) (2008) 139–157. [PubMed: 18595182]

[18]. Kong Y-Y, Mullangi A, Ding N, Differential modulation of auditory responses to attended and unattended speech in different listening conditions, Hearing research 316 (2014) 73–81. [PubMed: 25124153]

[19]. Power AJ, Lalor EC, Reilly RB, Endogenous auditory spatial attention modulates obligatory sensory activity in auditory cortex, Cerebral Cortex 21 (6) (2011) 1223–1230. [PubMed: 21068187]

[20]. Sheedy CM, Power AJ, Reilly RB, Crosse MJ, Loughnane GM, Lalor EC, Endogenous auditory frequency-based attention modulates electroencephalogram-based measures of obligatory sensory activity in humans, NeuroReport 25 (4) (2014) 219–225. [PubMed: 24231831]

[21]. O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC, Attentional selection in a cocktail party environment can be decoded from single-trial eeg, Cerebral Cortex (2014) bht355.

[22]. Mirkovic B, Debener S, Jaeger M, De Vos M, Decoding the attended speech stream with multi-channel eeg: implications for online, daily-life applications, Journal of neural engineering 12 (4) (2015) 046007. [PubMed: 26035345]

[23]. Horton C, Srinivasan R, DZmura M, Envelope responses in single-trial eeg indicate attended speaker in a cocktail party, Journal of neural engineering 11 (4) (2014) 046015. [PubMed: 24963838]

[24]. Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF, Reconstructing speech from human auditory cortex, PLoS Biol 10 (1) (2012) e1001251. [PubMed: 22303281]

[25]. Slaney M, Pay attention, please: Attention at the telluride neuromorphic cognition workshop, IEEE SLTC Newsletter.

[26]. Haghighi M, Moghadamfalahi M, Nezamfar H, Akcakaya M, Erdogmus D, Toward a brain interface for tracking attended auditory sources, in: 2016 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2016, pp. 1–6.

[27]. Friedman JH, Regularized discriminant analysis, Journal of the American statistical association 84 (405) (1989) 165–175.

[28]. Silverman BW, Density estimation for statistics and data analysis, Vol. 26, CRC press, 1986.

[29]. Orhan U, Erdogmus D, Roark B, Oken B, Fried-Oken M, Offline analysis of context contribution to erp-based typing bci performance, Journal of neural engineering 10 (6) (2013) 066003. [PubMed: 24099944]
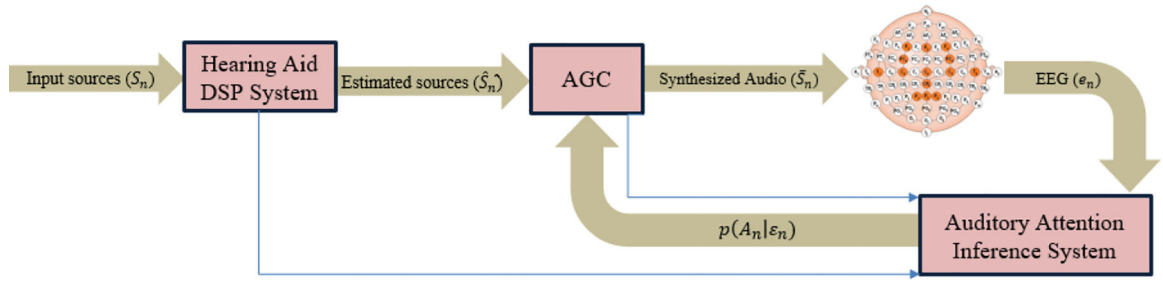
Figure 1:
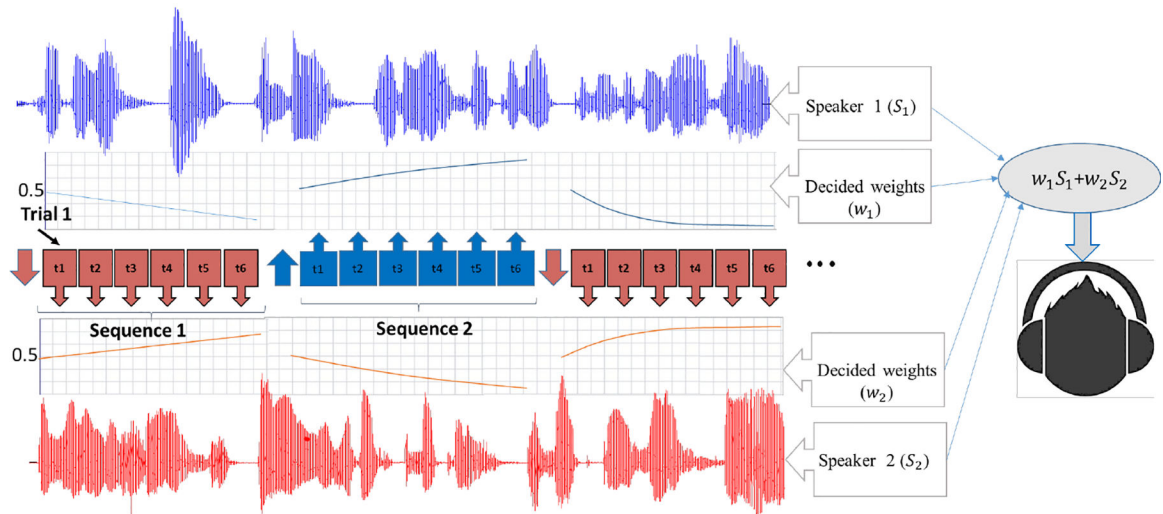EEG-augmented BCI sytem overview.

Figure 2:
Online session experimental paradigm visualization. Two sounds are diotically playing in both ears. participants attend to the instructed sound source in each sequence. Each sequence starts with an equal weight trial and in its following trials weights get updated using attention inference and AGC modules.
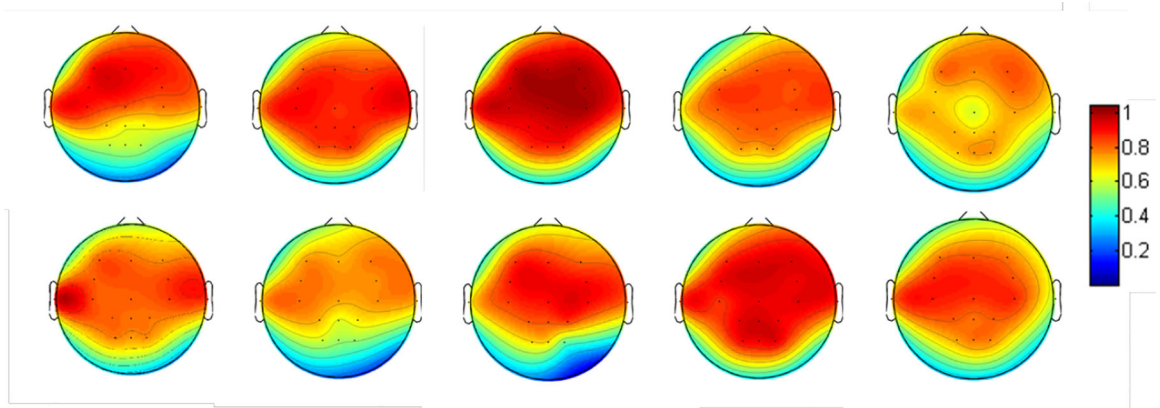
Figure 3:
Topographic map of classification performance over the scalp for classifying attended versus unattended speakers for all participants.
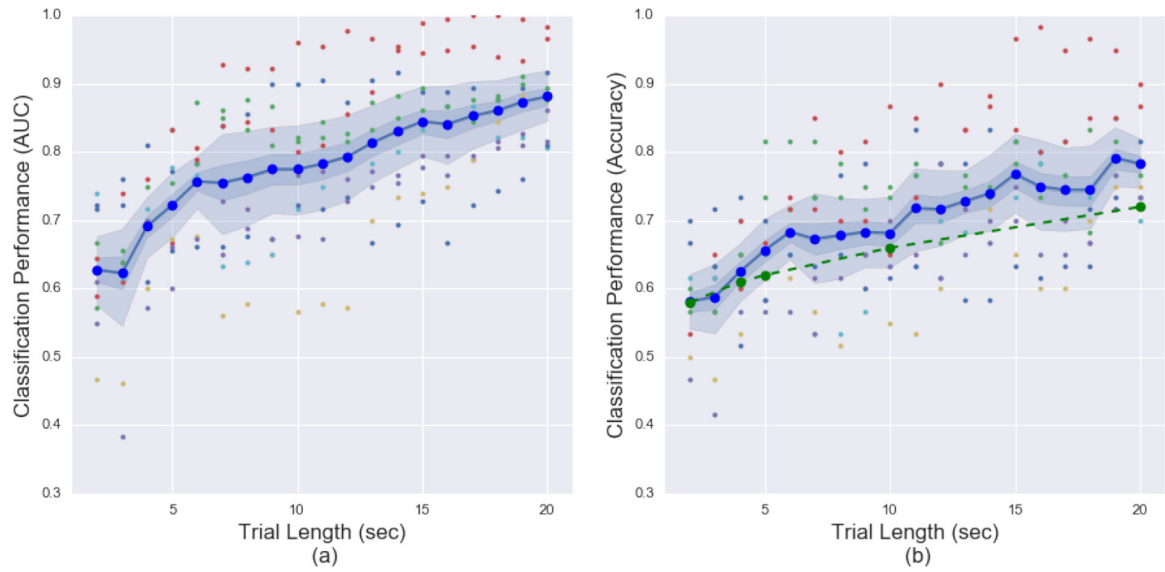
Figure 4:

Performance versus trial length curves considering (a) AUC and (b) Accuracy as performance metrics. Different colored dots are used to represent the performance of different participants. The blue curve is the average of performance values over all 10 participants. The green line presents the performance results of a previous approach.
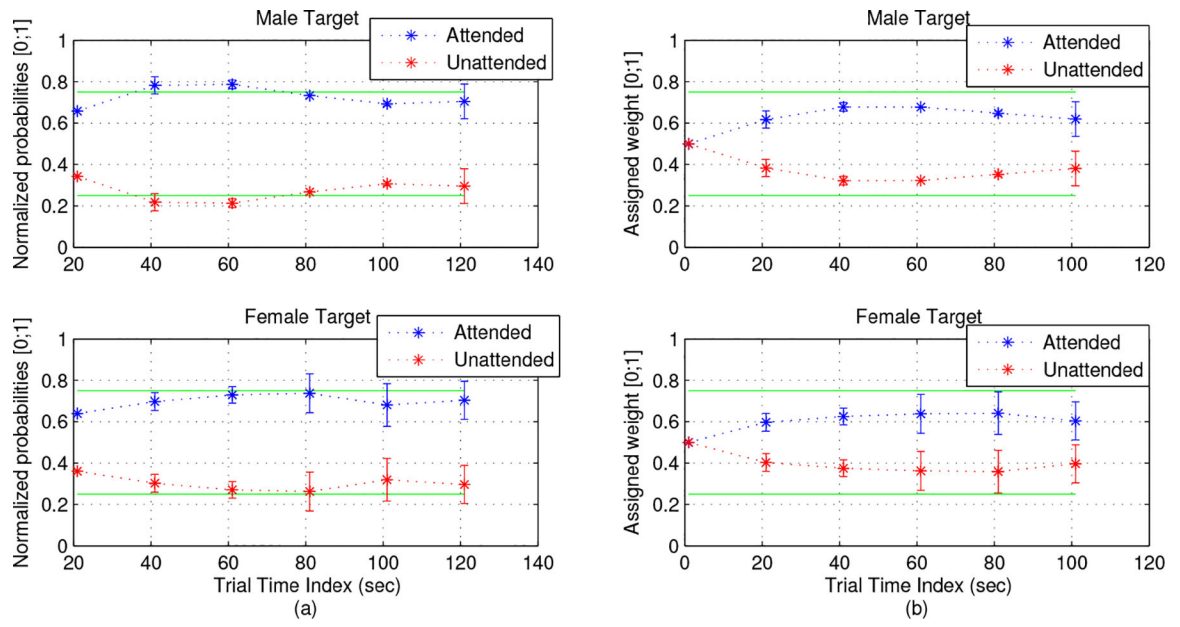
Figure 5:
Weight change in each trial at every 20 second which is averaged over trials and participants for female and male target separately.
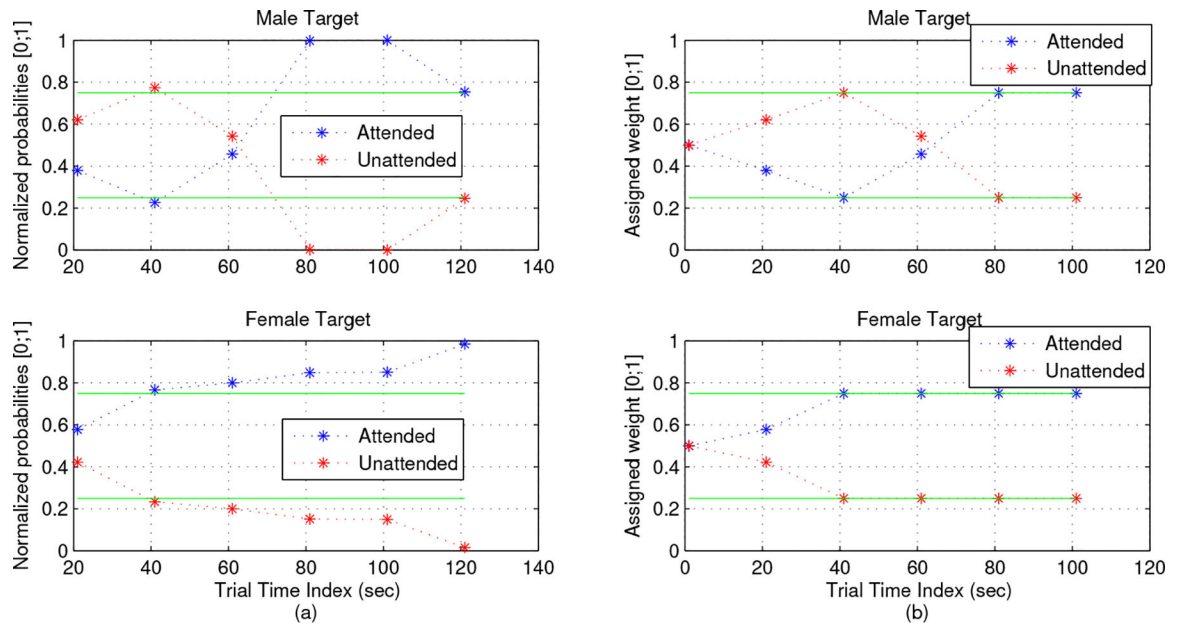
Figure 6:
Two examples for a normal versus mistake recovery case. First row is showing an example of mistake recovery by participant. Second row in an example of a normal sequence

Table 1:

Average of time latency and magnitude of peak in cross correlation responses across all participants.

Correlation Features	Positive Peak Magnitude Ratio	Time Lag of Peak (ms)	
Stimulus	Target / Distractor	Target	Distractor
Average for all Participants (mean ± sd)	2.08 ± 1.1	159.34 ± 11	225.78 ± 42.9

Table 2:

Channel with maximum performance and its corresponding AUC performance.

Participant	1	2	3	4	5	6	7	8	9	10
Best AUC	0.92	0.92	1	0.84	0.83	0.92	0.80	0.91	0.96	0.89
Best channel	Fz	C3	C4	Fc3	F4	T7	C3	C4	CPz	C3

Table 3:

AUCs for offline and online data independently and applying the learned model from offline data on online data.

AUC ＼ Participant	1	2	3	4	5	6	7	8	9	10
Calibration Data (offline)	0.91	0.89	1	0.81	0.88	0.82	0.77	0.89	0.94	0.83
Online Session Data	0.83	0.74	0.98	0.63	0.82	0.8	0.74	0.92	0.83	0.69
Calibration Model on Online Data	0.86	0.77	0.95	0.73	0.77	0.83	0.8	0.82	0.86	0.70