NAPLIB: AN OPEN SOURCE TOOLBOX FOR REAL-TIME AND OFFLINE NEURAL ACOUSTIC PROCESSING

Bahar Khalighinejad[1]*, Tasha Nagamine*[1]*, Ashesh Mehta*[2]*, Nima Mesgarani*[1]

[1]Department of Electrical Engineering, Columbia University, New York, USA
[2]Department of Neurosurgery, Hofstra Northwell School of Medicine and Feinstein Institute for Medical Research, Manhasset, New York, USA

ABSTRACT

In this paper, we introduce the Neural Acoustic Processing Library (NAPLib), a toolbox containing novel processing methods for real-time and offline analysis of neural activity in response to speech. Our method divides the speech signal and resultant neural activity into segmental units (e.g., phonemes), allowing for fast and efficient computations that can be implemented in real-time. NAPLib contains a suite of tools that characterize various properties of the neural representation of speech, which can be used for functionality such as characterizing electrode tuning properties, brain mapping and brain computer interfaces. The library is general and applicable to both invasive and non-invasive recordings, including electroencephalography (EEG), electrocorticography (ECoG) and magnetoecnephalography (MEG). In this work, we describe the structure of NAPLib, as well as demonstrating its use in both EEG and ECoG. We believe NAPLib provides a valuable tool to both clinicians and researchers who are interested in the representation of speech in the brain.

Index Terms— auditory neuroscience, EEG, ECoG, brain mapping, real-time processing

1. INTRODUCTION/BACKGROUND

Humans are unique in their ability to understand speech. As such, much research has gone into understanding the human auditory cortex, a brain region that plays an important role in the process of speech perception. However, progress in understanding the central auditory system has been hindered by lack of tools that can effectively and efficiently quantify the representation of speech at different stages of neural transformations, a problem that plagues both invasive and non-invasive recording methods.

Many non-invasive studies utilizing EEG and MEG in recent years have focused on understanding speech processing for applications to various hearing and language disorders [1, 2, 3]. Additionally, invasive recordings can be used to study auditory neuroscience, as well as performing brain mapping, an essential clinical procedure for epilepsy patients who go under surgical resection of seizure foci. Several approaches for brain mapping have already been developed, including electrical cortical stimulation (ECS) [4, 5], cortico-cortical evoked potentials (CCEP) [6], and mapping based on high gamma activity [7]. While these methods are effective for their intended applications, they do not provide information about tuning properties of electrodes or characterize the neural encoding of speech. Additionally, ECS can induce seizures in subjects.

Traditionally, quantification of the tuning properties of auditory brain regions is performed by calculating spectro-temporal receptive fields (STRFs), which are linear maps between stimulus and response that quantify a neuron's or neural population's ideal stimulus [8]. However, STRFs suffer from several drawbacks. First, STRFs assume a linear relationship between stimulus and response, an assumption which has been proven false, particularly in higher-level processes [9, 10]. Additionally, STRFs are dependent on the particular algorithm chosen for regularization (e.g., norm and sparsity constraints), which can limit their interpretability [11]. Finally, solving the linear regression typically requires the computation of the inverse of large matrices, making them computationally intractable for real-time systems.

In this paper, we present the Neural Acoustic Processing Library (NAPLib)[1], a library for studying brain regions involved in speech processing. Recent studies have shown the encoding of acoustic-phonetic features in speech cortices [12]; since each phonemic category has unique spectro-temporal properties, studying the responsiveness of neural activity to these categories informs us about spectro-temporal properties of responsive regions [13]. These methods do not make linear model assumptions, and they are computationally efficient meaning that they can be implemented in real-time [12]. We include both real-time and offline processing tools, and we demonstrate the use of this toolbox in both noninvasive and invasive neural recordings.

2. TOOLBOX DESCRIPTION

NAPLib is comprised of two main libraries for real-time and offline processing. The offline toolbox is developed in both MATLAB and Python, and provides functionality for source

[1]Available at http://naplab.ee.columbia.edu/NAPLib.

selectivity analysis [12], quantification of response delay, and analysis of phoneme similarity patterns in neural and acoustic space. The real-time toolbox, developed in Simulink, provides quantification of electrode responses to speech and shows the selectivity of sources to segmental units, such as phonemes. Additionally, we provide a small, open-source corpus of American English.

2.1. Speech stimuli

NAPLib quantifies spatial and temporal properties of neural responses to phoneme categories as subjects listen to continuous speech. In order to implement this technique, the continuous speech signal must be temporally aligned with the corresponding phoneme sequence. With the library, we provide a small, open-source corpus of American English with forced alignments generated using the Penn Phonetics Lab Forced Aligner [14]. We provide 25 minutes of speech, consisting of 148 utterances, 8450 phonemes and two speakers (a male and a female). NAPLib is generalizable to any phonetically aligned corpus (e.g. TIMIT [15]). Additionally, there are many open-source toolkits that can be used to generate forced alignments for existing and custom corpora [16, 17].

2.2. Offline processing

The offline toolbox is developed in both MATLAB and in Python. It contains three modules: data preprocessing, noise reduction and artifact rejection, and phoneme analysis.

2.2.1. Preprocessing

The preprocessing module aligns the phoneme labels (or other segmental unit) of the stimulus with the neural recording, allows the user to choose a scalp map (EEG) or electrode locations (ECoG) for visualization purposes, and performs filtering. For EEG, we provide zero-lag, FIR bandpass filter with cut-off frequencies of 2 and 15 Hz. For ECoG, we provide a filter bank to extract high gamma activity (envelop of 70 to 150 Hz), high gamma is correlated with neural spiking activity and encodes phonetic feature information [18, 12]. In offline processing, filters are non-causal and zero-phase.

2.2.2. Noise reduction

Users can choose from different noise reduction techniques including common average referencing, principal component analysis decomposition, and trial rejection based on visual inspection and setting a threshold.

2.2.3. Phoneme analysis

After preprocessing, denoising, and artifact rejection, the data can be fed into the phoneme analysis pipeline. Phoneme analysis can be used to perform brain mapping for speech selective regions, finding response delay and phonetic selectivity of electrodes, and quantifying the degree to which acoustic variability is reflected in neural data.

Selection of segmental unit. At the start of phoneme analysis, users can choose the unit that the rest of analysis will be based on. In addition to *individual phonemes* (default), we include functionality for grouping phones based on phonetic features (*manner of articulation*, *place of articulation*), *phone length*, and *speakers*. This allows for the study of acoustic, phonetic, and speaker features. In addition, this unit selection is easily generalizable and users can generate their own method for creating segmental units (e.g., syllables). When performing phonetic analyses, we recommend using individual phonemes for ECoG and EEG group analysis, while clustering labels into manner of articulation for single subject EEG due to noise concerns.

Average electrode response to phonemes. The average response elicited by each phoneme is an important tool for visualizing the feature selectivity of an electrode [12]. The average response $\overline{R}_{e,k}$ of electrode e for phoneme k occurring at time points of $T_1, T_2, ..., T_{N_k}$ in the stimulus is given by:

$$\overline{R}_{e,k}(t) = \frac{1}{N_k} \sum_{n_k=1}^{N_k} r(T_{n_k} + t, e), \tag{1}$$

where $r(t_{n_k}, e)$ is the neural response at time of phone onset and t defines the temporal window over which the average is computed.

Response delay. The latency between speech stimulus S and neural response R varies based on the brain region from which it was recorded [19]. We quantify the distinction between phonemes at each time point using the F-statistic (between-group variability divided by within-group variability) and define response delay as the time point that shows maximum distinction between categories. Consider $R_{e,k,n}(t)$ as the response of electrode e to the n^{th} instance of phoneme category k, where t denotes the sample time after the onset of phoneme. The response time is given by:

$$T_e = \underset{t}{\operatorname{argmax}} \left(\frac{\sum_k N_k (\overline{R}_{e,k}(t) - \overline{R}_e(t))^2/(K-1)}{\sum_{nk} (R_{e,k,n}(t) - \overline{R}_{e,k}(t))^2/(N-K)} \right), \tag{2}$$

where K denotes the number of phoneme categories, N denotes the total number of phones in the corpus (all categories), $\overline{R}_{e,k}(t)$ is defined in (1), and $\overline{R}_e(t)$ is the global mean of responses regardless of phoneme category. Categories are by default individual phonemes, but this can be generalized to any specified segmental unit.

Phoneme selectivity of channels. In order to characterize the selectivity of the neural response to phonemic categories at individual electrodes, we calculate the phoneme selectivity index (PSI) vectors as described in [12]. Each electrode is characterized by a [Kx1] vector, with each element corresponding to the PSI of one phoneme; each PSI has a value ranging from 0 to K that quantifies the number of phonemes that elicit a statistically distinguishable response from the target phoneme (Wilcoxon rank-sum test).

Quantification of distinction between phonemes. We calculate the distance (default is Euclidean) between the re-

sponses to phonemes for each phoneme pair at every time lag, yielding a time-varying pairwise phoneme distance matrix. This analysis focuses on the similarities and distinctions between categories rather than on individual items. We also provide functionality to visualize the distance matrices in two and three dimensions using multi-dimensional scaling (MDS) [20] and t-SNE [21].

Comparison of phoneme properties between stimuli and response. Speech is a continuous signal that changes over time; even within a single category, the acoustic properties change from the start to the end of the phone. In order to find how neural responses and acoustic properties of speech sounds are related through time, we define a *neural-acoustic covariance matrix*. This is a two-dimensional matrix that demonstrates the similarity between patterns of phones in the acoustic space and the corresponding neural responses at each time point. The acoustic similarity matrix is calculated using the acoustic spectrogram of phones [22].

Functional connectivity of electrodes. The functional connectivity of recording regions is quantified by finding the covariance between distinction patterns of different electrodes.

Group analysis. We provide an option of group analysis specifically recommended for analysis of EEG data when one subject does not provide sufficient signal to noise ratio. In this case, the response to the same phone is averaged between different subjects, after which all of the other analyses can be utilized.

2.3. Real-time processing

The real-time processing toolbox is implemented in Simulink and utilizes similar methods to the offline toolbox, which are simplified to create efficient computations.

Figure 1 illustrates the mechanism of real-time processing. Audio is a sound file that includes the stimulus as one channel and the phoneme labels as the other. As the subject listens to continuous speech, the phoneme labels are sent to the processor while the neural responses are recorded simultaneously. In preprocessing, users can choose between EEG and ECoG filters, then, through a rate-transition module, both phoneme labels and neural data will be resampled to 100 Hz. Next, neural data will be saved in a buffer with a window size of M samples (default: 600ms), and phoneme labels will be delayed by N samples. This defines the maximum number of samples after the phoneme onset which goes to phoneme response analyzer. For the data shown in Fig. 2A, M is equal to 60 samples (600 ms) and N is equal to 50 samples (500 ms).

In the phoneme response analyser block of the toolbox the following analysis are implemented: *selection of segmental unit*, *average electrode response to phonemes*, *response delay, and phoneme selectivity*.

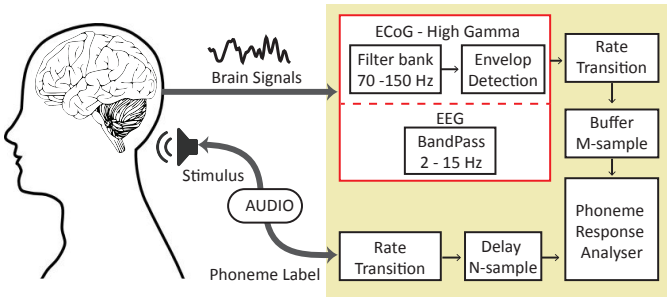

Fig. 1. Schematic of the real-time processing toolbox.

3. EXPERIMENTS

To show the efficacy of our toolbox, we demonstrate analyses from neural recordings in both EEG and ECoG.

3.1. Neural Recordings

We recorded neural activity from subjects as they listened to the provided NAPLib corpus. We recorded from 22 EEG participants with a 62-channel recording system. The three ECoG participants were undergoing neurological assessment for epilepsy surgery; one patient had a high-density micro-electrode grid array over temporal lobe, while the other had stereo EEG electrodes implanted. All subjects provided written informed consent. The Institutional Review board (IRB) of Columbia University at Morningside Campus approved all procedures.

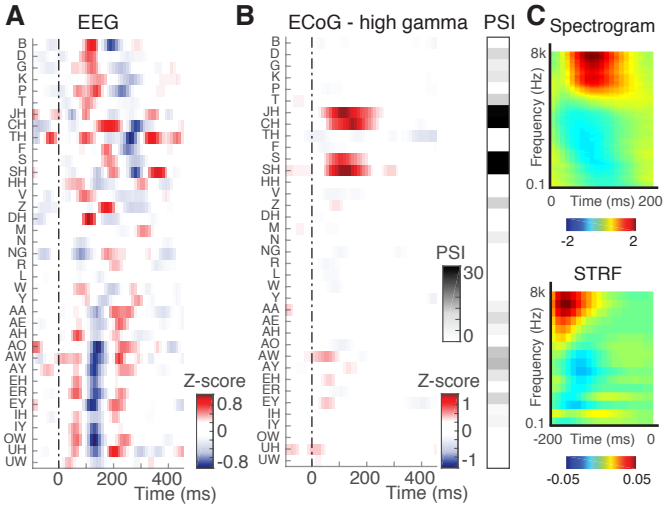

Fig. 2. Average response to phonemes in (A) EEG and (B) ECoG. The PSI vector for the ECoG electrode is shown at right. (C) Average spectrogram of combined phonemes /jh,ch,s,sh/ and the STRF of the ECoG electrode.

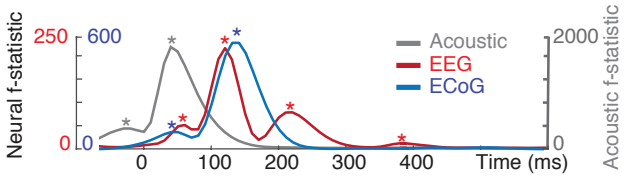

Fig. 3. F-test over time and response delay for acoustic phonemes (gray), one electrode in EEG (red) and one electrode in ECoG (blue).

3.2. Average electrode response and phoneme selectivity in EEG and ECoG

In this section we demonstrate how NAPLib can be used to visualize and quantify phonetic selectivity in both invasive and non-invasive recordings. Results shown are using the offline module, but we would like to emphasize that these analyses can also be implemented using the real-time module.

Figure 2A shows the average response of an example electrode (FCz) in EEG generated by group analysis including 22 subjects. Due to poor spatial resolution, it is typical to find broad responses to many phonemic categories. Because EEG recordings are also noisy, we also recommend averaging responses over subjects using group analysis.

ECoG recordings provide much higher spatial resolution, resulting in the average response and corresponding PSI vector shown in Figure 2B from a depth electrode in Heschl's gyrus. This electrode responds to unvoiced sibilants and affricates ($PSI > 25$), which all contain strong power in high frequency channels. This suggests that this electrode has broad tuning to high frequencies, and indeed, we can see that the STRF of this electrode closely matches the average spectrogram of these phonemes (Figure 2C).

3.3. Response delay

Figure 3 shows the F-test value at each time point based on the onse of phonemes. Phonemes are categorised based on manners of articulation. Figure illustrated the time differences between acoustic phonemes, ECoG (an electrode in Heschl's gyrus), and EEG (FCz, 22 subjects). The local maxima are denoted with asterisks.

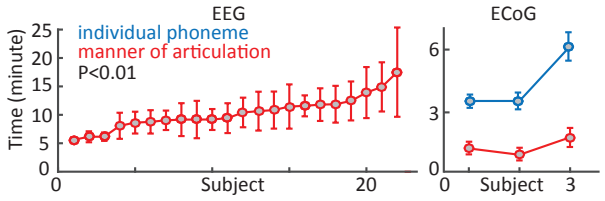

Fig. 4. Duration needed to find a significant electrode (ANOVA F-test). Error bars show standard deviation.

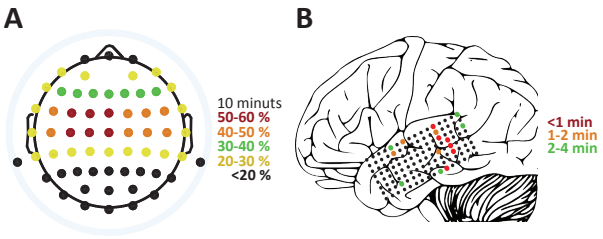

Fig. 5. (A) Scalp locations of responsive electrodes to speech over a duration 10 minutes. Percentages indicate the fraction of the recording time that an electrode displayed a statistically significant response to speech (ANOVA F-test). (B) The location of responsive grid electrodes to speech in a subject implanted with an ECoG array.

3.4. Mapping time

In order to quantify the duration which is needed to find a significant phoneme response, we used the ANOVA F-test. In EEG, the reported duration is based on significant distinction between *manners of articulation*. In ECoG, the duration for both *manner of articulation* and *individual phoneme categories* is reported. The p-value is assessed by the F-distribution with correction for multiple comparisons implemented via false discovery rate ($q < 0.01$). The calculated time duration does not include the natural silences of speech.

3.5. Locating speech-responsive regions

Determining the location of response is important for a variety of both clinical and research applications. Figure 5A shows the percentage of time that EEG electrode responses display a statistically significant response to speech over a 10 minute duration. Figure 5B shows the time needed to elicit a statistically significant response to speech a patient with an implanted ECoG microelectrode array.

4. CONCLUSIONS

In this paper we introduce the Neural Acoustic Processing Library (NAPLib), a free and open source toolbox for studying the neural representation of speech. The toolbox quantifies temporal and spectral responsiveness of electrodes based on responses to segmental linguistic categories (phonemes). Using such an approach allows for fast, efficient computations that can be implemented in real-time. As a proof of concept, we demonstrate use of the toolbox using both invasive (ECoG) and non-invasive (EEG) recordings.

5. ACKNOWLEDGEMENTS

This work was funded by a grant from National Institute of Health, NIDCD, DC014279 and the Pew Charitable Trusts, Pew Biomedical Scholars Program.

6. REFERENCES

[1] T. K. Guttorm, P. H. T. Leppänen, A.-M. Poikkeus, K. M. Eklund, P. Lyytinen, and H. Lyytinen, "Brain event-related potentials (ERPs) measured at birth predict later language development in children with and without familial risk for dyslexia," *Cortex*, vol. 41, no. 3, pp. 291–303, 2005.

[2] S. J. Johnstone, R. J. Barry, and A. R. Clarke, "Ten years on: a follow-up review of ERP research in attention-deficit/hyperactivity disorder," *Clinical Neurophysiology*, vol. 124, no. 4, pp. 644–657, 2013.

[3] R. Gil-da Costa, G. R. Stoner, R. Fung, and T. D. Albright, "Nonhuman primate model of schizophrenia using a noninvasive EEG method," *Proceedings of the National Academy of Sciences*, vol. 110, no. 38, pp. 15 425–15 430, 2013.

[4] H. W. Lee, S. B. Hong, D. W. Seo, W. S. Tae, and S. C. Hong, "Mapping of functional organization in human visual cortex Electrical cortical stimulation," *Neurology*, vol. 54, no. 4, pp. 849–854, 2000.

[5] A. Sinai, C. W. Bowers, C. M. Crainiceanu, D. Boatman, B. Gordon, R. P. Lesser, F. A. Lenz, and N. E. Crone, "Electrocorticographic high gamma activity versus electrical cortical stimulation mapping of naming," *Brain*, vol. 128, no. 7, pp. 1556–1570, 2005.

[6] C. J. Keller, C. J. Honey, P. Mégevand, L. Entz, I. Ulbert, and A. D. Mehta, "Mapping human brain networks with cortico-cortical evoked potentials," *Phil. Trans. R. Soc. B*, vol. 369, no. 1653, p. 20130528, 2014.

[7] C. Kapeller, M. Korostenskaja, R. Prueckl, P.-C. Chen, K. H. Lee, M. Westerveld, C. M. Salinas, J. C. Cook, J. E. Baumgartner, and C. Guger, "CortiQ-based Real-Time Functional Mapping for Epilepsy Surgery," *Journal of clinical neurophysiology*, vol. 32, no. 3, pp. e12–e22, 2015.

[8] F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant, "Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli," *Network: Computation in Neural Systems*, vol. 12, no. 3, pp. 289–316, 2001.

[9] C. K. Machens, M. S. Wehr, and A. M. Zador, "Linearity of cortical receptive fields measured with natural sounds," *The Journal of neuroscience*, vol. 24, no. 5, pp. 1089–1100, 2004.

[10] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nature neuroscience*, vol. 6, no. 11, pp. 1216–1223, 2003.

[11] G. B. Christianson, M. Sahani, and J. F. Linden, "The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields," *The Journal of Neuroscience*, vol. 28, no. 2, pp. 446–455, 2008.

[12] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic Feature Encoding in Human Superior Temporal Gyrus," *Science*, p. 1245994, 2014.

[13] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Phoneme representation and classification in primary auditory cortex," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 899–909, 2008.

[14] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[17] M. Najafian, A. DeMarco, S. J. Cox, and M. J. Russell, "Unsupervised model selection for recognition of regional accented speech." in *INTERSPEECH*, 2014, pp. 2967–2971.

[18] N. E. Crone, D. Boatman, B. Gordon, and L. Hao, "Induced electrocorticographic gamma activity during auditory perception," *Clinical neurophysiology*, vol. 112, no. 4, pp. 565–582, 2001.

[19] C. Liegeois-Chauvel, A. Musolino, J. M. Badier, P. Marquis, and P. Chauvel, "Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components," *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, vol. 92, no. 3, pp. 204–214, 1994.

[20] I. Borg and P. J. F. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[21] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[22] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE transactions on information theory*, vol. 38, no. 2, pp. 824–839, 1992.