TOWARDS DECODING SPEECH PRODUCTION FROM SINGLE-TRIAL MAGNETOENCEPHALOGRAPHY (MEG) SIGNALS

Jun Wang[1,2], Myungjong Kim[1], Angel W. Hernandez-Mulero[3], Daragh Heitzman[4], Paul Ferrari[5,6]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]MEG Center, Cook Children's Hospital, Fort Worth, Texas, United States
[4]MND/ALS Center, Texas Neurology, Dallas, Texas, United States
[5]MEG Laboratory, Dell Children's Medical Center, Austin, Texas, United States
[6]Department of Psychology, University of Texas at Austin, Austin, Texas, United States
wangjun@utdallas.edu

ABSTRACT

Patients with locked-in-syndrome (fully paralyzed but aware) struggle in their life and communication. Providing a level of communication offers these patients a chance to resume a meaningful life. Current brain-computer interface (BCI) communication requires users to build words from single letters selected on a screen, which is extremely inefficient. Faster approaches for their speech communication are highly needed. This project investigated the possibility to decode spoken phrases from non-invasive brain activity (MEG) signals. This direct brain-to-text mapping approach may provide a significantly faster communication rate than current BCIs can provide. We used dynamic time warping and Wiener filtering for noise reduction and then Gaussian mixture model and artificial neural network as the decoders. Preliminary results showed the possibility of decoding speech production from non-invasive brain signals. The best phrase classification accuracy was up to 94.54% from single-trial whole-head MEG recordings.

Index Terms— Brain computer interface, MEG, speech production, locked-in syndrome, neural decoding

1. INTRODUCTION

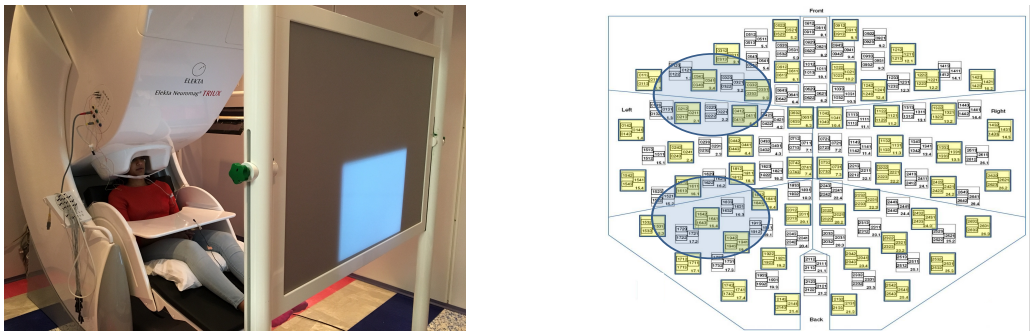
Brain damage or neurodegenerative disease (e.g., amyotrophic lateral sclerosis) may cause locked-in syndrome (fully paralyzed but aware) [1]. There is an incidence rate 0.7 / 10,000 for locked-in syndrome [2]. Patients with locked-in-syndrome struggle in their life and communication. Providing a level of communication offers these patients a chance to resume a meaningful life [3]. Brain activity may be the only pathway to facilitate the operation, control, and communication for these patients, because it bypasses the motor control mechanisms [4, 5, 6, 7].

Recent advances in BCIs have been applied for a number of potentially life-changing technologies for the disabled [8], including mechanical arm movement [9] and wheel-chair navigation [10]. Currently, the most widely used EEG-based BCIs for speech communication require users to select from a list of options (e.g., selecting single letters to build words) using visual or attention cues on a screen [8]. These approaches result in a slow communication rate of less than one word per minute [11, 12]. Faster approaches for speech communication are highly needed.

A direct mapping from brain activity signals to speech (text) will potentially provide a significantly faster communication rate than current BCIs can provide. Until recently, however, low spatial or temporal resolution of non-invasive neuroimaging devices (e.g., EEG and fMRI) has been a barrier for developing this efficient assistive technology. Moreover, the development of this direct mapping approach has been hindered by lack of effective computational methods.

MEG records the magnetic field changes produced by electrical current flows in the brain [13], and is able to obtain real-time resolution recordings of brain activity with higher spatial solution than EEG or fMRI [14]. The unique data characteristics makes MEG suitable for brain disorders that are sensitive to time and space. In addition, MEG is quiet and thus friendly for users. Prior MEG studies examining speech perception portend promising outlooks for decoding speech production within brain activity signals [15, 16, 17]. However, decoding speech production from MEG signals has rarely been studied.

The aim of this project is to decode latent speech from single trial MEG sensor data during an overt production task using machine learning approaches together with noise reduction techniques. In addition, the use of data from the whole-head (all sensors) or from speech related regions (i.e., Wernicke's area, Broca's area, and motor cortex) were compared.



(a) The MEG system (b) Sensor map (top view)

Fig. 1. *Setup in data collection.*

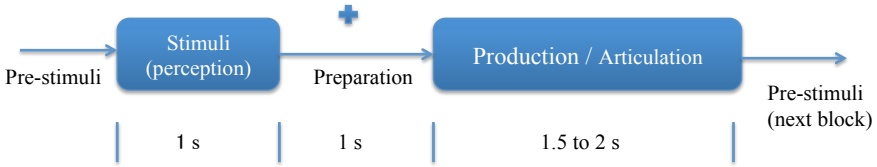


Fig. 2. *Segments within a trial (block or repetition) in the MEG recording*

2. DATA COLLECTION

2.1. Subjects and equipment

Two right-handed, male adults participated in this study. Both were fluent English speakers with normal or corrected to normal vision and no history of neurological or speech disorder.

Data acquisition was carried out at the MEG Center, Cook Children's hospital, Forth Worth, Texas. Neuromagnetic brain activity was recorded using a 306 channel Elekta Triux MEG machine, equipped with 204 planar gradiometers and 102 magnetometers (Figure 1a) and housed in a two layer magnetically shielded room (MSR). Data acquisition was performed at a 4kHz sampling rate with an online band-pass filter of 0.1 to 1300Hz. The sensor map is illustrated in Figure 1b. Subjects eye-blinks and cardiac signals were recorded via integrated bipolar EEG channels. Continuous head localization, a state-of-the-art technique for tracking subjects head position was used to monitor head motion. Voice production was recorded via a standard microphone attached to a transducer situated outside the MSR. Task related jaw movements were recorded by a custom air-pressure sensor connected to an air-filled bladder that was fixed to the subjects jaw in such a way as to cause depression during movement. Both voice and movement analog signals were fed into the MEG ADC channels and digitized in real-time as separate channels.

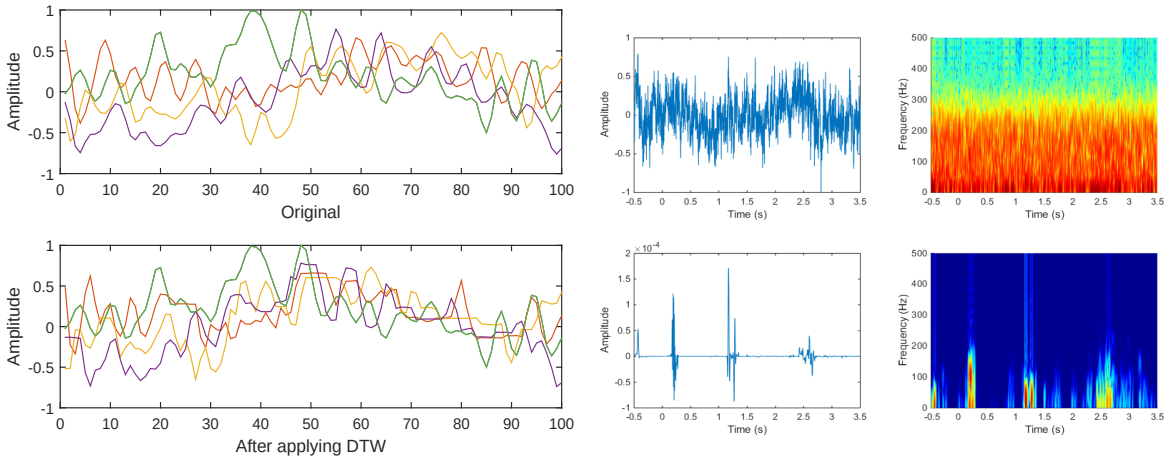
2.2. Procedure and task

Before recording, 3 fiducial points and 5 head-position-coils were digitized using Polhemus Fastrak for creating the subject coordinate system and head positioning in the MEG scanner, respectively. Subjects were then comfortably seated within the MEG unit with their arms resting on a table. A DLP projector connected to the stimulus computer displayed stimuli on a back-projection screen situated approximately 90cm in front of the subject. The task was a delayed overt reading task that consisted of pronouncing five short commonly used phrases: *how are you doing?*, *I am fine*, *I need help*, *That's good*, and *Good-bye*.

Phrase stimuli were presented on the screen for 1 second, followed by a 1 second fixation cross. Termination of the fixation heralded a blank screen that signaled the subject to overtly produce the phrase just previously shown. Subjects had 2 (up to 2.5) seconds to perform the speech before the next stimulus was presented. For most trials performance was accomplished within 1.5 seconds, providing for roughly 1 second of non-movement baseline before the start of the next trial (Figure 2). Each stimulus was presented 100 times in pseudo-randomized order to avoid response suppression to repeated exposure [18, 19]. Prior to starting the subjects were trained on sample stimuli to assure compliance. The entire experiment took approximately 45 minutes.

2.3. Data preprocessing

MEG sensor data were epoched into trials from -0.5 to +4.0 seconds centered on stimulus onset. Data preparation was performed in two steps. Data were visually inspected and trials containing erroneous movements that started either before the cue to speak or existed within the baseline period for the next trial were removed [18]. Of the remaining trials, those that contained excessive EOG or other large artifacts not related to movement were excluded. The data were then band-pass filtered between 1 and 250Hz and down-sampled to 1000Hz. These data were then forwarded for decoding analysis. Erroneous samples (e.g., due to wrong articulation) were



(a) Dynamic time warping (b) Wiener filtering

Fig. 3. *Illustration of (a) dynamic time warping and (b) Wiener filtering for MEG data processing.*

excluded. A total of valid 819 samples were collected from the two subjects.

3. DECODING METHODS

Before the MEG trials were fed into a classifier, three further processing steps were applied sequentially to remove the noise including dynamic time warping (DTW), wiener filtering, and gamma band filtering (30 - 100 Hz). Gamma band energy was estimated across the 50 ms time-series windows with 25 ms steps. Two machine learning classifiers - Gaussian mixture model (GMM) and artificial neural network (ANN) - were used as the classifiers/decoders.

3.1. Dynamic time warping (DTW)

DTW is arguably the best distance (or similarity) measure among time-series signals [20]. DTW calculates the summed Euclidean distances between the corresponding data points of two time-series signals, after aligning the peaks. DTW is particularly useful for signals with temporal variations (e.g., speech). In this paper, DTW was used as a data processing tool for removing the temporal variation in individual MEG trials. For each of the five phrases, the first trial was selected as the reference. Then, all other trials were warped to the references using DTW. Figure 3a shows a few (amplitude normalized) trials of one phrase (*how are you doing?*) in the preparation segment from one sensor (MEG1612) before and after applying DTW, where the green signal is the reference.

3.2. Wiener filtering

Wiener filtering is widely used in noise reduction for robust speech recognition. We adopted Wiener filtering to reduce MEG noise. A Wiener filter produces an estimate of target random process by linear-time invariant filtering of observed noisy process [21]. In this paper, we used the pre-stimuli segment data to estimate noise statistics. Figure 3b gives examples of Wiener filtering, where top row is the original signal (from sensor MEG1643, Wernicke's area) and its spectrogram; the bottom row is the filtered signal and filtered spectrogram (signals were amplitude normalized). The filtered signal and spectrogram clearly indicate the major events (visual, perception, and production).

3.3. Gaussian mixture model

Gaussian mixture model (GMM) is to model the data variation using Gaussian distribution, which has been used in speech recognition for decades [22]. Given Gaussian components, GMMs can model the relationship between features and target classes as a mixture of Gaussian density functions. GMM is a generative model and trained to represent as closely as possible the distribution (e.g., using means and variances) of training data. In this experiment, each class (phrase) has 10 Gaussians with diagonal covariances on average.

3.4. Artificial neural network

Artificial neural network (ANN) is a powerful non-linear computational modeling tool, used widely to model the complex relationship between inputs and targets. ANN is also widely used in pattern classification. In this paper, the input layer took 5 frames at a time (2 previous plus current plus 2 succeeding frames). Each frame had Gamma band energy from a frame size of 50 ms with a shift size of 25 ms. The output layer has 5 dimensions (5 phrases). The number of nodes in the hidden layer is 64 and the sigmoid activation function is used. The weights for nodes in the hidden layer at iteration $(t+1)$ are updated based on iteration (t) in a stochastic gradient descent way:

$$w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} \quad (1)$$

Table 1. Phrase classification results using data segments (%)

Method	Sensors (#)	Pre-stimuli	Perception	Preparation	Production
GMM	Whole-head (200)	48.59	77.39	77.98	88.88
	Speech related regions (20)	27.56	58.94	58.31	61.91
ANN	Whole-head (200)	36.34	86.38	90.38	94.54
	Speech related regions (20)	29.24	63.67	60.04	63.03

where w_{ij} is the weight between nodes i and j in neighboring layers, η is the learning rate, and C is the cross-entropy cost function. The implementation of GMM and ANN in Kaldi toolkit were used in this paper [23].

3.5. Classification experimental setup

Although the focus of this paper was decoding speech production, the classification was conducted on all segments including pre-stimuli (-0.5-0.0s), perception (0.0-1.0s), preparation (1.0-2.0s), and production (or articulation, 2.0-3.0s). Here, zero point is when stimulus was displayed. We hypothesized the results from the pre-stimuli segment will be low (close to chance level 20%); the results from other segments will be significantly higher than the chance level.

In addition, to compare the classification using the whole-head recording or the speech related regions, we compared the performances between using all 200 plannar gradiometers and only 20 selected sensors from the speech and motor related regions (i.e., Broca's area, Wernicke's area, and the motor cortex) [24]. Following are these sensors:

MEG1612, MEG1613, MEG1622, MEG1623, MEG1632, MEG1633, MEG1642, MEG1643, MEG0212, MEG0213, MEG0222, MEG0223, MEG0322, MEG0323, MEG0332, MEG0333, MEG0342, MEG0343, MEG0412, MEG0413

Four-fold cross validation was used in this experiment, where a quarter of the data for testing and the rest for training in a validation. The average classification accuracy (%) of the four validations was the overall performance. In this exploratory stage, the classification experiment was executed within each subject (speaker-dependent classification).

4. RESULTS AND DISCUSSION

Table 1 shows the average classification accuracy across the two subjects for each segment using the GMM and ANN methods with either 200 whole-head channels or the 20 selected channels. The accuracy for decoding during the pre-stimulus period was at or just above chance, whereas that of the perception, preparation, and production periods were all significantly higher [25], indicating the potential for decoding speech information from single-trial MEG signals. Additionally, results were the best using the ANN method with the whole head data during the production period (94.54%).

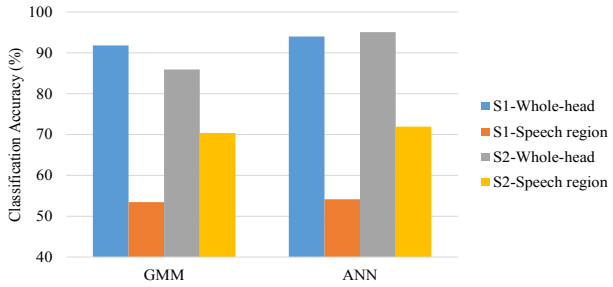


Fig. 4. *Phrase classification results on individual subjects (S1 and S2) using the production/articulation segment data.*

The fact that decoding with whole-head analysis outperformed selected sensors may indicate that a distributed network is involved in providing salient features for speech production [24]. On the other hand, we observed slightly higher (than chance level) decoding accuracy in the pre-stimulus segment for whole-head analysis as well, which suggested a possible slight over-fitting. The degree of over-fitting (if any) will be explored in further analysis. Last, ANN generally outperformed GMM in all configurations for all individual subjects (Figure 4).

5. RELATION TO PRIOR WORK

To our knowledge, this is the first non-invasive or MEG study to examine neural decoding of overt speech production. As mentioned previously, prior work using MEG has focused on speech perception (e.g., decoding speech from the perception segment) [15, 16, 17]. A recent study showed promising results in decoding speech production from invasive (ECoG) signals [26].

6. CONCLUSION AND FUTURE WORK

This paper demonstrated the possibility of decoding speech production from single trials of non-invasive (MEG) signals. A larger data set will be used to verify these findings.

7. ACKNOWLEDGMENTS

This project was supported by the University of Texas System Brain Research Initiative. We thank Dr. Mark McManis, Beiming Cao, Saara Raja, and the volunteering participants.

References

[1] E. Smith and M. Delargy, "Locked-in syndrome," *British Medical Journal*, vol. 330, no. 7488, pp. 406–409, 2005.

[2] R. Kohnen, J. Lavrijsen, J. Bor, and R. Koopmans, "The prevalence and characteristics of patients with classic locked-in syndrome in dutch nursing homes," *Journal of Neurology*, vol. 260, no. 6, pp. 1527–1534, 2013.

[3] D. Lulé, C. Zickler, S. Häcker, M.-A. Bruno, A. Demertzi, F. Pellas, S. Laureys, and A. Kübler, "Life can be worth living in locked-in syndrome," *Progress in Brain Research*, vol. 177, pp. 339–351, 2009.

[4] E. Buch, C. Weber, L. G. Cohen, C. Braun, M. A. Dimyan, T. Ard, J. Mellinger, A. Caria, S. Soekadar, A. Fourkas, et al., "Think to move: a neuromagnetic brain-computer interface (bci) system for chronic stroke," *Stroke*, vol. 39, no. 3, pp. 910–917, 2008.

[5] N. Birbaumer, "Brain–computer-interface research: coming of age," *Clinical Neurophysiology*, vol. 117, no. 3, pp. 479–483, 2006.

[6] N. Birbaumer and L. G. Cohen, "Brain–computer interfaces: communication and restoration of movement in paralysis," *The Journal of Physiology*, vol. 579, no. 3, pp. 621–636, 2007.

[7] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[8] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[9] D. J. McFarland, W. A. Sarnacki, and J. R. Wolpaw, "Electroencephalographic (EEG) control of three-dimensional movement," *Journal of Neural Engineering*, vol. 7, no. 3, pp. 036007, 2010.

[10] I. Iturrate, J. M. Antelis, A. Kubler, and J. Minguez, "A noninvasive brain-actuated wheelchair based on a p300 neurophysiological protocol and automated navigation," *IEEE Transactions on Robotics*, vol. 25, no. 3, pp. 614–627, June 2009.

[11] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain–computer interfaces for speech communication," *Speech Communication*, vol. 52, no. 4, pp. 367–379, 2010.

[12] F. Nijboer, E. Sellers, J. Mellinger, M. Jordan, T. Matuz, A. Furdea, S. Halder, U. Mochty, D. Krusienski, et al., "A p300-based brain–computer interface for people with amyotrophic lateral sclerosis," *Clinical Neurophysiology*, vol. 119, no. 8, pp. 1909–1916, 2008.

[13] D. Cohen, "Magnetoencephalography: detection of the brains electrical activity with a superconducting magnetometer," *Science*, vol. 175, no. 4022, pp. 664–666, 1972.

[14] M. Proudfoot, M. V. Woolrich, A. C. Nobre, and M. R. Turner, "Magnetoencephalography," *Practical Neurology*, vol. 14, pp. 336–343, 2014.

[15] P. Suppes and B. Han, "Brain-wave representation of words by superposition of a few sine waves," *Proceedings of the National Academy of Sciences*, vol. 97, no. 15, pp. 8738–8743, 2000.

[16] A. M. Chan, E. Halgren, K. Marinkovic, and S. S. Cash, "Decoding word and category-specific spatiotemporal representations from MEG and EEG," *Neuroimage*, vol. 54, no. 4, pp. 3028–3039, 2011.

[17] M. P. Guimaraes, D. K. Wong, E. T. Uy, L. Grosenick, and P. Suppes, "Single-trial classification of MEG recordings," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 3, pp. 436–443, 2007.

[18] D. Cheyne and P. Ferrari, "MEG studies of motor cortex gamma oscillations: evidence for a gamma fingerprint in the brain?," *Frontiers in Human Neuroscience*, vol. 7, no. 575, pp. 1–7, 2013.

[19] K. Grill-Spector, R. Henson, and A. Martin, "Repetition and the brain: neural models of stimulus-specific effects," *Trends in Cognitive Sciences*, vol. 10, no. 1, pp. 14–23, 2006.

[20] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing dtw to the multi-dimensional case requires an adaptive approach," *Data Mining and Knowledge Discovery*, pp. 1–31, 2016.

[21] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098–2108, 2006.

[22] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, vol. 14, PTR Prentice Hall Englewood Cliffs, 1993.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and V. K., "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, Waikoloa, USA, 2011, pp. 1–4.

[24] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation," *Nature*, vol. 495, no. 7441, pp. 327–332, 2013.

[25] E. Combrisson and K. Jerbia, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of Neuroscience Methods*, vol. 250, pp. 126–136, 2015.

[26] 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*, vol. 9, no. 217, 2015.