I. Cocktail Party Experiment

Daniel D.E. Wong, Enea Ceolini, Denis Drennan, Shih-Chii Liu, Alain de Cheveigné

MOTIVATION

In past years at the Telluride Neuromorphic Workshop, work has been done to develop EEG decoding methods to classify measures of auditory attention. Specifically, these were projects involving the classification of the attended speech envelope [O'Sullivan 2015] and the direction of a perceived sound source [Wong 2016]. The goal of the present experiment was to increase the complexity of the listening conditions. This will bring our experiments a step closer to a real cocktail party situation and to start to combine these projects into a practical application: brain-controlled acoustic processing for a hearing aid. The previous classification of the attended speech envelope did not involve the individual subject switching their attention, and was not performed in free-field. The previous localization experiment was only performed with attention to a single speaker, as such whether the direction of an attended sound source in a cocktail party environment can be decoded is still unknown. The applicability of these decoding methods to microphone array steering was explored.

METHODS

In this experiment, two Jules Verne stories were presented at the same time, each coming from one of two speakers positioned at ~ ±45 degrees azimuth: Journey to the Center of the Earth and Twenty Thousand Leagues Under the Sea. The subject was asked to listen to the right speaker on odd trials and the left speaker on even trials. Every two trials, the stories were swapped from one speaker to the other. This trial order was designed to avoid confounding speech envelope decoding with speaker location decoding, and speaker location decoding with talker identity decoding. In all 50 trials were recorded, each lasting 60 s. At the same time, EEG was recorded from the subject. Only a single subject was recorded. After the EEG recording, an array of 8 microphones were positioned around the room to record frequency sweeps presented through the speakers.

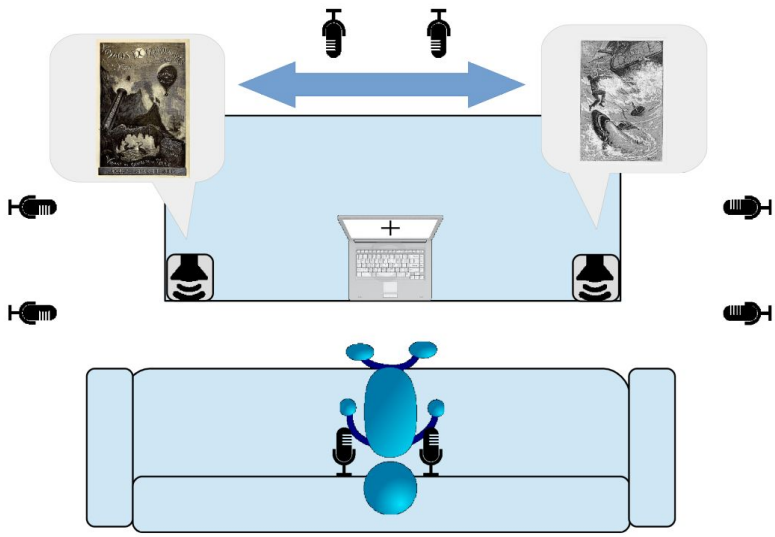


Figure 1. Experiment setup.

The following three sub-projects use this data to to tackle different aspects of implementing a hearing device that can be cognitively steered.

1. Multimicrophone processing
    a. Simulated data for sound separation and decoding attention
    b. Real data
2. Attended envelope classification
3. Classification of direction

IIa. Multi-Microphone Processing - Simulated Microphone Data

Sahar Akram and Behtash Babadi

MOTIVATION

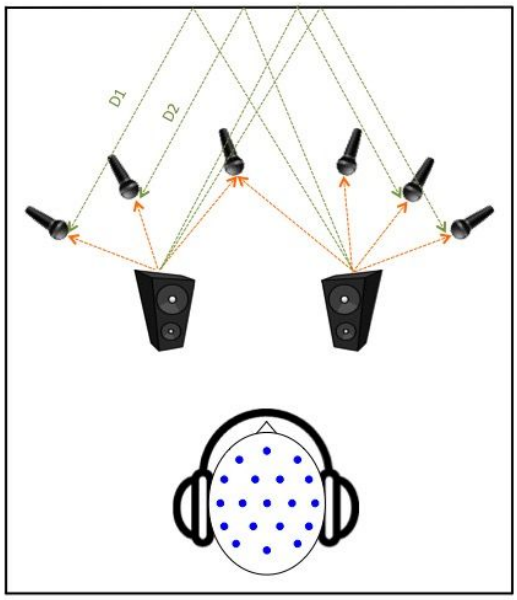
The goal of this study is to develop an auditory source segregation framework that is controlled by attention state of a listener. We are trying to answer (some of the) following questions:

1. Can we use ICA (or any other BSS techniques) to reliably recover the envelope of individual speeches from the speech mixtures?
2. Can we use the estimated envelopes, instead of the envelopes computed from the clean speeches to decode the attentional state?
3. Assuming that the ICA technique works well and we can help the listener attend to the speaker of interest (e.g. speaker 1), is there a way to facilitate attention switching to the second speaker?

METHODS

Subjects are required to listen to an audio mixture consisting of two talkers and attend to one of the talkers for a certain period of time, while their EEG signal is being recorded. Audio signals are played dichotically through headphones.

Simulated microphone signals were obtained by applying different delays and attenuation factors to the clean audio for each microphone, modelling the direct speaker-microphone path and the first reflection. The simulation modelled 2 speakers and 3 microphones with random delays ranging from 0-20 ms.



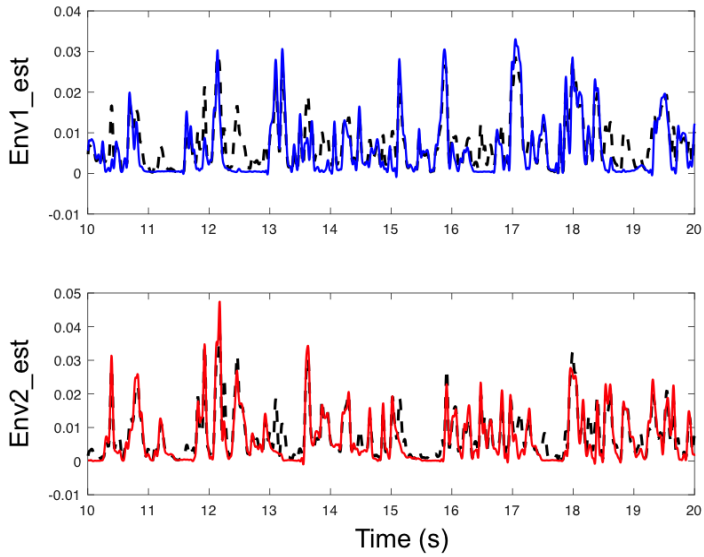
Speech Segregation: The first step is to segregate the two speech signals from the mixtures recorded by the simulated microphone array, using a BSS technique. Here we used, Fast-ICA (cite), Infomax (cite), ML-corrected ICA (cite), Time-Frequency-Masking ICA (cite), and M-NICA (cite).
All these techniques worked reasonably well in demixing the speech mixture and recover the original speech waveforms with 80-90% accuracy (correlation analysis) in simulated data. Delays and approximate impulse response function of a sample room are used for generating the speech mixtures in this simulation study. In the following equation, S1 and S2 are the sources of interest and M1, M2, and M3 are the mixed signals from three microphones in the room.

$$\begin{bmatrix} S_1 \\ S_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} * \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix}$$

We further apply a Hadamard transform to S1 and S2 to make the two signal more uncorrelated.

$$\hat{S}_1 = \frac{S_1 + S_2}{2} \qquad \hat{S}_2 = \frac{S_1 - S_2}{2}$$

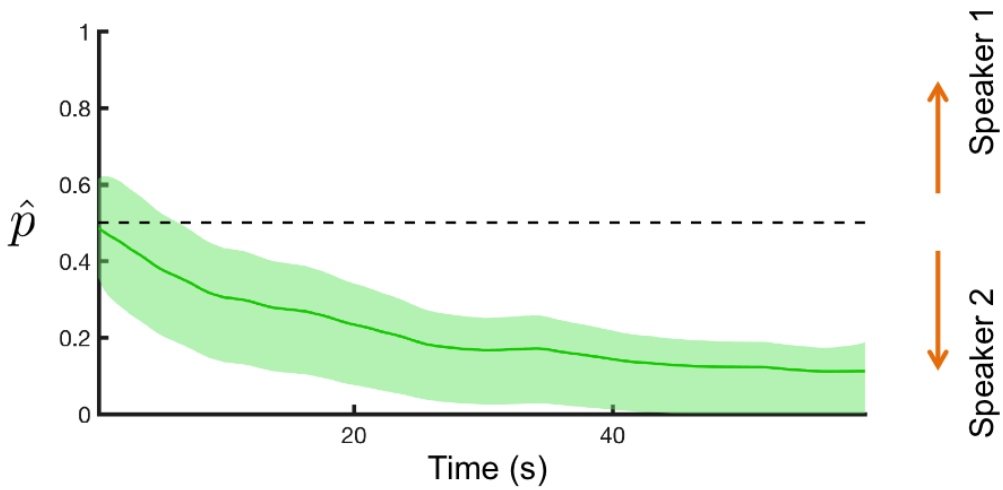
Figure below shows the original speech envelopes from the first and second speakers (blue and red solid curves, respectively), and those computed from the estimated sources above (black dashed-line) for each of the two speakers and for the first 20 ms of the trial. Correlation values between the original and estimated envelopes are .95 and .87, respectively for the first and second speakers.



Attention Decoding: The next step uses the recovered speech signals in an attention-decoding algorithm to estimate the attention state of the listener from the recorded EEG. Here, we have used state-space attention-decoding algorithm (cite) to obtain the probability of attending to speaker one as a function of time.

$$\hat{p} = \text{ATT-Decoder}(\hat{\text{Env}}_1, \hat{\text{Env}}_2, \text{EEG})$$

In this simulation study, we used the pre-recorded MEG data and the estimated envelopes from the previous step to perform the attention decoding. In this example, the listener attended to the second speaker and therefore the estimated probabilities of attending to speaker 1 are close to zero.



Adjusting Microphone Weights: The results of the attention decoding can be used to adjust the weights in the demixing matrix obtained from the ICA. The following equation can be used for computing time-varying demixing matrix that is changing with respect to the attentional state of the listener over time.

$$A_n := \begin{bmatrix} \hat{p}_n & 1-\hat{p}_n \end{bmatrix} * \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

Attended speech can therefore get extracted from the microphone recordings using the updated demixing matrix.

$$S_{Att} := \quad A_n * \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix}$$

The auditory files for the original and attention-modulated mixtures obtained using the described method are provided in the multimicrophone folder.

IIb. Multimicrophone Processing - Real Data

Daniel D.E. Wong, Sahar Akram, Behtash Babadi, Lucas Parra, and Alain de Cheveigné

MOTIVATION

In this section, various approaches were explored with the aim of separating a mixture of sound sources into their original streams. These streams can then be potentially used for EEG envelope decoding (section III), and eventually for acoustic feedback to the subject. ICA on simulated data was used previously as a proof-of-principle. Microphone data from the experiment in Section I is now used here to test several speech stream segregation algorithms under realistic conditions.

METHODS

The clean audio-to-microphone transfer function for each speaker was obtained by convolving the microphone recording of a frequency sweep with the spectral-power-inverse of the clean version [Müller 2001]. This transfer function, shown in Figure 1, was then used to recreate the microphone array signals that would have been recorded during the EEG experiment.

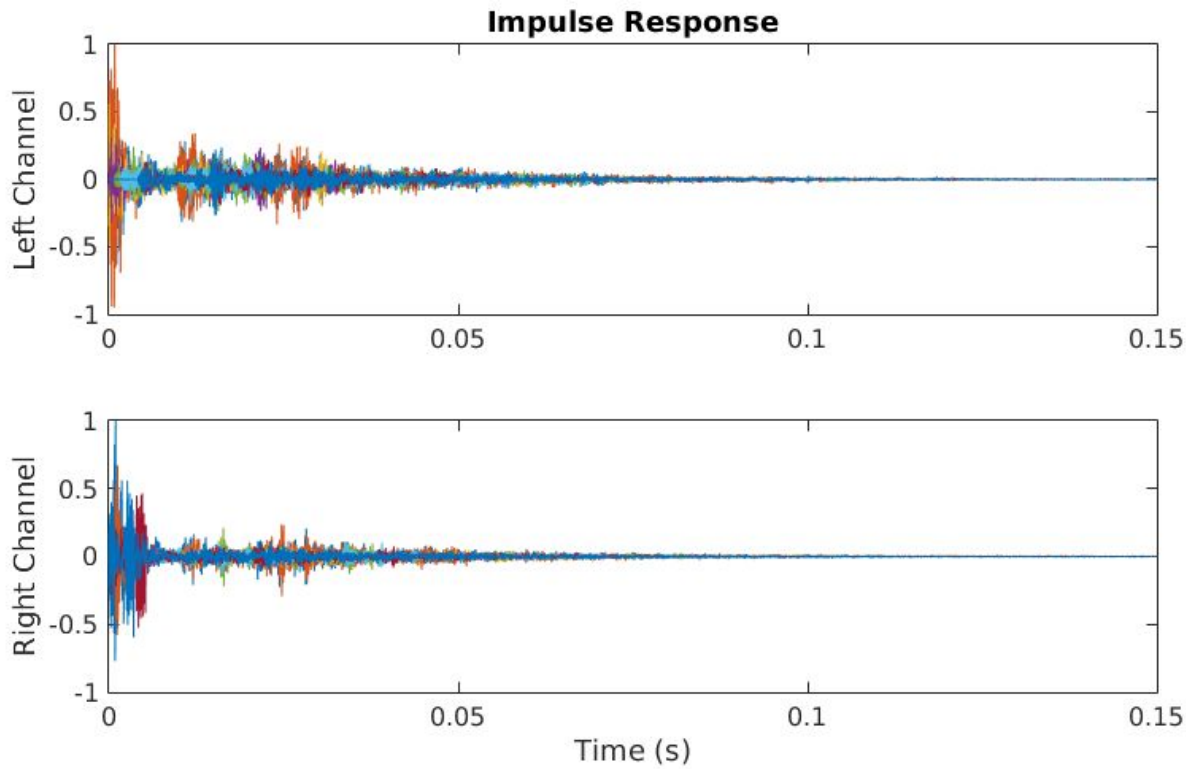


Figure 1: Speaker+room impulse responses from all 8 microphones for left and right (clean) audio channels.

Several algorithms were evaluated for separating the two talkers in the experiment using the microphone array:

A) Fast-ICA: This method attempts to find underlying independent components that contribute to the mixture, and was described in the previous section (IIa. Multimicrophone Processing - Simulated Data).

B) M-NICA on audio envelope: This method was used in [Van Eryndhoven 2016]. A caveat for this method is that the number of sources must be known. This number must be provided to M-NICA, which is anywhere from 1 to the number of sensors, N. This could potentially be addressed by performing M-NICA on all possible numbers of sources and then determining which of the resulting N*(N-1)/2 ICA components best matches the the EEG signal via an envelope decoding algorithm. However, the number of sources that can be handled by such an approach is limited to the number of sensors. This method of course only obtains the envelopes. Beamforming approaches are still required to estimate the separated sound source(s).

C) Linearly Constrained Minimum Variance (LCMV): This beamforming algorithm enforces a unit gain on a target source while minimizing the contribution of uncorrelated sources:

$$W^H L = I$$

$$min\ W^H R W$$

where $R = x^H x$ is the microphone signal covariance matrix, W is the weight matrix such that the source estimate $\hat{s} = W^H x$, and *L* is the source-to-microphone "forward" mapping [Van Veen 1988]. The minimization approach allows the algorithm to be more practical in a realistic situation where the number of sources may exceed the number of sensors. The Lagrangian solution to the minimization problem is:

$$W = L^H R^{-1} \left(L^H R^{-1} L \right)^{-1}$$

These calculations are performed in the frequency domain. The challenge with LCMV is that a sample of clean speech from the target source must be obtained in order to estimate *L*. One method that appears to be promising based on limited testing is to estimate the power in the residual components that were not part of the estimated source. Clean speech appears to have an estimated residual component that is about 3x smaller in initial experiments.

RESULTS

The envelopes of the separated source estimates and the clean audio were calculated by performing full-wave rectification and lowpassing at 8Hz. The correlation coefficient between the estimated sources and the clean audio are used as a measure of how cleanly the sources are separated.

A) FAST-ICA:

		Clean Speech	
		1	**2**
Estimated Sources	**1**	85.8 %	82.4
	2	84.6	83.2

B) M-NICA on Envelopes:

		Clean Speech	
		1	**2**
Estimated Sources	**1**	89.5 %	77.8
	2	56.7	88.6

C) LCMV:

		Clean Speech	
		1	**2**
Estimated Sources	**1**	94 %	57.7
	2	65.4	92.6

The correlation coefficient between the two clean speech envelopes was computed to be 59%.

DISCUSSION

ICA did not work as well with real data as it did with simulated data. LCMV beamforming worked the best; however, a caveat is that segments of clean speech are required for an ad-hoc array such as that used in the experiment in order to estimate the source-to-microphone forward mapping *L*. 0.5s appears to be sufficient to achieve correlation coefficients on the order of what was obtained in the results. Initial work showed that it may be possible to obtain these clean segments by estimating the residual signal of the beamformer. If this strategy can also be applied to individual frequency bands, it may be then be possible instead to just require a clean frequency band and interpolate the remaining frequency bands. An alternative strategy is to use a closely spaced sensor array to perform beamforming on different azimuths. Coupled with a voice activity detector (VAD), this method should be able to segregate speech streams provided that the speakers are sufficiently separated in space.

III. Attended Envelope Classification

Daniel D.E. Wong and Alain de Cheveigné

MOTIVATION

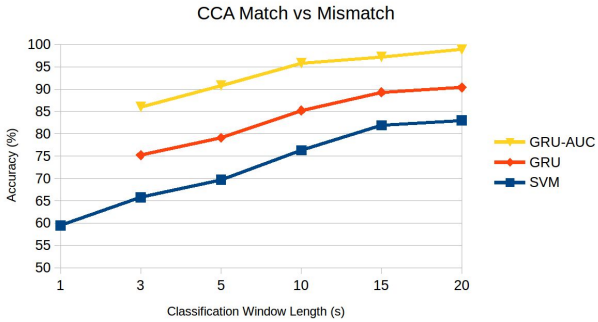
The aim of this sub-project was to classify which story the subject was attending to based on the relationship between the EEG and the envelope of the attended speech. The use of the paradigm described in Section I improves over past experiments by using free-field audio as opposed to dichotic stimuli, and avoids confounding attended envelope decoding versus sound location/talker identity by changing the attended location and talker throughout the experiment. By testing envelope decoding performance with the segregated audio from the microphone array, a better understanding can be gained of how different modules of the proposed cognitively steered hearing device will realistically interact.

METHODS

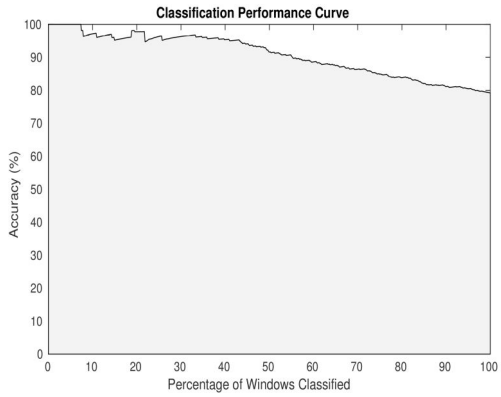
EEG and the clean audio streams were filtered into frequency bands using a log-frequency filter bank. Canonical correlation analysis was used to identify a subspace that maximized the correlation between the filter-banked EEG and attended audio stream. The correlation coefficients for the components, calculated over varying classification time windows, were used as classification features. A support vector machine (SVM) was trained on these features using a 3-1 training/testing split to classify:

a) the attended talker versus a random speech stream (clean speech) - Fig. 1.
b) the attended talker versus the unattended talker (clean speech) - Fig. 3.
c) the attended talker versus the unattended talker (LCMV beamformed audio) - Fig. 4.

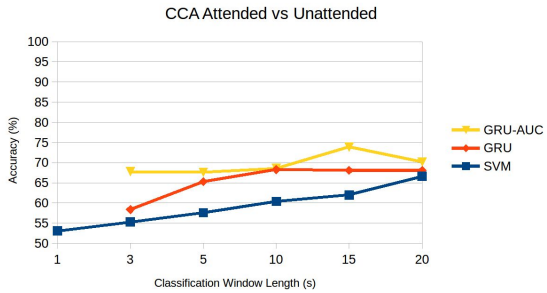
Improved accuracy could be achieved by dividing the time windows into 1s sub-windows and passing the features from these sub-windows into a gated recurrent unit (GRU) deep neural network. Additionally, if the discriminant value output of the classifier is thresholded so that some time windows are discarded (i.e. not classified), further accuracy improvement can be achieved. This tradeoff between the number of classified trials and classification accuracy can be described by the accuracy curve shown in Fig. 2. The area under this curve for the GRU classifier is shown as GRU-AUC in Figs. 1,3 and 4.



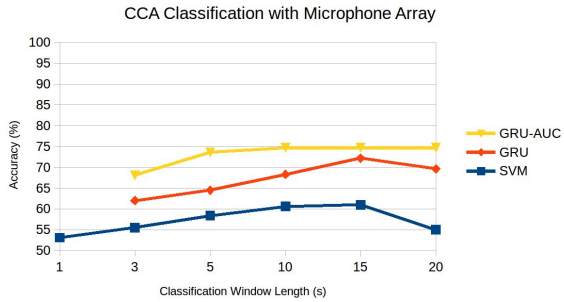
1



2



3



4

Figures 1-4: Classification performance.

DISCUSSION

Classification of attended-versus-unattended streams performed worse compared to match-versus-mismatch. The subject reported that it was difficult to maintain constant focus on individual streams due to the fact that two talkers were both male, and that English was not his first language. This could have potentially resulted in both streams being fairly well represented in the EEG, making envelope decoding difficult. Another explanation could be that classification of attended-versus-unattended streams is likely more sensitive to latency than match-versus-mismatch. Because only a single latency for all frequency bands was used for decoding, the best separation between attended and unattended classes may not have been achieved. Lastly, the EEG was quite noisy - robust PCA [Lin 2009] classified roughly half the EEG components as noise (and was thus not used for preprocessing). In short, there is some work to be done to improve the classification of attended-versus-unattended streams; however, it is promising that the classification performance is similar for both clean speech and segregated speech from the microphone array.

IV. Classification of Attended Sound Direction

Daniel D.E. Wong

MOTIVATION

The aim of this sub-project was to classify whether the attended audio was coming from the left or the right speaker. This expands on the experiment performed last year by using a competing talker instead of just a single talker. Classification of talker location would be useful for scenarios where a closely spaced array of microphones is used, allowing the array to be more easily steered based on azimuth.

METHODS

A support vector machine (SVM) classifier was designed to classify the location of the attended talker. The basic features used for the SVM were obtained using a variation of the filter-bank common spatial patterns (FBCSP) algorithm. EEG data was filtered into frequency bands between 0.5 and 32 Hz using an 11-channel log-filterbank. The common spatial patterns (CSP) dimensionality reduction method was then applied to the data. CSP computes components that maximize the variance between the two classes. Spatial topographies of the first four components are shown in Figure 1. The components were computed over short time windows, and the variance of the components within these windows were computed as features. These features were passed to the SVM using a 3-1 training/testing split. A gated recurrent unit (GRU) classifier was also used by dividing the time windows into 2.5s sub-windows and passing the features from these sub-windows into a GRU deep neural network.

RESULTS

For a 5s time window, 67.4% accuracy was achieved. Improved accuracy could be achieved by using a GRU deep neural network. The area under the GRU classification curve (described in section III) is indicated as GRU-AUC.

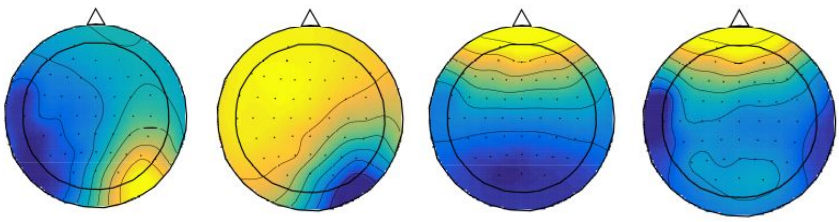


Figure 1. First four CSP components, all in the 1-8Hz frequency range.

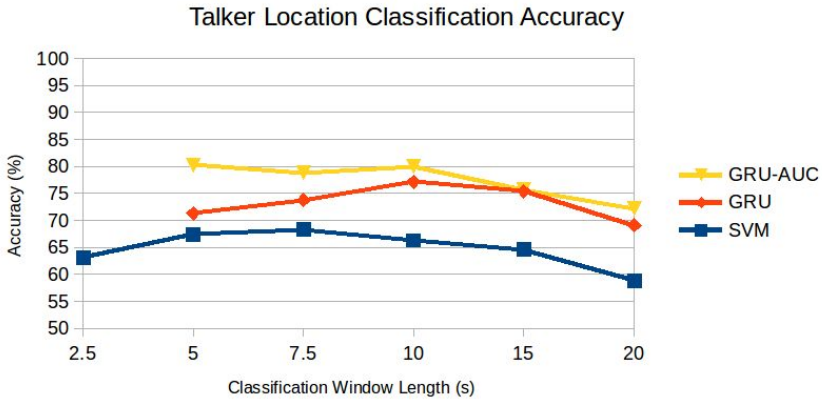


Figure 2. Classification accuracy.

DISCUSSION

Better accuracy was achieved, compared to CCA envelope decoding described in Section III. As discussed in Section III, it is possible that with less noisy EEG data, and with speech streams that are easier for the subject to attend to, even better classification can be attained. Further work can be done to classify additional positions.

V. Summary

The cocktail party experiment design offered a way to assess the performance of envelope decoding and location classification methods in a free-field multitalker environment, with minimal confounds. The results provided insight into how microphone array beamforming and EEG decoding strategies could be integrated into a cognitively steered hearing aid. It was demonstrated that microphone array beamforming could be used to obtain segregated speech envelopes that could be used for classifying which one was being attended to.

From an implementation standpoint, one possible configuration could use location decoding as a means to provide coarse beam-steering. In parallel, LCMV beamforming can be used to identify possible speech streams, potentially using information from location decoding to narrow down the number of streams. Envelope decoding can then be used to determine which stream to amplify. By combining information from both location classification and envelope decoding, it may be possible to achieve a higher classification accuracy than either of the two methods alone.

Reference

Lin Z., Chen M., Wu L. and Ma Y., “The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices,” UIUC Technical Report UILU-ENG-09-2215, Nov 2009.

Müller S. and Massarani P., “Transfer function measurement with sweeps,” J. Audio Engin. Soc., vol. 49, iss. 6, pp. 443-471, Jun 2001.

O’Sullivan J., Power A.J., Mesgarani N, Rajaram S., Foxe J.J., Shinn-Cunningham B.G., Slaney M., Shamma S.A. and Lalor E.C., “Attentional selection in a cocktail party environment can be decoded from single-trial EEG,” Cereb. Cortex, vol. 25, iss. 7, pp. 1697-706, Jul 2015.

Wong D.D.E., Pomper U., Alickovic E., Hjortkaer J., Slaney M., Shamma S., de Cheveigné A., “Decoding Speech Sound Source Direction from Electroencephalography Data,” ARO Mid-Winter Meeting, Feb 2016 [abstract].

Van Eryndhoven S., Francart T. and Bertrand A., “EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses,” IEEE Trans. Biomed. Eng., Jul 2016 [Epub].

Van Veen B.D., “Beamforming: a versatile approach to spatial filtering,” IEEE ASSP Mag., vol. 5, iss. 2, pp. 4-24, Apr 1988.