Decoding Speech Evoked Jaw Motion from Non-invasive Neuromagnetic Oscillations

Debadatta Dash
Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas, United States
debadatta.dash@utexas.edu

Paul Ferrari
Department of Psychology
University of Texas at Austin
Austin, Texas, United States
pferrari@utexas.edu

Jun Wang
Communication Sciences and Disorders
University of Texas at Austin
Austin, Texas, United States
jun.wang@austin.utexas.edu

Abstract*—Speech decoding-based brain-computer interfaces (BCIs) are the next-generation neuroprostheses that have the potential for real-time communication assistance to patients with locked-in syndrome (fully paralyzed but aware). Recent invasive speech decoding studies have demonstrated the possibility of speech kinematics decoding, where articulatory movements were decoded from the brain activity signals for speech synthesis, as an alternative solution to direct brain-to-speech mapping. As a starting point toward a non-invasive speech-neuroprosthesis, in this study, we investigated the decoding of continuous jaw kinematic trajectories directly from non-invasive neuromagnetic signals during speech production. The compensatory jaw behavior exhibited by patients with amyotrophic lateral sclerosis (ALS) is prevalent, hence, accurate decoding of the jaw kinematics could be a path for developing efficient communicative BCIs for these patients. Using magnetoencephalography (MEG), we recorded brain signals and jaw motions simultaneously from four subjects as they spoke short phrases. We trained a long short-term memory (LSTM) regression model to successfully map the brain activity to jaw motion with about 0.80 average correlation score across all four subjects. In addition, we also examined the decoding performance of specific frequency bands within the neural signals and found that the Delta ($0.3-4$ Hz) and high-gamma ($62-125$ Hz and $125-250$ Hz) frequencies independently can account for the major contributions in jaw motion decoding. Experimental results indicated that the jaw kinematics can be successfully decoded from non-invasive neural (MEG) signals.

***Index Terms*—BCI, LSTM, MEG, Brainwaves, Wavelets**

I. Introduction

Speech production is one of the most exquisite dynamically coordinated physiological phenomena in the human behavioral repertoire. It involves synergistic control between cortical brain regions and motor units and of overlapping, multi-articulatory vocal tract movements for transcribing thoughts into meaningful sounds. The brain orchestrates more than a hundred muscles and is continuously shaping and reshaping the articulators (lips, tongue, jaw, larynx, etc.) across time to produce unique vocal tract patterns, contextualizing communication [1] in form of a repertoire of overt speech sounds with simultaneous auditory feedback. Brain damage or neurodegenerative diseases (e.g., amyotrophic lateral sclerosis, ALS) may cause locked-in syndrome (completely paralyzed but aware) [2]. A brain-computer interface (BCI), that uses brain activity to control a computer without involving muscles, is currently a more preferred and reliable option [3], [4]. Yet, current commercially available BCIs use attention correlates from the users' brain to spell out words, letter by letter, which results in a very slow communication rate of under 10 words per minute, far slower than the normal speaking rate, which is about 200 words per minute. A major challenge but necessary requirement today is to move beyond these slow, error-prone, and laborious spelling based constrained technologies toward more efficient speech-BCIs with possibly normal communication rates.

Speech-BCI is a next-generation communication rehabilitative technology, which attempts to translate neural signals to speech in real-time. This transformative speech neuroprosthesis has the potential to offer an improved quality of life to neurologically impaired patients, potentially enabling independence, social interactions, and community involvement to some level by restoring lost communication [5]. Multiple research studies have proposed to decode both overt and covert speech directly from neural signals (neural speech recognition) either invasively with electrocorticography (ECoG, [6]–[9]) or non-invasively with electroencephalography (EEG, [10]–[13]) and magnetoencephalography (MEG, [14]–[18]). The majority of these decoding studies, however, have focused on classifying isolated speech units (phonemes/syllables) directly from the neural signal, which falls short of the ultimate goal of neural speech synthesis. Recently, a few ECoG studies have shown promise for neural speech synthesis [19]–[21]. In the ECoG study [21], discrete representations of articulatory movements were decoded from neural signals and then were used to synthesize speech (brain to articulation to speech).

Majority of articulation decoding studies have focused either on the classification of discrete articulatory features (e.g., opening vs closing) [22]–[24] or on decoding articulatory motions that were inversely mapped from acoustic data [21], [25]. An ECoG study for implantable BCIs [22] showed successful decoding of four different tongue movement directions (up, down, left, and right) with 85% classification accuracy by taking data from just $1\,\text{cm}^2$ area of the sensory-motor cortex from four subjects. Another study [23] produced a higher decoding performance for articulatory gesture classification

This work was supported by the University of Texas System Brain Research Grant under award number 362221 and the National Institutes of Health (NIH) under award numbers R03DC013990 and R01DC016621.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

than phonemes with electrocorticographic signals recorded from the pre-motor and motor cortex of two subjects, further strengthening the importance of articulatory kinematics decoding for a brain-machine interface. Decoding lip movements during speech production has also been investigated in a recent study [24] showing a 65% accuracy for open vs closed-lip position classification with ECoG signals. In the ECoG studies [21] and [25], the used articulatory kinematics were inversely mapped from acoustic signals that were trained from other speakers, because simultaneous acquisition of ECoG recordings and articulatory kinematics was not available. To our knowledge, there is no prior report on the successful decoding of real-continuous articulatory motion based on data recorded synchronously with brain activities.

In this study, we were able to collect simultaneous recordings of neural and jaw motion signals, which might provide a better and more reliable ground truth for the brain to kinematics mapping compared to the statistically estimated kinematics. Also, in contrast to the discrete classification-based articulation decoding as in most of the previous studies (except [21]), here, we performed continuous mapping of neural signal to articulatory (jaw) kinematics for each millisecond (sample). Moreover, instead of the conventional paradigm of collecting data during short speech unit production (phonemes/syllables/words), we collected the neuromagnetic signals corresponding to complete sentences, which is necessary to leverage coarticulation (dependency of articulatory gestures for current speech segment on previous and future speech segments) [26]. Furthermore, previous studies have only considered the high-gamma neural oscillations to perform the decoding tasks, which is justified, considering the large correlation of high-gamma ECoG activity with multi-unit firing rates [27]. However, articulatory movements have also been shown to correlate with frequencies under 40 Hz [28], [29]. Thus, we also explored the contribution of multiple oscillatory frequencies for jaw motion decoding.

We used magnetoencephalography (MEG) to record the neural signals during speech production, synchronously with jaw motion and acoustic speech. MEG is a non-invasive functional neuroimaging modality that records the post-synaptic neuronal current-induced magnetic fields with a very high spatial ($3-10$ mm) and temporal (1 ms) resolution. Although current MEG is not-portable, recent studies on next-generation movable MEG devices based on optically pumped magnetometers [30] have the potential to be used for BCI applications in the near future. Moreover, our previous works on both overt and imagined speech decoding have resulted in high classification accuracy [16]–[18] on closed-set classification tasks with MEG signals. Here, we used a sequential deep learning regression model with long short-term memory (LSTM)-recurrent neural networks as the decoder considering their efficacy in analyzing sequential time series data.

In summary, the major contributions of this study are:

- Decoding of continuous articulatory kinematics (jaw motion) was attempted from non-invasive neural (MEG) signals, in contrast to the previous works with invasive ECoG signals [21]–[25].
- Simultaneously acquired neural signals (via MEG) and jaw motion (via a customized pressure sensor) were used contra prior studies that obtained articulatory data with inverse mapping model [21], [25], or videos [24].
- A continuous regression model was developed to synthesize the neural jaw motion in real-time instead of the prior works of classifying discrete movement events [22]–[24].
- Investigations on the role of each neural oscillations for jaw kinematics decoding was performed to find significant contributions by Delta ($0.3-4$ Hz) and high-gamma ($62-125$ Hz and $125-250$ Hz) frequencies.

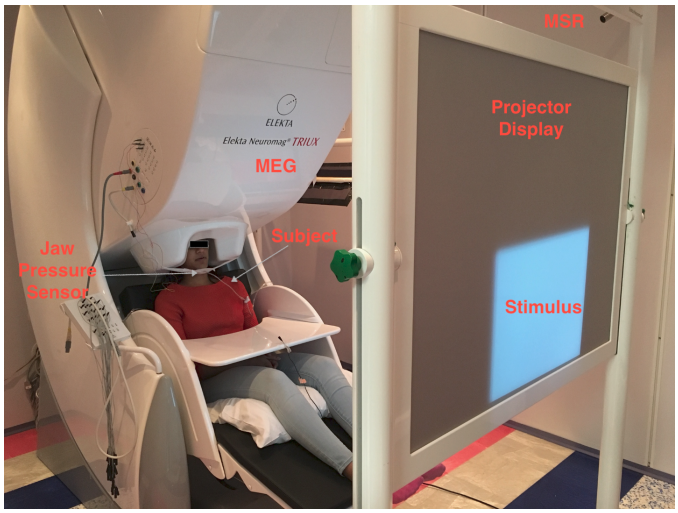

Fig. 1. The MEG unit with a subject

II. Data Collection and Signal Processing

A. Data Acquisition

Four healthy subjects (age: 48 ± 14; 1 female) participated in the study with informed consent in compliance with the institutional ethics review boards. We used a 306 channel (204 gradiometers + 102 magnetometers) MEG device (MEGIN, LCC) to collect the neuromagnetic signals from the subjects (Figure 1). The MEG unit is housed inside a magnetically shielded room (MSR) for discarding unwanted magnetic interference. We designed a time-locked protocol consisting of a 0.5 ms of pre-stimuli stage, followed by a 1 s of perception stage, where a sample stimulus (phrase) was presented to the subjects written in English via a projector display situated at about 90 cm from the subject. This stage was followed by a preparation (imagination) stage of 1 s where a fixation cross (+) appeared on the screen, after which the subjects overtly produced the previously shown phrase. The overt production stage was designed to be of 2.5 s (except for the very first subject it was 1.5 s). This 4-stage procedure constituted a trial, and for each subject, we collected data for 100 trials per phrase. We kept a non-movement baseline of $1-1.5$ s within successive trials. The subjects spoke 5 different phrases: *1. Do you understand me; 2. That's perfect; 3. How are you; 4. Good-bye; 5. I need help*; which were displayed on the

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

screen one at a time in a pseudo-randomized order to avoid response suppression due to repeated exposure [31]. Acoustic output during the speech production stage was recorded via a standard built-in microphone connected to a transducer placed outside the MSR. Jaw movement was recorded through a custom-made air bladder connected to an air pressure sensor. The depression in that bladder provided the trajectory for jaw motion. Both speech and jaw movement analog signals were then digitized by feeding into the MEG ADC in real-time as separate channels. All sensors were checked for noise and tuned prior to data collection. The whole experiment lasted for about 45 minutes per subject, excluding the break time (if any) and the preparation time. All default safety practices for MEG were followed [32].

B. Preprocessing

The signals were epoched into trials from -0.5 to 5 s (subject 1: -0.5 to 4 s) centered on stimulus onset. Data, only from gradiometer sensors, were used considering their effectiveness in noise suppression over magnetometers. Out of 204 gradiometers, data from 8 sensors were discarded due to high channel noise. Through Visual inspection, trials containing high artifacts or irregularities were discarded. On average, 75 trials per phrase per subject were retained. The signals were recorded with 4 kHz sampling frequency with an online filter of $0.3-1000$ Hz, low-passed to 250 Hz with a 4^{th} order Butterworth filter, and resampled to 1 kHz. Line noise (60 Hz) and their harmonics were removed with notch filters.

C. Wavelet Decomposition

The preprocessed signals were decomposed with discrete wavelet transform (DWT) using a Daubechies wavelet (db-4) with 7 levels and reconstructed back to each level for generating distinct neural oscillations. The use of db4 to generate distinct neural oscillations has been employed previously in numerous MEG studies [33]–[35]. After decomposition, the reconstructed signal from the low-pass approximation coefficient at 7^{th} level was the Delta (D) frequency band ($0.3-4$ Hz), and the reconstructed signals from the high-pass coefficients from each level starting from level 7 to level 2 corresponded to Theta (T) ($4-8$ Hz), Alpha (A) ($8-16$ Hz), Beta (B) ($16-30$ Hz), Gamma (G) ($31-58$ Hz), lower high-gamma (L-HG) ($62-125$ Hz), and upper high-gamma (U-HG) ($125-250$ Hz) brainwaves respectively. The frequency ranges for brainwaves vary across studies, particularly with ECoG, high-gamma frequency has been considered to be in the range of $70-200$ Hz. However, we considered the above-mentioned frequency ranges and divided the high gamma frequency range into 2 parts as L-HG ($62-125$ Hz), and U-HG ($125-250$ Hz).

III. Method

We developed an LSTM regression model to map the gradiometer sensor signals to the corresponding jaw motion irrespective of the phrases. Considering the cognitive variance across subjects [16], [36], in this study, we only focused on subject-dependent jaw kinematics decoding. Previous articulatory kinematics decoding studies have [21], [25] used recurrent models (with Bidirectional-LSTM, BLSTM). Since both MEG and the corresponding jaw motion signals are sequential in nature, we used the LSTM recurrent neural network, eying to the future work on real-time decoding. However, for comparison, BLSTM regression was also performed.

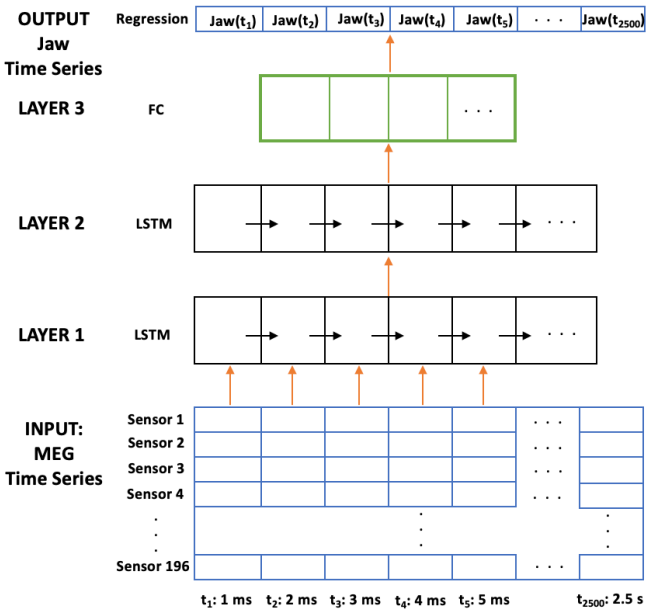

Fig. 2. Architecture of the LSTM regression model.

A. Model Architecture

We developed a 4-layered stacked LSTM architecture with initial 2 layers containing recurrent LSTM units feeding to a fully connected (FC) layer on the top followed by an output regression layer (Figure 2). The individual memory blocks of LSTM contained the default architecture of cell-gate structure and activations. Hidden layers 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 the respective order. A hard-sigmoid activation function was used to update the gate state and tanh for the cell and hidden states. The input data to the model was the z-score normalized 196-dimensional gradiometer signals trained to regress to 1-dimensional jaw signal at each sample. The ground truth for the regression model was the jaw motion signal, smoothed with a 50 ms window-based moving average filter and z-score normalized. Out of 75 averagely retained trials per phrase, 50 trials per phrase were used for training, and 12 trials per phrase for testing. The remaining trials were used as development data for hyperparameter tuning. A separate regression model was developed for each subject. Hyperparameters, including the number of LSTM units in each layer varied for subject to subject based on least RMSE score on development data. For

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE I
MODEL ARCHITECTURE AND HYPERPARAMETERS

Components	Details			
	Sub 1	**Sub 2**	**Sub 3**	**Sub 4**
Input	MEG Signals			
Input layer dimension	196			
Sampling rate	1 kHz			
Input time points	1500	2500	2500	2500
Num. of training samples	300	300	300	300
Number of dev samples	100	25	55	85
Output	Jaw motion signal			
Output layer dimension	1			
Output time points	1500	2500	2500	2500
Number of test samples	60	60	60	60
Regression Model	LSTM-RNN			
Depth	4 layers			
LSTM Units in Layer 1	512	640	576	640
LSTM Units in Layer 2	64	256	256	192
Number of FC nodes	50	50	50	50
Dropout	0.5	0.5	0.5	0.5
Batch size	32	32	32	32
Maximum epochs	100	40	40	50
Optimizer	ADAM			
Training method	BPTT			
Initial learning rate	0.005	0.005	0.008	0.007
Learning rate drop factor	0.5	0.5	0.5	0.5
Learning rate drop period	10 epochs			
Gradient threshold method	L2 Norm			
Gradient threshold value	.1			
beta1	0.9			
beta2	0.999			
epsilon	1.00E-08			
Loss function	RMSE			

unbiased splitting, the model was trained on 3 different splits with the same number of trials in each set. A 5-fold cross-validation on training data was also performed to check for model overfitting on development data. For performance comparison with bidirectional recurrent models, LSTM units were replaced with BLSTM units, with optimized hyperparameters.

B. Model Hyperparameters

Hyperparameters play a crucial role in optimizing model performance. Since this is a subject-dependent study and the neural data corresponding to each subject was different from one another, model development and hyperparameter tuning were performed separately for each subject. Table I enlists the details of model architecture and hyperparameters for each subject. The loss function was root mean square error (RMSE), trained with an ADAM optimizer via back-propagation through time (BPTT) with fixed $\beta 1$, $\beta 2$, and ϵ values of 0.9, 0.999, and 1E-8 respectively. The initial learning rate was tuned with a coarse to fine setting within the range of values of 0.1, 0.01, 0.001, 0.0001. The learning rate was set to be halved with 10 epochs. Batch size was tuned for values of 16, 32, 64, 128, and 256. The number of LSTM units in the recurrent layers were tuned with a grid search within ranges between 64 to 1024 with increments of 64. The number of FC nodes was tuned for values from 10 to 100 with increments of 10. A 50% dropout probability was also used on the FC layer for regularization, which was tuned from 10% to 60% with 10% steps. The maximum number of epochs was tuned for values from 40 to 150 with increments of 10 up to loss convergence. L_2-Norm based gradient threshold was used with a threshold value of 0.1. The final hyperparameter values were chosen based on the least RMSE score on development data.

IV. RESULTS

Figure 3 shows the best-predicted jaw kinematics plotted on top of originally recorded signal in a z-score normalized space for all 4 subjects. The corresponding Pearson correlation score and RMSE values in z-score space are given as the titles. For a more intuitive understanding of the prediction performance, correlation score (*r*) was taken as the objective measure, similar to previous studies [21], [24]. All results reported are the average of 3 different data splits, each with 3 runs. The mean correlations across all test trials for each subject $(1-4)$ were 0.90 ± 0.089, 0.72 ± 0.17, 0.74 ± 0.20, 0.82 ± 0.15, leading to an average of about 80% correlation score across 4 subjects (also shown as the diagonal elements of Table III). This result and visualizing the best predictions ($r > 0.95$) (Figure 3) show that it's possible to directly map the neural signals onto the jaw kinematics space. However, there were a few trials for which the predictions were extremely bad, even negative. There were about 4 trials on an average per subject for which the correlation scores were below 0.5, which brought down the mean scores.

The best prediction from each subject was for the trial when subjects spoke '*Good-bye*'. Compared to the 5 phrases used in this study this phrase was the shortest and probably LSTM was better able to predict the jaw kinematics of short lengths compared to the long ones. This led us to investigate the phrase level correlation score for each subject which is shown in Table II. We found that there was no significant difference between predicting '*Good-bye*', the shortest phrase, and '*Do you understand me*', the longest phrase (2-tail t-test, $p > 0.05$). Also, the average prediction score per phrase varied across subjects. For instance, the best prediction for subject 1 was for the phrase 'How are you' ($r = 0.9613$) but for subject 2 it was 'I need help' ($r = 0.7967$). These results (Figure 3; Table II) are when MEG signals with all neural oscillations $(0.3 - 250\,\text{Hz})$ were trained.

We also examined the contribution of each neural oscillations: Delta (D) $(0.3-4\,\text{HZ})$, Theta (T) $(4-8\,\text{Hz})$, Alpha (A) $(8-16\,\text{Hz})$, Beta (B) $(16-30\,\text{Hz})$, Gamma (G) $(31-58\,\text{Hz})$, lower high-gamma (L-HG) $(62-125\,\text{Hz})$, and upper high-gamma (U-HG) $(125-250\,\text{Hz})$, for jaw motion regression. The results are shown in Figure 4 which represents the distribution of correlation scores across trials obtained with each neural oscillations individually and combined (All) as box-plots. The mean and median correlation score was always the highest when all the frequencies were combined to train the model, compared to when trained with individual brainwaves. Across 4 subjects, the mean correlations were 0.80 ± 0.15, 0.51 ± 0.25, 0.49 ± 0.22, 0.41 ± 0.29, 0.26 ± 0.27, 0.23 ± 0.30, 0.27 ± 0.32, and 0.77 ± 0.16 for all, upper high-gamma, lower high-gamma,

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

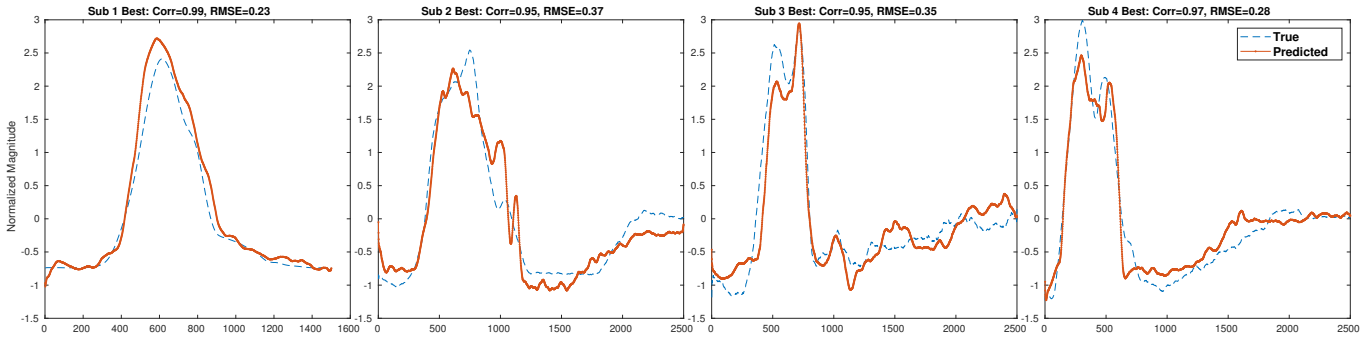

Fig. 3. Trials with the best regression performance for the four subjects

TABLE II
PHRASE LEVEL MEAN CORRELATION SCORES FOR ALL FOUR SUBJECTS

	Do you understand me	**That's perfect**	**How are you**	**Good-bye**	**I need help**
Subject 1	0.8715	0.7797	0.9613	0.9570	0.9194
Subject 2	0.7348	0.6574	0.7491	0.6405	0.7967
Subject 3	0.7967	0.8070	0.7693	0.7964	0.5346
Subject 4	0.6528	0.9225	0.8533	0.8221	0.8692
Average	***0.7640***	***0.7917***	***0.8333***	***0.8040***	***0.7800***
STD	*0.0928*	*0.1088*	*0.0966*	*0.1298*	*0.1712*

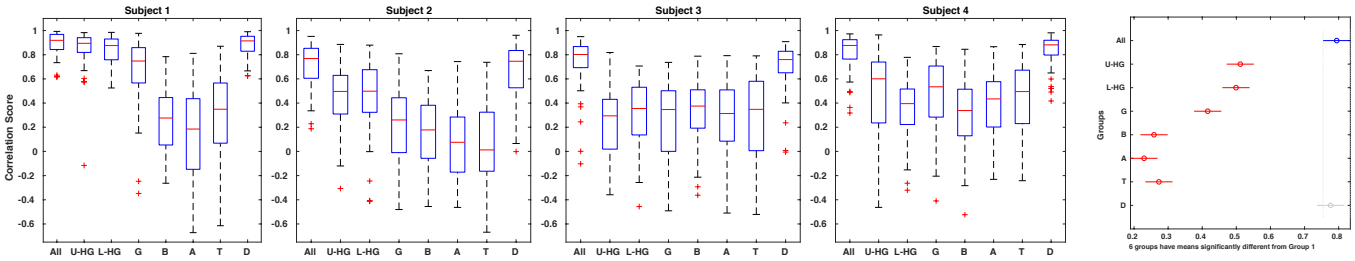

Fig. 4. Distribution of correlation values between true and predicted jaw motion signals across all test trials when trained with MEG signals containing all neural oscillations (All) (group 1), and individual neural oscillations (group $2-8$): upper-high gamma (U-HG), lower-high gamma (L-HG), Gamma (G), Beta (B), Alpha (A), Theta (T), and Delta (D), are shown for subject 1 to 4 in the left 4 figures. Right most figure represents the statistical differences between 8 groups across 4 subjects computed with 1-way ANOVA

Gamma, Beta, Alpha, Theta, and Delta frequency respectively. Interestingly, Delta frequency performed equally good as all brainwaves combined. The results with gamma frequency bands (U-HG, L-HG, and G) were satisfactory but with the rest, the predictions were not as good.

Statistically, a 1-way ANOVA with the 8 groups (All, U-HG, L-HG, G, B, A, T, D) across 4 subjects (Figure 4 rightmost) showed a significant difference between the decoding performances obtained with Delta and other oscillations ($p < 0.05$). Also, a statistical difference was found between 'All' and the rest of the oscillations ($p < 0.05$) except when 'All' was compared to Delta ($p = 0.9974$). U-HG and L-HG were not significantly different ($p = 0.9997$) but were different than the rest ($p < 0.05$). The performances of Theta, Alpha, and Beta were the lowest, significantly different than the rest ($p < 0.05$). The mean correlation score was better for subject 1 compared to the rest 3 subjects, probably due to the task of predicting a shorter period (subject 1: 1.5 s, subject $2-4$: 2.5 s)

V. DISCUSSION

A. *Efficacy of Recurrent LSTM Regression*

The unique arrangement of memory blocks in the LSTM recurrent neural network model helps in performing additive interactions to improve gradient flow over long sequences of time series. Thus, for modeling sequential time series data LSTM has been a popular approach, and has been applied for modeling EEG [37]–[39] or MEG [40]–[42] based neural signals. However, using LSTM-RNN as a regression model for MEG signals has not been explored before. A BLSTM decoder was used in [21] for ECoG to articulation mapping. BLSTM considers both future and past samples to predict the outcome of the present whereas LSTM only considers the past samples and thus can be modeled for a real-time decoder. However, since our ultimate goal is to synthesize speech in real-time, we focused on an LSTM model. For comparison, we also trained a BLSTM model with optimized hyperparameters for jaw kinematics decoding.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

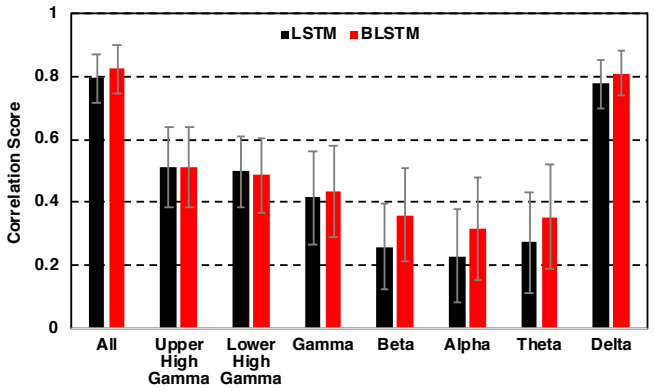

Fig. 5. Decoding Performance of LSTM v. BLSTM. Error bars indicate standard errors across 4 subjects.

Figure 5 shows the comparison of correlation scores averaged across 4 subjects with LSTM and BLSTM decoders trained at specific neural oscillations. Although there was no statistically significant difference between the two decoders, BLSTM provided better correlation scores than LSTM, except for the high gamma oscillations. Comparing our LSTM-RNN results with the EcoG study in [21], we observed that the results obtained in this study were equally good (perhaps better) which further strengthens the possibility of using non-invasive neuromagnetic signals for speech neuroprostheses. However, the decoder architectures are different in these two studies, and also, we utilized non-invasive neural signals across the whole brain in contrast to the selective invasive sampling of the ECoG study where only a part of the brain was analyzed.

B. Importance of Decoding Jaw Kinematics

Aiming for articulatory-speech neuroprostheses as the next-generation speech-BCIs [21] is the key to restoring communication for ALS patients. In contrast to the difficulty of collecting simultaneous articulatory data along with neural signals via ECoG, a MEG setup has the potential to collect both articulatory kinematics and non-invasive neuromagnetic signals in parallel with a MASK (Magnetoarticulography for the Assessment of Speech Kinematics) [43]. Similarly, in this study, we used the traditional MEG set up and used a custom air bladder connected to an air pressure sensor to simultaneously record the jaw movement and brain activity during continuous speech production. Although only jaw motion data were investigated in this study, it is important for ALS patients. Jaw kinematics have been studied in ALS patients which show specific compensatory changes during disease progression [44]–[47]. Both transient and non-vowel specific changes in jaw kinematics have been shown to be more prevalent than the other articulators [45] in ALS patients. Thus, accurate decoding of jaw kinematics will be extremely valuable for developing speech-BCIs for these patients. The mean performance obtained with this study ($r = 0.80$) is significant enough to motivate additional research using non-invasive methods of articulatory speech decoding. More data

TABLE III
PERFORMANCE OF EACH SUBJECT WITH DIFFERENT SUBJECT MODEL

	Model Hyperparameter Choice			
Evaluation Data	**Subject 1**	**Subject 2**	**Subject 3**	**Subject 4**
Subject 1	**0.8978**	0.8799	0.8898	0.8821
Subject 2	0.6284	**0.7157**	0.5826	0.6516
Subject 3	0.6831	0.6974	**0.7402**	0.7275
Subject 4	0.7823	0.8183	0.8237	**0.8240**

with better models (e.g. Sequence-to-Sequence translation) or deep neural network features [48] could provide better performance. With MEG, both jaw and lips data can be collected simultaneously either with MASK [49] or with the current MEG set up (MEG+ Jaw pressure sensor) along with a video camera (equivalent to [24]), after which continuous real-time speech synthesis can be possible.

C. Role of Neural Oscillations in Jaw Kinematics Decoding

As shown in Figure 4, combining all the neural oscillations resulted in the best performance. This result was expected as the contribution of Gamma and high gamma frequencies for speech-motor movement [27] as well as the effectiveness of low-frequencies in speech decoding [13], has been previously shown in the literature. The interesting observation was the high performance obtained with the upper high-gamma and Delta frequency bands. Although it can be argued that the high-frequency muscle artifacts and the low-frequency movement artifacts might have influenced this behavior with upper high-gamma and Delta frequency respectively, the consistency of high performance across trials and subjects for these two brainwaves were significant. Including these signals in the analysis would thus still be beneficial. Also, it should be noted that most of the decoding results (correlation scores) were positively skewed (median $>$ mean) and the outliers were negative. This begs for a better automatic trial rejection strategy, which may improve the performance significantly.

D. Efficacy of Hyperparameter Tuning

The hyperparameters are a crucial factor in getting optimal performance. We tuned all the hyperparameters separately for each subject based on the performance with the development data. The learning rate and the number of nodes in LSTM layer 1, were found to be the most important hyperparameters for jaw motion regression. Based on the coarse-to-fine tuning strategy of learning rates, an increase in mean correlation score (and decrease in RMSE) was observed when the learning rate was decreased from 0.1 to $0.005 - 0.008$ and below that the performance started to decrease. An average of 0.08 increase in correlation score was observed from standard to the optimized setting. In regards to the LSTM units, layer 1 units were found to be more contributing than layer 2. On average, an increase of 0.12 correlation score was observed when the number of LSTM units was increased from 64 to $512 - 640$ in the first layer. The results were better when the number of nodes for the second layer was less compared to the first. Additional LSTM layers did not increase performance,

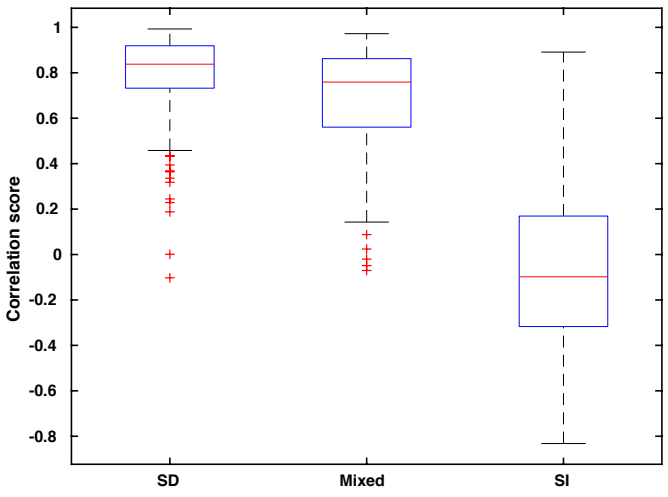

Fig. 6. Subject-dependent (SD) v. mixed v. subject-independent (SI) jaw decoding performance

probably due to the small number of data samples. Adding dropouts to LSTM layers did not contribute to better performance, however, adding dropouts after the FC layer increased the correlation score by 0.03. A dropout probability of 0.5 was found to be the best. The rest of the hyperparameters were insignificant in tuning. It was necessary to tune hyperparameters for each subject separately as evident from Table III which shows the average correlation score of the predicted and recorded jaw motion signal, obtained by training the MEG signals with all oscillations ($0.3 - 250$ Hz) of each subject with the tuned model of other subjects. Taking subject 2 as an example, the average decoding performance was 0.72 with its own model (learning rate=0.005, layer 1 nodes=640, and layer 2 nodes=256), which decreased to 0.58 when trained with the hyperparameters tuned for subject 3 (learning rate=0.008, layer 1 nodes=576, and layer 2 nodes=256). This also shows the significance of the learning rate and the number of nodes in layer 1 in the used LSTM regression model.

E. Subject Independent Jaw Kinematics decoding

Cognitive-behavioral variance across subjects makes it difficult to generalize a subject-independent neural data-based model [16], [36]. Thus, most of the BCI works in the field are done by developing decoders for each subject. To highlight this issue in our data, we developed a subject-independent (SI) model with the same architecture and evaluated with 3 fold cross-validation (train with 2 subjects and test with 1, repeated for 3 unique shuffles). Only 3 subjects (subject $2-4$) were used in this SI experiment since for subject 1, the number of time samples was different (subject 1: 1.5 s, rest 3 subjects: 2.5 s). We also developed a mixed model where we combined the data from all 3 subjects and held-out 20% data for testing (similar to subject dependent testing: $60/300$). A comparison of the performances of these 3 models (SD, mixed, and SI) is shown in Figure 6. Clearly, for SI model, the average correlation score was not as good ($r = -0.06$). The mean correlation score of the mixed model was satisfactory ($r = 0.69$) but still less than the SD model ($r = 0.80$). All of these results were with MEG signals including all the neural oscillations. More data and better adaptation strategy might improve the SI model performance, nonetheless, the motivation for a subject-dependent model is clear, and appropriate when considering the variance that is inherent in patient-specific disease progression for which these models can be tailored by online learning.

VI. Summary

In this study, we investigated the possibility of synthesizing jaw kinematics from non-invasive neural (MEG) signals as a crucial first step toward developing speech neuroprostheses for ALS patients. We acquired the jaw motion and neural signals simultaneously with an MEG setup and developed an LSTM regression model to efficiently predict jaw kinematics from the MEG signals with an average correlation score of 0.80 across four subjects. Our analysis of the contribution of specific neural oscillations indicated high efficacy when using high-gamma and Delta frequencies in jaw motion decoding. This study only included healthy subjects, and the approaches used here are needed to be tested with data from ALS subjects. Future work will focus on decoding articulatory kinematics including lips and tongue (in addition to the jaw) from the MEG signals of patients with ALS.

Acknowledgment

We thank Dr. Angel Hernandez, Dr. Saleem Malik, Kristin Teplansky, Hannah Weiner, Saara Raza, Dr. Alan Wisler, Beiming Cao, and the volunteering participants.

References

[1] W. J. Levelt, "Models of word production," *Trends in Cognitive Sciences*, vol. 3, no. 6, pp. 223 – 232, 1999.

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

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

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

[5] E. F. Chang and G. K. Anumanchipalli, "Toward a Speech Neuroprosthesis," *JAMA*, 12 2019.

[6] D. Moses, M. Leonard, J. Makin, and E. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nature Communications*, vol. 10, 12 2019.

[7] 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, p. 217, 2015.

[8] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "Direct classification of all American English phonemes using signals from functional speech motor cortex," *Journal of Neural Engineering*, vol. 11, no. 3, p. 035015, may 2014.

[9] S. F. Martin, P. Brunner, I. Iturrate, J. d. R. Millán, G. Schalk, R. T. Knight, and B. N. Pasley, "Corrigendum: Word pair classification during imagined speech using direct brain recordings," in *Scientific reports*, vol. 7, no. 44509, 2016.

[10] M. D'Zmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, "Toward EEG sensing of imagined speech," in *Human-Computer Interaction. New Trends*, J. A. Jacko, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 40–48.

[11] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 992–996.

[12] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features," *Journal of Neural Engineering*, vol. 15, no. 1, p. 016002, nov 2017.

[13] X. Chi and H. John, "EEG-based discrimination of imagined speech phonemes," *International Journal of Bioelectromagnetism*, vol. 13, no. 4, pp. 201–206, 01 2011.

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

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

[16] D. Dash, A. Wisler, P. Ferrari, and J. Wang, "Towards a Speaker Independent Speech-BCI Using Speaker Adaptation," in *Proc. Interspeech 2019*, 2019, pp. 864–868.

[17] D. Dash, P. Ferrari, D. Heitzman, and J. Wang, "Decoding speech from single trial MEG signals using convolutional neural networks and transfer learning," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2019, pp. 5531–5535.

[18] D. Dash, P. Ferrari, and J. Wang, "Decoding imagined and spoken phrases from non-invasive neural (MEG) signals." *Frontiers in neuroscience*, vol. 14, p. 290, 2020.

[19] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ECoG using densely connected 3D convolutional neural networks," *Journal of Neural Engineering*, vol. 16, no. 3, p. 036019, apr 2019.

[20] C. Herff, L. Diener, M. Angrick, E. Mugler, M. C. Tate, M. A. Goldrick, D. J. Krusienski, M. W. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Frontiers in Neuroscience*, vol. 13, p. 1267, 2019.

[21] G. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, pp. 493–498, 04 2019.

[22] E. Salari, Z. V. Freudenburg, M. P. Branco, E. J. Aarnoutse, M. J. Vansteensel, and N. F. Ramsey, "Classification of articulator movements and movement direction from sensorimotor cortex activity," in *Scientific Reports*, vol. 9, 2019, p. 14165.

[23] E. M. Mugler, M. Goldrick, J. M. Rosenow, M. C. Tate, and M. W. Slutzky, "Decoding of articulatory gestures during word production using speech motor and premotor cortical activity," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2015, pp. 5339–5342.

[24] S. Lesaja, C. Herff, G. D. Johnson, J. J. Shih, T. Schultz, and D. J. Krusienski, "Decoding lip movements during continuous speech using electrocorticography," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, March 2019, pp. 522–525.

[25] J. Chartier, G. K. Anumanchipalli, K. Johnson, and E. F. Chang, "Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex," *Neuron*, vol. 98, no. 5, pp. 1042 – 1054.e4, 2018.

[26] W. Hardcastle, N. Hewlett, and K. Munhall, "Coarticulation: Theory, data, and techniques," *The Journal of the Acoustical Society of America*, vol. 109, pp. 19–19, 01 2001.

[27] N. Crone, L. Hao, J. Hart, D. Boatman, R. Lesser, R. Irizarry, and B. Gordon, "Electrocorticographic gamma activity during word production in spoken and sign language," *Neurology*, vol. 57, no. 11, pp. 2045–2053, 2001.

[28] J. Wang, J. R. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[29] A. L. Crowell, E. S. Ryapolova-Webb, J. L. Ostrem, N. B. Galifianakis, S. Shimamoto, D. A. Lim, and P. A. Starr, "Oscillations in sensorimotor cortex in movement disorders: an electrocorticography study," *Brain*, vol. 135, no. 2, pp. 615–630, 01 2012. [Online]. Available: https://doi.org/10.1093/brain/awr332

[30] E. Boto, N. Holmes, J. Leggett, G. Roberts, V. Shah, S. S. Meyer, L. D. Muñoz, K. J. Mullinger, T. M. Tierney, S. Bestmann, G. R. Barnes, R. Bowtell, and M. J. Brookes, "Moving magnetoencephalography towards real-world applications with a wearable system," *Nature*, vol. 555, no. 7698, pp. 657–661, mar 2018.

[31] 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.

[32] J. Gross, S. Baillet, G. R. Barnes, R. N. Henson, A. Hillebrand, O. Jensen, K. Jerbi, V. Litvak, B. Maess, R. Oostenveld, L. Parkkonen, J. R. Taylor, V. van Wassenhove, M. Wibral, and J.-M. Schoffelen, "Good practice for conducting and reporting MEG research," *NeuroImage*, vol. 65, pp. 349 – 363, 2013.

[33] F. Siebenhhner, S. A. Weiss, R. Coppola, D. R. Weinberger, and D. S. Bassett, "Intra- and inter-frequency brain network structure in health and schizophrenia," *PLOS ONE*, vol. 8, no. 8, pp. 1–13, 08 2013.

[34] A. S. Ghuman, J. R. McDaniel, and A. Martin, "A wavelet-based method for measuring the oscillatory dynamics of resting-state functional connectivity in MEG," *NeuroImage*, vol. 56, no. 1, pp. 69 – 77, 2011.

[35] D. Dash, P. Ferrari, S. Malik, A. Montillo, J. A. Maldjian, and J. Wang, "Determining the optimal number of MEG trials: A machine learning and speech decoding perspective," in *Brain Informatics*. Cham: Springer International Publishing, 2018, pp. 163–172.

[36] D. Dash, P. Ferrari, and J. Wang, "Spatial and Spectral Fingerprint in the Brain: Speaker Identification from Single Trial MEG Signals," in *Proc. Interspeech 2019*, 2019, pp. 1203–1207. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3105

[37] N. Michielli, U. R. Acharya, and F. Molinari, "Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals," *Computers in Biology and Medicine*, vol. 106, pp. 71 – 81, 2019.

[38] P. R. Davidson, R. D. Jones, and M. T. R. Peiris, "Detecting behavioral microsleeps using EEG and LSTM recurrent neural networks," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Jan 2005, pp. 5754–5757.

[39] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *Journal of Neural Engineering*, vol. 16, no. 3, p. 031001, apr 2019.

[40] D. Dash, P. Ferrari, S. Malik, and J. Wang, "Automatic speech activity recognition from MEG signals using seq2seq learning," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, March 2019, pp. 340–343.

[41] D. Kostas, E. Pang, and F. Rudzicz, "Machine learning for MEG during speech tasks," *Scientific Reports*, vol. 9, p. 1609, 02 2019.

[42] D. Dash, P. Ferrari, S. Dutta, and J. Wang, "NeuroVAD: Real-time voice activity detection from non-invasive neuromagnetic signals," *Sensors*, vol. 20, no. 8, p. 2248, 2020.

[43] N. Alves, C. Jobst, F. Hotze, P. Ferrari, M. Lalancette, T. Chau, P. van Lieshout, and D. Cheyne, "An MEG-compatible electromagnetic-tracking system for monitoring orofacial kinematics," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1709–1717, Aug 2016.

[44] B. J. Perry, R. Martino, Y. Yunusova, E. K. Plowman, and J. R. Green, "Lingual and jaw kinematic abnormalities precede speech and swallowing impairments in ALS," *Dysphagia*, vol. 33, no. 6, pp. 840–847, Dec 2018.

[45] S. Shellikeri, Y. Yunusova, D. Thomas, J. R. Green, and L. Zinman, "Compensatory articulation in amyotrophic lateral sclerosis: Tongue and jaw in speech," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 060061, 2013.

[46] E. M. Wilson, J. R. Green, and G. Weismer, "A kinematic description of the temporal characteristics of jaw motion for early chewing: Preliminary findings," *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 2, pp. 626–638, 2012.

[47] Y. Yunusova, J. R. Green, M. J. Lindstrom, G. L. Pattee, and L. Zinman, "Speech in ALS: Longitudinal changes in lips and jaw movements and vowel acoustics." *Journal of medical speech-language pathology*, vol. 21 1, pp. 1–13, 2013.

[48] T. Horikawa, S. C. Aoki, M. Tsukamoto, and Y. Kamitani, "Characterization of deep neural network features by decodability from human brain activity," *Scientific data*, vol. 6, p. 190012, 2019.

[49] 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, p. 575, 2013.