Automatic Speech Activity Recognition from MEG Signals Using Seq2Seq Learning

Conference Paper · March 2019

DOI: 10.1109/NER.2019.8717186

4 authors:

Debadatta Dash
National Institutes of Health
35 PUBLICATIONS **271** CITATIONS
SEE PROFILE

Ferrari Paul
Helen DeVos Children's Hospital
64 PUBLICATIONS **1,669** CITATIONS
SEE PROFILE

Saleem I Malik
Cook Children's Health Care System
23 PUBLICATIONS **245** CITATIONS
SEE PROFILE

Jun Wang
University of Texas at Austin
112 PUBLICATIONS **2,548** CITATIONS
SEE PROFILE

9th International IEEE EMBS Conference on Neural Engineering
San Francisco, CA, USA, March 20 - 23 , 2019

Automatic Speech Activity Recognition from MEG Signals Using Seq2Seq Learning

Debadatta Dash[1], Paul Ferrari[2], Saleem Malik[3], and Jun Wang[4]

Abstract— **Accurate interpretation of speech activity from brain signals is critical for understanding the relationship between neural patterns and speech production. Current research on speech activity recognition from the brain activity heavily relies on the region of interest (ROI) based functional connectivity analysis or source separation strategies to map the activity as a spatial localization of a brain function. Albeit effective, these methods require prior knowledge of the brain and expensive computational effort. In this study, we investigated automatic speech activity recognition from brain signals using machine learning. Neural signals of four subjects during four stages of a speech task (i.e., rest, perception, preparation, and production) were recorded using magnetoencephalography (MEG), which has an excellent temporal and spatial resolution. First, a deep neural network (DNN) was used to classify the four whole tasks from the MEG signals. Further, we trained a sequence to sequence (Seq2Seq) long short-term memory-recurrent neural network (LSTM-RNN) for continuous (sample by sample) prediction of the speech stages/tasks by leveraging its sequential pattern learning paradigm. Experimental results indicate the effectiveness of both DNN and LSTM-RNN for automatic speech activity recognition from MEG signals.**

I. INTRODUCTION

Activity recognition from the brain deals with the accurate interpretation of human actions through a series of observations via representative neural signals. The complexities and variances involved in the brain functions make the activity recognition task extremely challenging. Researchers use various neuroimaging modalities to acquire task-evoked neural signals and then map it to the structural brain image to observe the spatial localization of the actions. Based on the observations and prior neuroscientific knowledge, task related conclusions are derived. Although there are multiple studies on brain activity recognition for the understanding of the human brain during various activities such as emotion [1] face recognition [2], biometrics [3], memory [4] and age-specific effects [5]; it has been particularly focused on healthcare applications. Especially in detection and understanding of neurologic disorders such as down syndrome [6], Schizophrenia [7], autism [8], and brain injury [9], brain activity recognition is crucial.

Recognizing different types of speech activities, including speech perception, preparation, and articulation, from neural signals are particularly useful for the building of neural decoding based brain computer interface (BCI) with the potential for higher communication rate than the current ones [10], which are based on visual/attention cues in the neural signals. Neural decoding-based BCIs translate neural pattern to speech (text) directly [11][12]. Recent work has shown the feasibility of recognizing isolated phonemes [11] or phrase production [12] from invasive or noninvasive neural signals. To move this type of BCI work further, automatic recognition of different types of speech activity is needed, so that the continuous speech recognition performance can be maximized with the known speech activity boundaries. This task, however, is extremely challenging because speech activities have very fast temporal dynamics.

Prior research on speech-evoked brain activity recognition is primarily based on functional magnetic resonance imaging (fMRI)[13][14] as it has a very high spatial resolution [15] and can effectively locate the sources of activation in the brain [16][17]. Functional near-infrared spectroscopy (fNIRS) has also been used as a low-cost alternative to fMRI for task recognition [18]. However, the low temporal resolution of both fMRI and fNIRS hinders in efficiently modeling the high frequency non-stationary characteristics of speech and hence, are not preferable for neural speech task recognition. Neural patterns of speech have been also studied with positron emission tomography (PET) [19] and Electrocorticography (ECoG) [20][21]. But, they are invasive and thus, practically unsuitable. Although Electroencephalography (EEG) is non-invasive and is popularly applied for neural speech task ('Yes' and 'No') classification [22], its low spatial resolution results in intermediate accuracy. Moreover, these traditional approaches suffer from experimental bias and expensive computational modeling. We believe that, by leveraging the recent advances in machine learning, the neural speech activity recognition task can be automated which will result in huge computational gain and generalization.

In this study, we investigated automatic speech activity recognition from MEG signals. MEG is a non-invasive neuroimaging modality, that has a higher spatial resolution than EEG and a better temporal resolution than fMRI and fNIRS. Recent speech studies with MEG [12][23][24][25] suggest its effectiveness for capturing the neural speech information. Leveraging these advantages of MEG over other neuroimaging modalities, this study aimed to automatically

[1] Debadatta Dash is a Ph.D. Student with the Department of Bioengineering, The University of Texas at Dallas, TX 75080, USA. debadatta.dash@utdallas.edu

[2] Paul Ferrari is Research Associate Professor with the Department of Psychology, The University of Texas at Austin, and the Research Director of the MEG lab, Dell Children's Medical Center, Austin, TX 78712, USA. pferrari@utexas.edu

[3] Saleem Malik is a neurologist and the director of the MEG lab at Cook Children's Hospital, TX 76104, USA. saleem.malik@cookchildrens.org

[4] Jun Wang is an Assistant Professor in the Department of Bioengineering and Callier Center for Communication Disorders, The University of Texas at Dallas, TX 75080, USA. wangjun@utdallas.edu

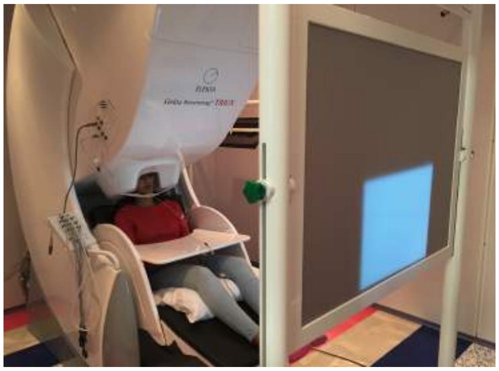

Fig. 1. The Neuromag Elekta MEG unit with a subject

predict speech perception, preparation, production and resting stages from MEG signals using two latest machine learning algorithms including DNN and Seq2Seq LSTM-RNN.

II. DATA COLLECTION

A. The MEG Unit

The brain activity signals were recorded via a 306 channeled (204 gradiometers and 102 magnetometers) Elekta Triux Neuromag MEG device housed inside a magnetically shielded room (MSR) (Figure 1). A computer interfaced DLP projector displayed the speech stimulus on a screen situated at about 90 cm distance from the device. All the sensors were calibrated for noise levels prior to data collection.

B. Subjects

Four healthy subjects (2 females) participated in this pilot study. All the subjects had normal or corrected to normal vision. No speech/language/hearing or cognitive disorder history was reported. Written consent from each subject was taken before the experiment.

C. Stimuli

Five common daily used English phrases are selected as the stimuli of the experiment. They are: *1. Do you understand me? 2. That's perfect. 3. How are you? 4. I need help. 5. Good-bye.* These sentences are typically used for alternative and augmentative communication (AAC) and hence, were chosen as stimuli in this study.

D. Protocol

The experiment was designed in four continuous stages as rest (pre-stimuli), perception, preparation (imagination) and production (articulation) as shown in Figure 2. The first (rest) stage did not involve any task and was fixed to be of 0.5 s prior to the stimulus onset. The second stage was designed as a visual speech perception task, where a single stimulus (phrase) was displayed on the screen for 1 s. The phrase was then replaced with a fixation cross on the screen during the next stage of preparation where the subjects were signaled to prepare for the articulation of the phrase. After 1 s of preparation stage, the final stage of speech production was initiated by displaying a blank screen which heralded the participant to overtly articulate the phrase at their normal speaking rate which lasted for an average of 2 seconds. This protocol was repeated 100 times for all the 5 phrases. The stimuli were presented on the screen in a pseudo-randomized order to avoid response suppression due to repeated exposure [26]. Training of subjects on sample stimuli was done prior to data collection to ensure compliance.

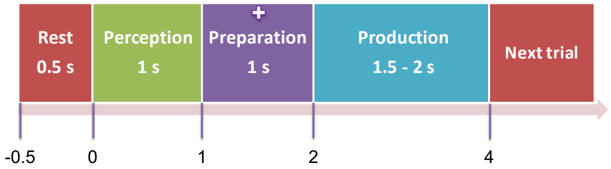

Fig. 2. Protocol of the Time-locked Experiment

E. Data Preprocessing

MEG signals were acquired with 4 kHz sampling frequency and then down-sampled to 1 kHz for further analysis. The recorded data of each stimulus was epoched into trials from -0.5 to $+4.0$ s centered on stimulus onset as shown in Figure 2. Head motion, eye blinks, and cardiac signals were recorded with default continuous head localization technique, EOG and ECG respectively and were removed from the MEG data. Further remaining artifacts and untimely articulated signals were visually inspected and removed. After preprocessing, a total of 1635 valid trials were remained out of 2000 (4 subjects x 5 phrases x 100 repetitions) acquisitions. Out of 204 gradiometers, 4 sensors contained high artifacts and hence were not considered for analysis.

III. METHODS

A. Wavelet Analysis

In our prior work on speech decoding from neuromagnetic signals [23][27], wavelets were proved to be efficient for denoising of MEG signals. Hence, we used Daubechies (db)-4 discrete wavelet with a three level decomposition to denoise and restrict the brain activity signals within the gamma frequency bandwidth range.

B. Feature Extraction

Three types of features were extracted from the MEG signals at each millisecond of time. Considering the whole brain as a single source of the task at a particular time, the sum of energy produced by all the 200 sensors at each ms of time was used as the first feature. To emphasize the role of negative MEG signal values, the cumulative sum of values given by all the 200 sensors was calculated as the second feature. Further, based on the activation, we computed the most active sensor at each ms of the task and its index $(1-200)$ was used as the third feature (Figure 3).

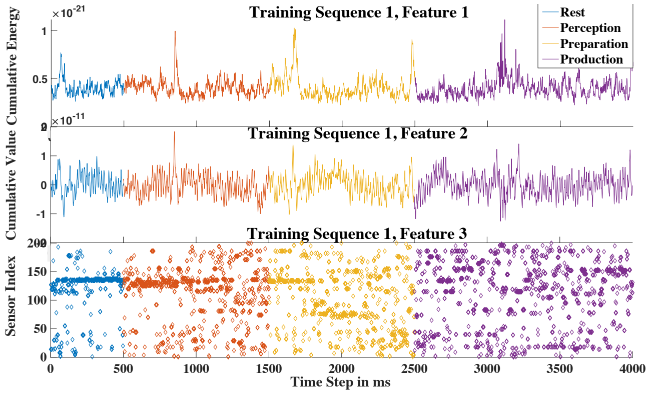

Fig. 3. Extracted MEG features

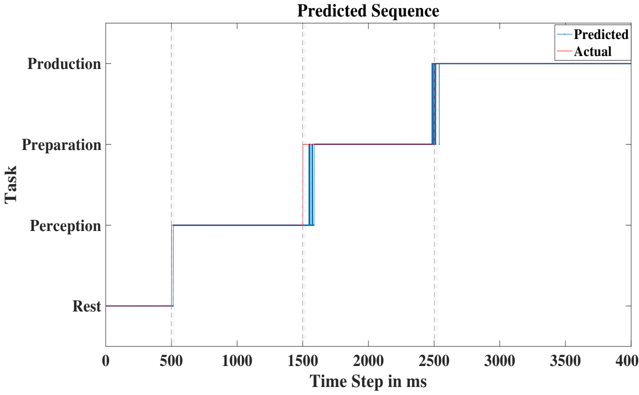

Fig. 4. Actual v. Predicted Sequence with Seq2Seq Leaning

C. Isolated Classification Using DNN

The features obtained during each of the four stages were separated and trained with a three-layered DNN for isolated classification of the stages. The input layer was designed to take the extracted features (3 feature dimension × 1000/1500 temporal dimension) of all the valid trials per stage. The 3 hidden layers consisted of reducing number of nodes as 256, 128 and 64. The output layer consisted of a fully connected softmax layer with 4 nodes to compute the cross-entropy score of each class. The network was trained using backpropagation with 70% training data. The remaining data was divided as 15% for validation (to check for data over-fitting) and 15% for testing. The initial learning rate was fixed at 0.01 with a drop factor of 0.1 per 20 epochs. The hyper-parameters providing the best validation results were selected for the test data set. Results from the test set are reported in the Results section.

D. Continuous Prediction Using LSTM-RNN

A Seq2Seq LSTM can be used to continuously predict the classes at each time point on a sequential data [28]. We designed a Seq2Seq LSTM to take the temporal sequence of the three-dimensional features as the input. It consisted of a hidden layer with 100 nodes (memory blocks) for encoding the features. This hidden layer contained the learnable parameters (input and recurrent weights), designed as a vertical concatenation of the input/recurrent weight matrices for the components (gates) of the LSTM layer namely, input gate, forget gate, cell candidate and output gate in respective order. A hard-sigmoid activation function was used to update the gate state. The tanh activation function was used to update the cell and hidden states. This unique arrangement of memory blocks helps in performing additive interactions to improve gradient flow over long sequences during training [28]. The hidden LSTM layer was then followed by a fully connected, a softmax and a classification layer, each with 4 nodes to decode the class information by classifying the four classes and updating the network states simultaneously. The network was trained via an Adam optimizer with a gradient threshold of 0.1, gradient threshold method of L_2 norm, initial learning rate of 0.01, learning rate drop factor of 0.1 after 20 epochs and a maximum epoch of 400. 70% of the data (irrespective of the phrase stimulus) was used for training and 15% each of the data was used for validation and testing. Hyper-parameters were chosen based on the best validation results.

IV. Results and Discussion

Figure 4 represents the comparison of a test sequence (red line) with the corresponding Seq2Seq prediction (blue line). It is clear that LSTM is able to efficiently predict the speech production stages as both the lines are highly overlapping. It can also be observed that for some time steps, the preparation stage has been predicted as perception and production stage has been predicted as preparation. However, these inaccurate predictions occurred at the stage boundaries, indicating boundaries between stages are more of a continuum rather than binary. That there are similarities in neural behavior among post perception and starting of speech preparation stages is quite reasonable. Additionally, although the articulation segment was sharply defined at 1 s post stimulus, the occurrence of true articulation was slightly variable and could be a reason for the misclassification during initial production time steps. An average of 93.53% prediction accuracy was obtained through LSTM (Figure 5).

Speech production is continuous with the perception and preparation stages inherent to the process. In this study, DNN was used for isolated classification of the speech production stages and the results can be seen in Figure 5.

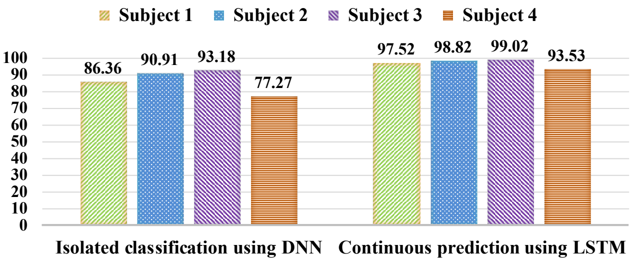

Fig. 5. Subject Level Isolated and Continuous Prediction Accuracy

An average of 86.93% classification accuracy was obtained across subjects. This indicates that the neural dynamics of speech is distinguishable across stages. It can be observed that the accuracies obtained by both DNN and LSTM are subject dependent. For both DNN and LSTM, subject 4 resulted in a relatively lower accuracy, probably due to the presence of more artifacts compared to the other subjects.

Current MEG has limitations to make it ready for BCI purposes, such as high cost, unportability, and large size. However, continuous advancements of cost-effective MEG acquisition techniques including a recently developed portable MEG [29] strengthens the potential of MEG as a neural device for BCI applications in the future.

V. CONCLUSIONS

In this study, we demonstrated the possibility of automatic recognition of speech perception, preparation, production, and resting stage of the brain from MEG signals using machine learning. DNN was successfully implemented for isolated classification of the four speech stages, whereas, LSTM-RNN based sequence to sequence learning was proved to be highly effective for continuous prediction of the speech task-evoked brain stages at each millisecond of time. This efficient isolation and prediction of speech stages from the brain signal will enable in the building of the next-generation, neural decoding-based brain computer interfaces.

ACKNOWLEDGMENT

This study was supported by the University of Texas System Brain Research grant, the National Institutes of Health (NIH) under award number R03 DC013990, and the American Speech-Language-Hearing Foundation through a New Century Scholar Grant. We thank Dr. Daragh Heitzman, Dr. Mark McManis, Dr. Ted Mau, Dr. Angel W. Hernandez-Mulero, Kristin Teplansky, and the volunteering participants.

REFERENCES

[1] R. Horlings, D. Datcu, and L. J. M. Rothkrantz, Emotion Recognition Using Brain Activity, 9^{th} International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing, Comp. Sys.Tech., Gabrovo, Bulgaria, pp. 6:II.1–6:1, 2008.

[2] C. A. Nelson, The development and neural bases of face recognition, Inf. Child Develop., 10, pp. 3-18, 2001.

[3] S. Yang and F. Deravi, On the Usability of Electroencephalographic Signals for Biometric Recognition: A Survey, IEEE Transactions on Human-Machine Systems, 47(6), pp. 958-969, 2017.

[4] G. R. I. Barker, F. Bird, V. Alexander and E. C. Warburton, Recognition Memory for Objects, Place, and Temporal Order: A Disconnection Analysis of the Role of the Medial Prefrontal Cortex and Perirhinal Cortex, Journal of Neuroscience, 27(11), pp. 2948-2957, 2017.

[5] R. Cabeza, S. M. Daselaar, F. Dolcos, S. E. Prince, M. Budde and L. Nyberg, Task-independent and Task-specific Age Effects on Brain Activity during Working Memory, Visual Attention and Episodic Retrieval, Cerebral Cortex, 14(4), pp. 364-375, 2004

[6] J. H. Karrer, R. Karrer, D. Bloom, L. Chaney and R. Davis, Event-related brain potentials during an extended visual recognition memory task depict delayed development of cerebral inhibitory processes among 6-month-old infants with Down syndrome, International Journal of Psychophysiology, 29(2), pp. 167-200, 1998.

[7] H. H. Holcomb, A. C. Lahti, D. R. Medoff, M. Weiler, R. F. Dannals, and C. A.Tamminga, Brain Activation Patterns in Schizophrenic and Comparison Volunteers During a Matched-Performance Auditory Recognition Task, American Journal of Psychiatry, 157(10), pp. 1634-1645, 2000.

[8] A. Nowicka, C. B. Hanna, P. Tacikowski, P. Ostaszewski, and R. Kuś, Name recognition in autism: EEG evidence of altered patterns of brain activity and connectivity, Molecular Autism, 7(1), p. 38, 2016.

[9] R. K. Yin, Face recognition by brain-injured patients: A dissociable ability?, Neuropsychologia, 8(4), pp. 395 - 402, 1970.

[10] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, Brain-computer interfaces for communication and control, Clinical Neurophysiology, no. 113, pp. 767-791, 2002.

[11] C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, Brain-to-text: decoding spoken phrases from phone representations in the brain, Frontiers in Neuroscience, 9(217), 2015.

[12] J. Wang, M. Kim, A. W. Hernandez-Mulero, D. Heitzman, and P. Ferrari, Towards decoding speech production from single-trial magnetoencephalography (MEG) signals, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3036-3040, 2017.

[13] V. L. Gracco, P.Tremblay, and B.Piketitle, Imaging speech production using fMRI, NeuroImage, 26(1), pp. 294 - 301, 2005.

[14] K. Okada, J. H. Venezia, W. Matchin, K. Saberi, G. Hickok, An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex, PLOS ONE 8(6): e68959, 2013.

[15] D. Dash, V. Abrol, A. Sao, and B. Biswal, Spatial sparsification and low rank projection for fast analysis of multi-subject resting state fMRI data, IEEE 15th International Symposium on Biomedical Imaging (ISBI), pp. 1280-1283, 2018.

[16] D. Dash, V. Abrol, A. Sao, and B. Biswal, The model order limit: Deep sparse factorization for resting brain, IEEE 15th International Symposium on Biomedical Imaging (ISBI), pp. 1244-1247, 2018.

[17] D. Dash , B. Biswal, A. K. Sao, J. Wang, Automatic Recognition of Resting State fMRI Networks with Dictionary Learning. In: Brain Informatics (BI), Lecture Notes in Computer Science, vol 11309, Springer, pp. 249-259, 2018.

[18] N. Wan, A. S. Hancock, T. K. Moon, R. B. Gillam, A functional near-infrared spectroscopic investigation of speech production during reading. Hum Brain Mapp, 39, pp. 1428-1437, 2018.

[19] J. M. Rumsey, K. Nace, B. Donohue, D. Wise, J. M. Maisog, and P. Andreason, A Positron Emission Tomographic Study of Impaired Word Recognition and Phonological Processing in Dyslexic Men, Arch Neurol., 54(5), pp. 562-573, 1997.

[20] V. L. Towle, H. Yoon, M. Castelle, J. C. Edgar, N. M. Biassou, D. M. Frim, J. P. Spire, and M. H. Kohrman, ECoG gamma activity during a language task: differentiating expressive and receptive speech areas, Brain, 131(8), pp. 2013-2027, 2008.

[21] V. G. Kanas, I. Mporas, H. L. Benz, K. N. Sgarbas, A. Bezerianos and N. E. Crone, Real-time voice activity detection for ECoG-based speech brain machine interfaces, 19th International Conference on Digital Signal Processing, pp. 862-865, 2014.

[22] A. R. Sereshkeh, R. Trott, A. Bricout and T. Chau, EEG Classification of Covert Speech Using Regularized Neural Networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(12), pp. 2292-2300, 2017.

[23] D. Dash, P. Ferrari, S. Malik, and J. Wang, Overt speech retrieval from neuromagnetic signals using wavelets and artificial neural networks, in IEEE Global conference on Signal and Information Processing (GlobalSIP), pp. 489-493, 2018.

[24] G. S. Panagiotis , J. I. Breier, G. Zouridakis, and A.C. Papanicolaou, Identification of Language-Specific Brain Activity Using Magnetoencephalography, Routledge, Journal of Clinical and Experimental Neuropsychology, 20(5), pp. 706-722, 1998.

[25] R. Salmelin, P. Kiesil, K. Uutela, E. Service, and O. Salonen, Impaired visual word processing in dyslexia revealed with magnetoencephalography. Ann. Neurol., 40, pp. 157-162, 1996.

[26] K. Grill-Spector, R. Henson, and A. Martin, Repetition and the brain: neural models of stimulus-specific effects, Trends in Cognitive Sciences, 10(1), pp. 14-23, 2006.

[27] D. Dash, P. Ferrari, S. Malik, A. Montillo, J. Maldjian, and J. Wang, Determining the Optimal Number of MEG Trials: A Machine Learning and Speech Decoding Perspective, in Brain Informatics (BI), Lecture Notes in Computer Science, vol 11309, Springer, pp. 163-172, 2018.

[28] X. Zhu, P. Sobhani, and H. Guo, Long Short-term Memory over Recursive Structures, Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML), 37, pp. 1604-1612, 2015.

[29] E. Boto, N. Holmes , J. Leggett et al., Moving magnetoencephalography towards real-world applications with a wearable system, Nature, 555(7698), pp. 657-661, 2018.