A Graphical Model for Online Auditory Scene Modulation Using EEG Evidence for Attention

Marzieh Haghighi, Mohammad Moghadamfalahi, *Student Member, IEEE*, Murat Akcakaya, *Member, IEEE*, Barbara G. Shinn-Cunningham, and Deniz Erdogmus, *Senior Member, IEEE*

***Abstract*— Recent findings indicate that brain interfaces have the potential to enable attention-guided auditory scene analysis and manipulation in applications, such as hearing aids and augmented/virtual environments. Specifically, non-invasively acquired electroencephalography (EEG) signals have been demonstrated to carry some evidence regarding, which of multiple synchronous speech waveforms the subject attends to. In this paper, we demonstrate that: 1) using data- and model-driven cross-correlation features yield competitive binary auditory attention classification results with at most 20 s of EEG from 16 channels or even a single well-positioned channel; 2) a model calibrated using equal-energy speech waveforms competing for attention could perform well on estimating attention in closed-loop unbalanced-energy speech waveform situations, where the speech amplitudes are modulated by the estimated attention posterior probability distribution; 3) such a model would perform even better if it is *corrected* (linearly, in this instance) based on EEG evidence dependence on speech weights in the mixture; and 4) calibrating a model based on population EEG could result in acceptable performance for new individuals/users; therefore, EEG-based auditory attention classifiers may generalize across individuals, leading to reduced or eliminated calibration time and effort.**

***Index Terms*— EEG, auditory attention detection, brain interface.**

I. Introduction

MOST listeners solve the cocktail party problem (CPP) of attending to one sound in the presence of competing sounds with ease [1], [2]. Healthy adult listeners achieve this by selectively attending to discriminating auditory features of the desired source, such as spectral profile, harmonicity, spatial position, and temporal modulation, in order to differentiate it from other sources in the auditory scene [3]–[6]. Listeners with hearing loss have difficulty focusing selective attention [7], leading to a need for hearing aids to assist with communication settings that may exhibit the CPP. Recent research findings indicate that neural activity reveals auditory cortical mechanisms of selective attention in the CPP using spatial [8]–[11], and spectral or pitch [12] cues of sound sources. (For more information, please see reviews on this topic [13]).

Cortical responses have been shown to entrain to the temporal envelope of attended speech [14]–[16]. These effects can be seen in non-invasive magnetoencephalography (MEG) [17]–[19] and electroencephalography (EEG) [15], as well as during invasive electrocorticography (ECoG) [20]. These observations have lead to efforts to use recorded cortical measurements to reconstruct an attended speech stimulus [21]. Recent studies have quantified the quality of speech reconstructed from such recordings, and have explored how the temporal properties of cortical responses track those of both attended and unattended speech signals [22]–[26].

Differences in cortical responses to attended versus unattended sound sources can be used to develop algorithms that determine what source a listener is attending in the CPP. For example, using EEG [22], [23], MEG [24], or ECoG [25], a stimulus reconstructed from cortical recordings correlates more strongly with an attended source versus an unattended source. Other approaches for categorizing auditory attention during the CPP have also been proposed and tested. In [27], a biophysically inspired state space model that tracked auditory sources using 160 MEG neural measurements was able to categorize attentional focus within the order of a few seconds. In [28], three types of discriminant features were used to train and test a linear classifier that aims to identify attended and unattended sound sources using EEG. They achieved *high* classification accuracy using 20 seconds of 128-channel EEG data.

Previously, we achieved successful classification of attended versus unattended speech in a two-speaker scenario using 60 seconds of single-channel EEG [29] and 20 seconds [30] of 16-channel EEG measurements. In that earlier study, we had also tested the idea of attention-estimate-based modulation of source amplitudes in order to assess the possibility of closed-loop auditory scene modulation and estimation of attention. The introduced system performed well in real-time testing, using 20 seconds of data for inference [30].

In this study, we extend our previously introduced EEG-assisted sound source modulation framework in two

Manuscript received February 13, 2017; revised April 14, 2017; accepted May 14, 2017. Date of publication June 6, 2017; date of current version November 6, 2017. This work was supported by NSF under Grant CNS-1136027, Grant IIS-1149570, and Grant CNS-1544895, in part by NIDLRR under Grant 90RE5017-02-01, and in part by NIH under Grant R01DC009834. *(Corresponding author: Marzieh Haghighi.)*

M. Haghighi is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: haghighi@ece.neu.edu).

M. Moghadamfalahi is with Honeywell Laboratories, Plymouth, MN 55441 USA.

M. Akcakaya is with the University of Pittsburgh, Pittsburgh, PA 15260 USA.

B. G. Shinn-Cunningham is with Boston University, Boston, MA 02215 USA.

D. Erdogmus is with Northeastern University, Boston, MA 02115 USA.

Digital Object Identifier 10.1109/TNSRE.2017.2712419

novel ways: (1) a recursive maximum a posteriori (MAP, i.e., Bayesian with unit cost for errors, zero cost for correct decisions) auditory attention inference procedure based on a generative signal model relating sound envelopes to EEG; in contrast, the previous model used MAP inference with class conditional likelihoods of EEG/envelope cross-correlation sequence based features. (2) a probabilistic model for recursive MAP auditory attention inference extended with temporal dynamics of attention in a given context, as well as the impact of sound source powers on attention. These augmentations make the model more appropriate for auditory attention tracking in a closed-loop system where (estimated) source amplitudes may be modulated.

The new signal-modeling approach did not change performance significantly in test cases where conditions were similar to calibration conditions, but when the auditory scene deviated from calibration conditions, the model based approach allowed us to employ a corrective action that resulted in improved performance. The model based approach also achieved competitive accuracy with a feature vector of lower dimensionality. In addition, the performance of individual-specific and population-based calibration of attention inference models are compared using both cross-correlation and signal-modeling based feature extraction approaches; the results demonstrate that while individual-specific calibration outperforms, subject-to-subject transfer of models is viable and future model can exploit this observation by using hierarchical models of EEG (sessions $<$ individuals $<$ population) to achieve transfer learning benefits towards reduced calibration effort.

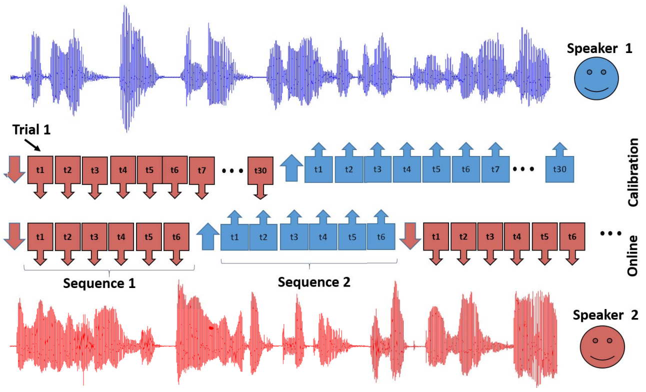
II. Participants and Data Acquisition

A. EEG Acquisition and Preprocessing

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 a Northeastern IRB-approved protocol. EEG signals were recorded using a g.USBamp biosignal amplifier using active g.Butterfly electrodes with cap application from g.Tec (Graz, Austria) at 256 Hz. Sixteen EEG channels (P1, PZ, P2, CP1, CPZ, CP2, CZ, C3, C4, T7, T8, FC3, FC4, F3, F4 and FZ according to International 10/10 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. The acoustic envelope of speech stimulus signals were calculated using the Hilbert transform and then both EEG brain activity measurements and speech envelopes were filtered by a linear-phase bandpass filter ([1.5, 10]Hz). Then, t_x seconds of EEG and acoustic envelope signals following every stimulus and time locked to the stimulus onset were extracted. This dataset has been used in our previous work using a different analysis approach [30].

B. Experimental Paradigm

The experimental paradigm is summarized in Figure 1. Each subject completed one calibration session and one online session. Calibration sessions were approximately 30 minutes and they consisted of 60 trials of 20-second-long stimulus presentations separated by 4-second-long rest periods. In each trial, two speech waveforms were presented simultaneously and diotically (both sounds playing in both ears) as auditory stimuli through earphones. Speech waveforms were selected from audio books of literary novels. One male and one female speaker narrated their stories in each trial for all sessions. The session was divided into two subsessions of 30 trials. In each subsession, subjects were instructed to direct their attention to one of the speakers. For example, if one subject was requested to direct his/her attention to the male speaker in the first subsession, the attention was on the female speaker in the second subsession. The order of the attended speaker was randomized across subjects. The target speaker was indicated to the subjects by displaying the letters F or M on a monitor during each trial. In all speech waveforms, silent portions longer than 0.2 second were truncated to be 0.2 second long. The amplitude of each speech waveform was scaled to yield equal energy in calibration trials.



(a)

Fig. 1. Experimental paradigm visualization: [First Row] Calibration [Second Row] Online (quasi-real-time) sessions. Direction of the arrows indicate the designated attended speaker.

The online session consisted of 10 sequences that were each 2 minutes long. Each sequence contained 6 trials that were each 20 seconds long; the speech waveform energies were normalized such that each trial had equal energy for both speakers. Before each sequence, by displaying the letter F or M on the monitor, subjects were asked to attend to the designated target speech waveform for that sequence. In each sequence, the speech mixture weights were initialized to be equal (0.5 vs 0.5), and they were updated 5 times (after each 20-second-long segment like the calibration trials) during 0.5-second-long pauses in which inference calculations were carried out.

III. Proposed Method

A. System Framework Overview

The overarching BCI framework used in this work was previously introduced [30], and it consists of three main modules: Digital Signal Processing (DSP), Automatic Gain Control (AGC), and Auditory Attention Inference. The system takes a mixture of sounds from the auditory environment as the input, modifies the power of each source sound, and produces a new mixture as the output. In anticipated applications, such

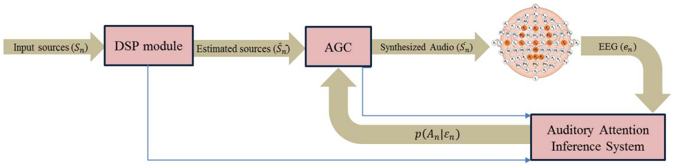

Input sources (S_n)
DSP module
Estimated sources ($\hat{S}_n$)
AGC
Synthesized Audio ($\tilde{S}_n$)
EEG (e_n)
$p(A_n|\varepsilon_n)$
Auditory Attention Inference System

Fig. 2. EEG-based Auditory Scene Modulation.

as hearing aids or augmented/virtual reality, the output of this system would be delivered to the user's ear, closing the loop in online scenarios. The DSP module estimates individual/independent sound sources from the mixture of sounds in the environment (which could involve blind source separation or adaptive interference cancellation in a realistic setting). 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 power of each specific sound. As we also emphasize in the introduction of this manuscript, we extend the probabilistic framework for the Attention Inference module in this manuscript. The details of the extended Attention Inference and AGC modules are provided in the following section.

B. Auditory Attention Inference and AGC Modules

We employ the graphical model presented in Figure 3 to built the Auditory Attention Inference module. In this graphical model, A_n represents the unknown attention of the subject during the n^{th} trial; c is the contextual prior defined over the subject's attention; ε_n is the EEG evidence obtained in response to the attended source during the n^{th} trial; and $\mathbf{w_n}$ is the weight vector that modulates the sound sources.

The graph illustrated in Figure 3 extends the previous model presented in earlier work [30]. The previous model assumed that the unknown attention of the subject was the only factor that affected the EEG evidence and the attention of the subject only depended on a contextual prior. The current model relaxes this assumption to include the dependency of attention A_n to attention at the previous trial A_{n-1}, and the weights that modulate the sound sources during the n^{th} trial, $\mathbf{w_n}$. As the weights modulating different sound sources will be different, this difference may affect the attention of a subject; therefore, we think that this extension in the graphical model is essential to account for the effect of modulation in intent inference.

Using the graphical model presented in Figure 3, for each trial the conditional posterior distribution for the attention of the subject, A_n conditioned on the EEG evidence ε_n, prior attention A_{n-1}, and modulation weights $\mathbf{w_n}$ can be computed: $p(A_n|A_{n-1}, \varepsilon_n, \mathbf{w_n}, \mathbf{c})$. Under certain Independence assumptions, this posterior simplifies to

$$p(A_n|A_{n-1}, \varepsilon_n, \mathbf{w_n}, \mathbf{c}) \propto \\ p(\varepsilon_n|A_n)p(A_n|\mathbf{w_n})p(A_n|A_{n-1}, \mathbf{c})/p(A_n) \qquad (1)$$

where

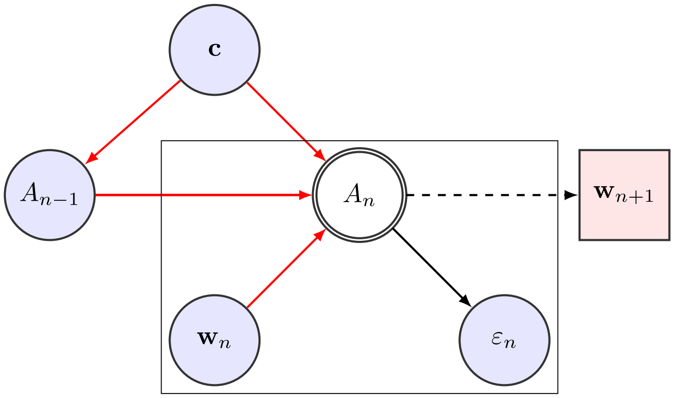


Fig. 3. Proposed graphical model.

- $P(\varepsilon_n|A_n)$ is the conditional distribution of the EEG evidence conditioned on the attended sound source. This conditional distribution is calculated based on EEG evidence features extracted and learned for each sound source (class) as explained in Section III-D.
- $P(A_n|\mathbf{w}_n)$ is the conditional distribution of the attention conditioned on the weights that modulate the sound sources. In the decision framework, this conditional probability models how the relative power of each sound source influences the probability of attending to that specific source. In Section IV-C, we describe how we develop an approximate model for this conditional distribution.
- $P(A_n|A_{n-1}, \mathbf{c})$ is the conditional distribution of the attention in the n^{th} trial conditioned on the attention on the previous trial and context prior. Based on context, transition of the attention from one state to another can be prompt or slow. In our experiments, we assume that this conditional distribution is uniform due to the lack of any contextual evidence.
- $P(A_n)$ is the prior distribution on the attention of a subject during the n^{th} trial. In our experiments, since the trials are independent from each other, we assume a uniform prior over the attention during each trial.

Next we use the posterior distribution computed in (1) to calculate the weights for the $(n+1)^{\text{th}}$ trial, $\mathbf{w_{n+1}}$, that are used by the AGC module to modulate the sound sources. More specifically, lets assume that $S_n = (\mathbf{s}_{1,n}, ..., \mathbf{s}_{i,n}, ..., \mathbf{s}_{M,n})$ is a matrix containing original sound sources presented during the n^{th} trial such that each column corresponds to a different sound source with M as the number of sounds sources. Denote $\hat{S}_n = (\hat{\mathbf{s}}_{1,n}, ..., \hat{\mathbf{s}}_{i,n}, ..., \hat{\mathbf{s}}_{M,n})$ as the matrix containing the estimated sound sources that is obtained for example after blind source separation. Moreover, $\mathbf{w_n} = (w_{1,n}, ..., w_{i,n}, ..., w_{M,n})^{\intercal}$ is the vector of gain control weights with $w_{i,n}$ as the gain of the i^{th} estimated sound source; and e_n is the vector of EEG measurements which is collected during the n^{th} trial. Note that according to this notation, $A_n = i$ indicates that the attention is on the i^{th} sound source. Initially, the system presents all sound sources with equal weights, but as e_n is observed and EEG evidence, ε_n is extracted based on e_n and $\hat{S}_n$, the posterior distribution of the attention is updated with the observed EEG evidence and the AGC module updates the weights of the sound sources as a function of this updated posterior distribution. We assume that the updated weight, in general, is a function of past and present weights and attention posterior distributions.

Since, in this paper, we do not focus on developing strategies for how weights should be controlled, as in previous work [30], the experiments below specifically use weights that are instantaneously obtained by applying a saturating linear function to the current attention posterior distribution.

$$\mathbf{w_{n+1}} \propto p(A_n|A_{n-1}, \varepsilon_n, \mathbf{w_n}, \mathbf{c}), \tag{2}$$

with a limitation factor on $\mathbf{w_{n+1}}$,

$$w_{i,n+1} = \begin{cases} w_{max} & \text{if } p(A_n = i|A_{n-1}, \varepsilon_n, \mathbf{w_n}, \mathbf{c}) \geq w_{max} \\ w_{min} & \text{if } p(A_n = i|A_{n-1}, \varepsilon_n, \mathbf{w_n}, \mathbf{c}) \leq w_{min} \end{cases}$$

These limitations on weight range were imposed to ensure the audibility of all sources, to enable mistake correction in the event of algorithm/human errors, and to allow shifting attention if desired.

C. Signal Modeling for Feature Extraction

We propose to use a signal model that assumes EEG is a function of the estimated sound sources. This model is inspired by previous results that indicated there is high correlation between EEG and envelopes of the (attended and unattended) sounds sources at various time lags [30]. Specifically, EEG measurements were correlated with both attended and unattended speech envelopes. Correlation patterns were similar except that the correlation between the EEG and unattended source was delayed and had a lower amplitude compared to the correlation between the EEG and attended source. Based on these observations, we suggest the following signal model:

$$e_n(t) = w_{i,n}\hat{\mathbf{s}}_{i,n} * h^i_{+,n}(t) + \sum_{j \neq a} w_{j,n}\hat{\mathbf{s}}_{j,n} * h^j_{-,n}(t) + n(t) \tag{3}$$

where sound source index i is attended to, while others $j \neq i$ are unattended, $h^j_{-,n}(t)$ is the impulse response function for unattended sound source with index j and $h^i_{+,n}(t)$ is the impulse response function for the attended source. As defined previously, $\hat{\mathbf{s}}_{i,n}$ is the estimated speech envelope of the attended source and $\hat{\mathbf{s}}_{j,n}$ is that of unattended source j. Finally, we assume that $n(t)$ is (temporally) white Gaussian noise with zero mean and covariance $E(nn^T) = \sigma^2\ I$.

D. Feature Extraction and Classification

This section describes how to extract EEG evidence, ε using the proposed signal model and EEG calibration data. Specifically, here we discuss how to learn the class/sound-source conditional EEG evidence distribution, $P(\varepsilon_n|A_n)$, which is used by the Auditory Attention Inference module to compute the posterior distribution of the attention as described in (1). Assuming that there are N sources, first, using leave-one-out cross validation over the calibration data, we learn the impulse responses $\widehat{\mathbf{h}_i} = [\mathbf{h}^1_-, .., \mathbf{h}^i_+, ..., \mathbf{h}^j_-, ..., \mathbf{h}^N_-]$ through least-square estimation, assuming that the i^{th} source is the attended or target and the others $j \neq i$ are the unattended or distractors. For a two speaker scenario, $\widehat{\mathbf{h}_1} = [\mathbf{h}^1_+, \mathbf{h}^2_-]$ and $\widehat{\mathbf{h}_2} = [\mathbf{h}^1_-, \mathbf{h}^2_+]$ are estimated assuming that speaker-1 and speaker-2 are the attended sources, respectively. The details of the estimation are given in Appendix A. These impulse response functions are then used in (3) to estimate the EEG, and the correlation coefficient between measured (and preprocessed/filtered) EEG and estimated EEG (from the signal model assuming i^{th} source being target, $\widehat{\mathbf{h}_i}$) are calculated for each EEG channel. Accordingly, in a two speakers scenario, we defined $x^{ch} = [\boldsymbol{\rho}_{\mathbf{e^{ch}},\hat{\mathbf{e}}^{\mathbf{ch}}_{\mathbf{A=1}}}, \boldsymbol{\rho}_{\mathbf{e^{ch}},\hat{\mathbf{e}}^{\mathbf{ch}}_{\mathbf{A=2}}}]^{\intercal}$ as a 2×1 dimensional vector for each channel with $\boldsymbol{\rho}_{\mathbf{e^{ch}},\hat{\mathbf{e}}^{\mathbf{ch}}_{\mathbf{A=i}}}$ for $i = 1, 2$ as the correlation between the estimated and raw EEGs when the i^{th} source is the attended source. Assuming that the number of the EEG channels is N_c, final feature vector for each trial, $\mathbf{x} = (x^1, ..., x^{ch}, ..., x^{N_c})^{\intercal}$ is a 32×1 dimensional vector resulting from concatenation of all the features of all EEG channels.

Once the feature vector is extracted for each trial, for dimensionality reduction, we use the regularized discriminant analysis (RDA). The RDA defines a quadratic projection of high dimensional features to a one-dimensional evidence. Assuming that there are only two sound sources, if the distributions of the features for both classes (EEG features corresponding to the attended and unattended sound sources) were Gaussians, then this projection would be the result of log-likelihood ratio which optimizes the Bayesian Risk. For each EEG feature vector, we denote the RDA projections as $s_{RDA}(\mathbf{x})$, and use these projected values as the EEG evidence, ε. The details of the RDA projection is provided in our previous work [30]. Once the EEG evidence is extracted for class i; that is when the i^{th} sound source is the target (i.e., $A = i$), using kernel density estimation we learn $p(\boldsymbol{\varepsilon} = \epsilon|A = i)$. We use a Gaussian kernel the bandwidth of which is estimated using Silverman's rule of thumb.

IV. Experimental Results

All analyses are performed on data from calibration (source energies equal in each trial) and online (source energies in each trial are modulated based on attention inference) sessions. The calibration session data will also be referred to as the equal-energy/weight dataset (DT1); and the online session data as the modulated-energy/weight dataset (DT2). Within equal-energy or modulated-energy datasets, 5-fold cross-validation is employed to estimate classification performance, which is quantified using area-under-ROC-curve (AUC). Classification performance using AUC by calibrating (training) with the equal-energy dataset and testing on the modulated-energy dataset is also analyzed. In all analyses both cross-correlation (CC) and signal model (SM) based features are used. For both CC and SM features, parameters for the extraction method are optimized as explained in Appendix B.

A. Single-Channel Classifier Is Competitive With the 16-Channel Classifier

Using the selected order of h, SM feature vector x^{ch} is formed as described in Section III-D. Using these

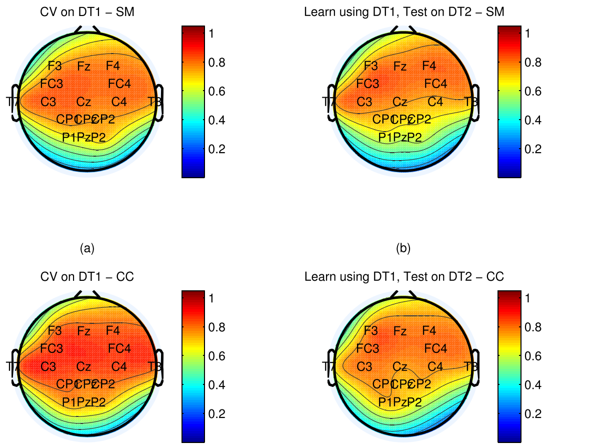


Fig. 4. Topographic map of classification performance over the scalp for classifying attended versus unattended speakers, median values over all participants. (a). cross-validation on dataset-1 or equal weights dataset using SM and CC features. (b). model trained using dataset-1 and validated on dataset-2 or variable weights dataset for SM and CC features.

TABLE I
AUC FOR CLASSIFICATION USING SINGLE BEST CHANNEL VERSUS ALL 16 CHANNELS FOR EACH SUBJECT

Participant	Best (CC)	All (CC)	Best (SM)	All (SM)
1	Fz: 0.94	0.92	F3: 0.92	0.93
2	C3: 0.93	0.89	C3: 0.92	0.91
3	C4: 0.99	0.97	C4: 0.99	0.97
4	Fc3: 0.87	0.87	Fc4: 0.81	0.92
5	F4: 0.88	0.93	C4: 0.71	0.64
6	T7: 0.92	0.88	Cz: 0.88	0.87
7	C3: 0.84	0.84	T7: 0.82	0.84
8	C4: 0.90	0.88	F4: 0.88	0.92
9	CPz: 0.95	0.94	F4: 0.94	0.93
10	C3: 0.89	0.84	C3: 0.82	0.93

EEG-channel-specific features individually, attended source posteriors are evaluated. Figure 4 (a) shows the median 5-fold cross-validation AUC (across 10 participants) for each EEG channel in the form of a topographical map in equal-energy dataset (dataset 1). Figure 4 (b) shows the median AUC resulting from learning the model using the equal-energy dataset (dataset 1) and test it on the modulated-energy dataset (dataset 2).

To complement these channel-specific results, Table I shows the 5-fold cross-validation AUC for CC and SM features on equal-energy dataset (DT1) when using the best channel and when using all 16 channels. This table indicates that while best scalp locations for EEG acquisition may be subject specific, a small number of well-positioned electrodes can be competitive with respect to many channels, and this is a significant practical consideration to be explored further.

B. AUC Improves With Increasing Trial Length

Different lengths of EEG is used to infer the attended speech source by varying the length of EEG used from 2 seconds to 20 seconds. Figure 5 shows the median classifier AUC (across 10 participants) when using all 16 channels, with both CC and SM features; shaded areas are 90% confidence intervals.

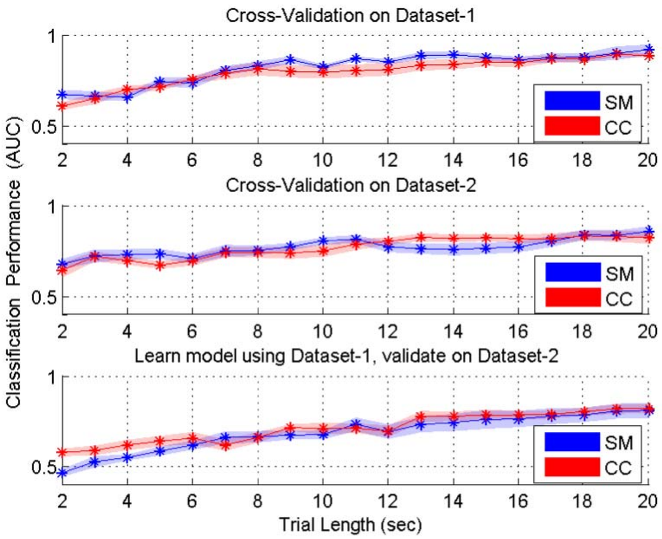


Fig. 5. Attended speech classifier AUC increases with trial length using both CC and SM features, according to 5-fold cross-validation in both (top) equal- and (middle) modulated-energy datasets, as well as (bottom) when calibrating with equal-energy trials and testing on modulated-energy trials. Median curves and 90% confidence intervals are shown.

As expected, accuracy increases when longer EEG evidence windows (more samples) are used for inference.

C. Compensating for Sound Source Power Variation Improves Classification Accuracy

The assumption that EEG-evidence statistics are invariant under changing relative power of input sources is not realistic. In calibration, energy of both speech sources were kept equal in each trial. However, in online sequences where source amplitudes are modulated based on attention inference posterior, the relative energies of sources deviate from being equal in each trial. Testing a classifier calibrated with equal-energy dataset on online dataset when trial energy is modulated by weights controlled by inference posteriors, the EEG feature distribution drifts away from the calibration model and as expected results in a performance drop. One could design a calibration session that explores various source energy levels in trials to sample this parameter space sufficiently, but in practice this may lead to prohibitively long calibration sessions and will be undesired. A model based approach would enable corrections for these effects to be incorporated into feature extraction, therefore calibration procedures can be reasonable in terms of time and effort requirements, while performance in online use can be maintained.

Figure 6 shows, with data pooled from the online (modulated-energy) datasets of the population of 10 participants, the probability of attending to male or female speakers as estimated by the classifier (using SM or CC features) versus the modulation weight for the corresponding trial. The average of these probabilities in weight-bins, along with the best line fits indicate that there is a dependency of the classifier estimates of attention probabilities on source energy. Considering this effect of modulation weights on classifier assessment of EEG is expected to improve performance.

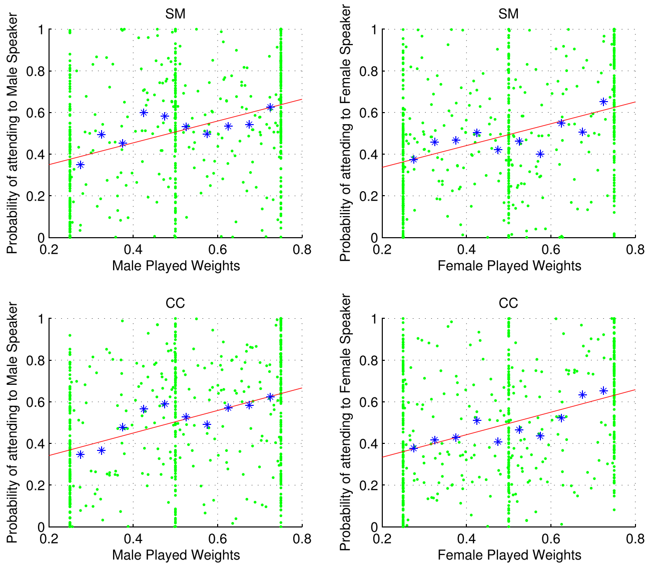


Fig. 6. For SM (top) and CC (bottom) features, the (estimated) probability of attending to the male speaker (left) versus the female speaker (right) versus modulation weight of the respective speaker in each trial, presented as a (green) scatter plot. The concentration of weights at 0.25, 0.5, and 0.75 are due to initial weights being equal at 0.5 and lower and upper bounds of the weights being set to $w_{min} = 0.25$ and $w_{max} = 0.75$, respectively. The (blue) stars indicate average probabilities in one of eleven weight intervals, while the (red) lines indicate best linear least square fits to these averages.

As a simple first-order approximation, using $p(A_n = i|\mathbf{w}_n) = a w_{i,n} + b$ derived from the linear least squares fits indicated above, the dependence of attention on source weights is approximately corrected in this preliminary study using $p(\varepsilon_n|A_n)p(A_n|\mathbf{w}_n)$. Clearly, this is not a proper posterior, since the likelihood of attention given weight is not a proper probability distribution function; further modeling effort must be put into this component in the future. With this simplistic corrective action, classification AUC improves in a statistically significant fashion, as demonstrated in Figure 7.

D. Population Classifier Is Competitive With Individual-Specific Classifiers

Subject to subject generalization of classification accuracy is a desirable feature in brain interfaces, as it would allow pooling data from multiple users and calibrating a population classifier, or using such a population model as the prior for an individual classifier, in order to eliminate or reduce calibration time for a new user. Using leave-one-subject-out cross-validation scheme, the performance of a population classifier is tested. Training the classifier using calibration (equal-energy) data from all subjects but one, and evaluating the equal-energy data from the left-out subject reveals that such subject to subject generalization is feasible. Validation AUCs of population classifiers obtained by leaving each subject out during calibration are reported in Table II. In contrast, as an estimate of individual-specific classifier performance, 5-fold cross-validation AUCs for classifiers calibrated on equal-weight datasets for each participant are evaluated and included in the table. The process is repeated with SM and CC features for comparison. While we acknowledge that this process focused only on equal-weight calibration datasets, these results could be considered as an upper bound of performance expectations when models are calibrated on equal energy trials and tested online with modulated-energy trials, based on previously discussed results. Qualitatively, the results indicate that population data is representative of individual data, and a hierarchical model that considers the individual as an instance of the population could lead to a calibration procedure that allows effective evidence pooling from multiple users to reduce calibration time or improve calibration quality for a particular individual.

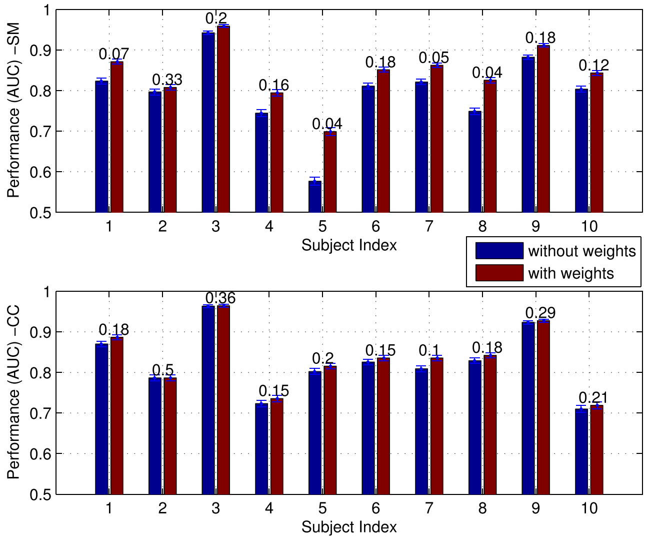


Fig. 7. AUC and 90% confidence intervals with/without considering the linear approximate $p \triangleright A_n|\mathbf{w}_n \triangleleft$ correction for both SM and CC features. The numbers above the two bars for each subject indicate the p-value for the null hypothesis that the AUC after correction is less than or equal to the AUC before correction. For SM features, the linear corrective term improves AUC in a statistically significant fashion for almost all subjects, while for CC features, this is not the case. This result indicates that the signal model based approach may be improved further and generalize with appropriate corrective terms superimposed on calibration models with limited source energy variability.

TABLE II
PERFORMANCE (AUCS) OF THE POPULATION CLASSIFIER VERSUS INDIVIDUAL CLASSIFIER ON EQUAL WEIGHTS DATASET (DATASET-1)

Participant AUC	1	2	3	4	5	6	7	8	9	10
Population classifier (SM)	0.86	0.78	0.97	0.85	0.72	0.84	0.78	0.86	0.68	0.81
Individual classifier (SM)	0.93	0.91	0.97	0.92	0.64	0.87	0.84	0.92	0.93	0.93
Population classifier (CC)	0.93	0.94	0.95	0.88	0.85	0.85	0.87	0.89	0.86	0.90
Individual classifier (CC)	0.92	0.89	0.97	0.87	0.93	0.88	0.84	0.88	0.94	0.84

V. CONCLUSION

EEG has been demonstrated to exhibit useful evidence regarding auditory attention in the presence of multiple competing speech waveforms. A data-driven cross-correlation feature and a signal model based feature is considered in two scenarios. The results demonstrate that dense electrode arrays as used in previous reports are not necessary; competitive binary auditory attention classification results with at most 20 seconds of EEG from 16 channels, or even a single well-positioned channel can be obtained. It is also shown that a model calibrated using equal-energy speech waveforms can perform well in closed-loop unbalanced-energy speech waveform conditions, where the speech amplitudes are modulated by the estimated attention posterior probability distribution. Further analysis demonstrate that such a model would perform even better if it is corrected (in this case, linearly) to account for EEG evidence dependency on speech weights in mixture. Finally, results indicate that calibrating a model based on population EEG could result in acceptable performance for new individuals/users; more interestingly, population data can be pooled to form a prior model for individual classifiers, thereby reducing calibration time significantly for new users. The results presented in this paper contribute to the field of auditory-attention driven manipulation of auditory scenes in hearing aid and virtual/augmented reality applications.

Appendix A
Signal Model Derivations

In this appendix we demonstrate the estimation method for the model parameter h as defined in equation (3). Define L as the order of auto-regressive signal model, N as the number of time samples in each trial (both for sound source envelopes and EEG signal), e as EEG time samples, and s^1 and s^2 as the sound sources envelope time samples for source 1 and 2 respectively. Then according to our model we can define,

$$\underset{\mathbf{e}}{\begin{pmatrix} e_{L+1}\\ e_{L+2}\\ \cdot\\ \cdot\\ \cdot\\ e_N \end{pmatrix}} = \underset{S^1}{\begin{pmatrix} s_1^1 & s_2^1 & \cdot\ \cdot & s_L^1\\ s_2^1 & s_3^1 & \cdot\ \cdot & s_{L+1}^1\\ \cdot & \cdot & \cdot\ \cdot & \cdot\\ \cdot & \cdot & \cdot\ \cdot & \cdot\\ \cdot & \cdot & \cdot\ \cdot & \cdot\\ s_{N-L}^1 & \cdot & \cdot\ \cdot & s_N^1 \end{pmatrix}} \underset{\mathbf{h}^1}{\begin{pmatrix} h_1^1\\ h_2^1\\ \cdot\\ \cdot\\ h_L^1 \end{pmatrix}}$$

$$+ \underset{S^2}{\begin{pmatrix} s_1^2 & s_2^2 & \cdot\ \cdot & s_L^2\\ s_2^2 & s_3^2 & \cdot\ \cdot & s_{L+1}^2\\ \cdot & \cdot & \cdot\ \cdot & \cdot\\ \cdot & \cdot & \cdot\ \cdot & \cdot\\ \cdot & \cdot & \cdot\ \cdot & \cdot\\ s_{N-L}^2 & \cdot & \cdot\ \cdot & s_N^2 \end{pmatrix}} \underset{\mathbf{h}^2}{\begin{pmatrix} h_1^2\\ h_2^2\\ \cdot\\ \cdot\\ h_L^2 \end{pmatrix}} + \underset{n}{\begin{pmatrix} n_{L+1}\\ n_{L+2}\\ \cdot\\ \cdot\\ \cdot\\ n_N \end{pmatrix}}$$

Then, the model at n^{th} trial is $\mathbf{e}_n = C_n\mathbf{h} + n_n$, where,

$$C_n = \begin{bmatrix} S_n^{1^T} & S_n^{2^T} \end{bmatrix} \text{ and } \mathbf{h} = \begin{bmatrix} \mathbf{h}^1\\ \mathbf{h}^2 \end{bmatrix}.$$

Afterward, we define the following optimization problem to estimate the parameter values at which the sum of squared estimation error in all trials is minimized.

$$\widehat{\mathbf{h}} = \arg\min_{\mathbf{h}} \sum_n \|\mathbf{e}_n - C_n\mathbf{h}\|_2^2$$

The solution to our convex optimization problem is

$$\widehat{\mathbf{h}} = \sum_n (C_n^\top C_n)^{-1} \sum_n C_n^\top e_n.$$

In our model we propose that model parameter values are dependent on target sound source hence, we solve the above optimization problem two times to estimate the $\widehat{\mathbf{h}}$: (1) using trials in which speaker 1 is target ($\widehat{\mathbf{h}_1}$) and (2) using trials in which speaker 2 is target ($\widehat{\mathbf{h}_2}$).

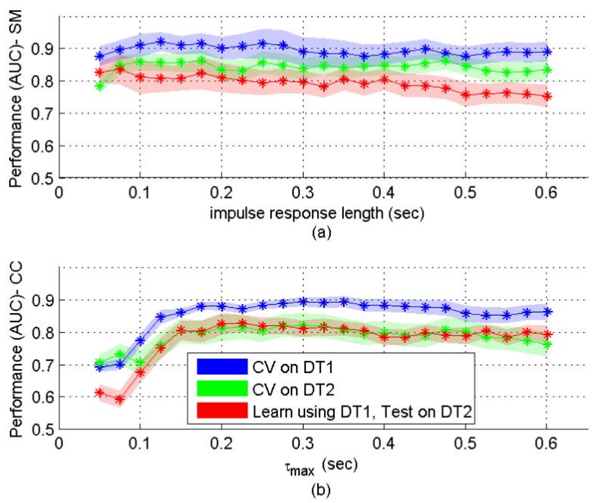


Fig. 8. Median of classification performance (AUCs) over participants versus; (a) impulse response length of the SM model (h length in seconds or L samples) (b) τ_{max} in seconds for extracting CC features in $\mathcal{T}0, \tau_{max}\mathcal{U}$ range.

Appendix B
Model Order Selection

The model order L of our proposed model in Section III-C, needs to be optimized for best performance. Moreover, we need to define the optimum time window duration $[0, \tau_{max}]$ for CC features. In this appendix we present the effect of these parameters on classification performance using cross validation in Figure 8. We believe that the comparison of different methods in our manuscript is fair because the presented results are obtained at optimum values of these parameters, according to the similar performance gains. More specifically, we selected the parameter values for which the average of AUCs over three types of analysis shown in Figure 8 are maximized.

In Figure 8, red curves show median of cross-validation performances on dataset-1 over all participants. Shaded areas around each curve are showing the corresponding standard error for that curve. Green curves show median of cross-validation performances on dataset-2 over all participants. And blue curves are showing the result of applying the learned model using dataset-1, on dataset-2.

Based on the explained criteria, impulse response length is set to 0.125 seconds ($L = 32$) and $\tau_{max} = 0.275$ seconds.

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[2] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica United Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[3] W. A. Yost, "The cocktail party problem: Forty years later," in *Binaural and Spatial Hearing in Real and Virtual Environments*. 1997, ch. 17, pp. 329–347.

[4] W. A. Yost and S. Sheft, "Auditory perception," in *Human Psychophysics*. New York, NY, USA: Springer, 1993, pp. 193–236.

[5] A. W. Bronkhorst, "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," *Attention, Perception, Psychophys.*, vol. 77, no. 5, pp. 1465–1487, 2015.

[6] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends Cognit. Sci.*, vol. 12, no. 5, pp. 182–186, 2008.

[7] B. G. Shinn-Cunningham and V. Best, "Selective attention in normal and impaired hearing," *Trends Amplif.*, vol. 12, no. 4, pp. 283–289, 2008.

[8] L. Dai and B. G. Shinn-Cunningham, "Contributions of sensory coding and attentional control to individual differences in performance in spatial auditory selective attention tasks," *Frontiers Human Neurosci.*, vol. 10, no. 10, p. 530, 2016.

[9] A. K. Lee *et al.*, "Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch," *Frontiers Neurosci.*, vol. 6, p. 190, Jan. 2013.

[10] J. C. Middlebrooks and P. Bremen, "Spatial stream segregation by auditory cortical neurons," *J. Neurosci.*, vol. 33, no. 27, pp. 10986–11001, 2013.

[11] J. Dong, H. S. Colburn, and K. Sen, "Cortical transformation of spatial processing for solving the cocktail party problem: A computational model," *Eneuro*, vol. 3, no. 1, p. 0086, 2016.

[12] K. T. Hill and L. M. Miller, "Auditory attentional control and selection during cocktail party listening," *Cerebral cortex*, vol. 20, no. 3, pp. 583–590, 2009.

[13] A. K. C. Lee, E. Larson, R. K. Maddox, and B. G. Shinn-Cunningham, "Using neuroimaging to understand the cortical mechanisms of auditory selective attention," *Hearing Res.*, vol. 307, pp. 111–120, Jan. 2014.

[14] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: Functional roles and interpretations," *Frontiers Human Neurosci.*, vol. 8, p. 311, 2014.

[15] S. J. Aiken and T. W. Picton, "Human cortical responses to the speech envelope," *Ear Hearing*, vol. 29, no. 2, pp. 139–157, 2008.

[16] Y.-Y. Kong, A. Mullangi, and N. Ding, "Differential modulation of auditory responses to attended and unattended speech in different listening conditions," *Hearing Res.*, vol. 316, pp. 73–81, Oct. 2014.

[17] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich, "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 23, pp. 13367–13372, 2001.

[18] H. Luo and D. Poeppel, "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," *Neuron*, vol. 54, no. 6, pp. 1001–1010, 2007.

[19] D. A. Abrams, T. Nicol, S. Zecker, and N. Kraus, "Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech," *J. Neurosci.*, vol. 28, no. 15, pp. 3958–3965, 2008.

[20] K. V. Nourski *et al.*, "Temporal envelope of time-compressed speech represented in the human auditory cortex," *J. Neurosci.*, vol. 29, no. 49, pp. 15564–15574, 2009.

[21] B. N. Pasley *et al.*, "Reconstructing speech from human auditory cortex," *PLoS Biol.*, vol. 10, no. 1, p. e1001251, 2012.

[22] J. A. O'Sullivan *et al.*, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2014.

[23] B. Mirkovic, S. Debener, M. Jaeger, and M. De Vos, "Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications," *J. Neural Eng.*, vol. 12, no. 4, p. 046007, 2015.

[24] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *J. Neurophysiol.*, vol. 107, no. 1, pp. 78–89, 2012.

[25] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[26] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, May 2017.

[27] S. Akram, A. Presacco, J. Z. Simon, S. A. Shamma, and B. Babadi, "Robust decoding of selective auditory attention from meg in a competing-speaker environment via state-space modeling," *NeuroImage*, vol. 124, pp. 906–917, Jan. 2016.

[28] C. Horton, R. Srinivasan, and M. D'Zmura, "Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party,'" *J. Neural Eng.*, vol. 11, no. 4, p. 046015, 2014.

[29] M. Haghighi, M. Moghadamfalahi, H. Nezamfar, M. Akcakaya, and D. Erdogmus, "Toward a brain interface for tracking attended auditory sources," in *Proc. IEEE 26th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2016, pp. 1–5.

[30] M. Haghighi, M. Moghadamfalahi, M. Akcakaya, and D. Erdogmus. (Nov. 2016). "EEG-assisted modulation of sound sources in the auditory scene." [Online]. Available: https://arxiv.org/abs/1612.00703

Marzieh Haghighi received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2011. She is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, Northeastern University, and a Research Assistant with the Cognitive Systems Laboratory, Boston, MA, USA. Her research interests include brain–computer interfaces, machine learning, and statistical signal processing for biomedical and neural applications.

Mohammad Moghadamfalahi (S'14) received the B.Sc. degree in electrical engineering from Amirkabir University, Tehran, Iran, in 2008, and the Ph.D. degree in electrical engineering from Northeastern University, Boston, MA, USA, in 2016.

He is currently with Honeywell Laboratories as a Research and Development Engineer. His current research interests are signal processing and machine learning for neural and aerospace applications.

Murat Akcakaya (S'07–M'12) received the B.Sc. degree from the Electrical and Electronics Engineering Department, Middle East Technical University, Ankara, Turkey, in 2005, and the M.Sc. and Ph.D. degrees in electrical engineering from Washington University in St. Louis, MO, in 2010, respectively. He is currently an Assistant Professor with the Electrical and Computer Engineering Department, University of Pittsburgh. His research interests are in the area of statistical signal processing and machine learning with applications in non-invasive brain computer interface design. He received the student paper contest awards at the 2010 IEEE Radar Conference, the 2010 IEEE Waveform Diversity and Design Conference, and the 2010 Asilomar Conference on Signals, Systems and Computers.

Barbara G. Shinn-Cunningham received the B.Sc. degree in electrical engineering from Brown University and the master's and Ph.D. degrees in electrical and computer engineering from the Massachusetts Institute of Technology. She was with Bell Communications Research, MIT Lincoln Laboratory, and Sensimetrics. She joined the Faculty at Boston University. She is an Auditory Neuroscientist best known for her work on attention and the cocktail party problem, sound localization, and the effects of room acoustics and reverberation on hearing.

Deniz Erdogmus received the B.S. degrees in EE and mathematics and the M.S. degree in EE from the Middle East Technical University, Turkey, in 1997 and 1999, respectively, and the Ph.D. degree in ECE from the University of Florida in 2002. He was a Post-Doctoral Research Associate with the University of Florida until 2004. He was an Assistant Professor of Biomedical Engineering with Oregon Health and Science University until 2008. Then, he joined Northeastern University, where he is currently an Associate Professor with the Electrical and Computer Engineering Department. His research focuses on statistical signal processing and machine learning with applications to contextual signal/image/data analysis with applications in cyberhuman systems, including brain–computer interfaces and technologies that collaboratively improve human performance. He has served as an Associate Editor and a Program Committee Member for a number of journals and conferences in these areas, including IEEE SIGNAL PROCESSING LETTERS, and the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, and the IEEE TRANSACTIONS ON NEURAL NETWORKS.