Interspeech 2018
2-6 September 2018, Hyderabad

Revealing Spatiotemporal Brain Dynamics of Speech Production Based on EEG and Eye Movement

Bin Zhao[1], *Jinfeng Huang*[2], *Gaoyan Zhang*[1], *Jianwu Dang*[1, 2], *MinboChen*[1], *YingjianFu*[1], *Longbiao Wang*[1]

[1]Tianjin key Laboratory of Cognitive Computing and Application, Tianjin University, China
[2]Japan Advanced Institute of Science and Technology, Japan

zhanggaoyan@tju.edu.cn, jdang@jaist.ac.jp

Abstract

To understand the neural circuitry associated with speech production in oral reading, it is essential to describe the whole-range spatiotemporal brain dynamics in the processes including visual word recognition, orthography-phonology mapping, semantic accessing, speech planning, articulation, self-monitoring, etc. This has turned out to be extremely difficult because of demanding resolution in both spatial and temporal domains and advanced algorithms to eliminate severe contamination by articulatory movements. To tackle this hard target, we recruited 16 subjects in a sentence reading task and measured multimodal signals of electroencephalography (EEG), eye movement, and speech simultaneously. The onset/offset of gazing and utterance were used for segmenting brain activation stages. Cortical modeling of causal interactions among anatomical regions was conducted on EEG signals through (i) independent component analysis (ICA) to identify cortical regions of interest (ROIs); (ii) multivariate autoregressive (MVAR) modeling of representative cortical activity from each ROI; and (iii) quantification of the dynamic causal interactions among ROIs using the Short-time direct Directed Transfer function (SdDTF). The resulting brain dynamic model reveals a widely connected bilateral organization with left-lateralized semantic, orthographic and phonological sub-networks, right-lateralized prosody and motor sequencing sub-networks, and bi-lateralized auditory and multisensory integration sub-networks that cooperate along interlaced and paralleled temporal stages for speech processing.

Index Terms: dynamic neural network, speech production, speech planning, EEG source information flow, eye movement

1. Introduction

Uncovering how the brain perceives, plans, executes, and monitors continuous speech in oral reading has been a long-anticipated, while extremely challenging, goal [1-3]. A classic theory of speech production [4, 5] characterized this process as occurring in stages: after the initial text presentation, word generation proceeds through lexical retrieval, phonological/phonetic encoding, motor planning, articulation, and certain phases of output control of self-produced speech. Normally the fast and accurate process from visual word presentation to utterance could be easily initiated within 600 ms [6]. To capture the underlying brain dynamics in such a short sub-second timescale, both delicate equipment and effective algorithms are needed to transiently record and analyze the spatial and temporal dynamics pertaining to speech. So far, a major impediment to progress is technical constraints. Previous studies employing functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) focused on cerebral activation patterns and their regional functionality [7], yet missing temporal details as to how the regions are involved as the utterance progresses. Electroencephalography (EEG) and magneto-encephalography (MEG), in contrast, are well suited to describing millisecond dynamics [5]. However, EEG/MEG signals could easily be buried in electromagnetic artifacts from muscular actions in articulation, thus interfering with the analysis [8]. Electrocorticography (ECoG) is free from scalp noises, whereas, probably due to the complex and costly procedures, only the frontal-motor areas and superior temporal gyrus in the left hemisphere were investigated in an ECoG study on this topic [1]. However, it has been increasingly recognized that both left and right hemispheres contribute to language and speech functions [9-13]. In particular, Chinese prosody was suggested to be bilaterally processed with a relative right bias [14-16]. Such a whole-brain speech processing network is yet to be constructed. In addition, to circumvent the complexity of multisensory interaction and integration involved in continuous oral reading, most existing studies used words as stimuli, leaving the topic of sentence production rarely pursued.

Fortunately, recent advances in blind source separation, such as the independent component analysis (ICA), paved a new way of component decomposition that is capable of excluding muscle activities and eye blinks from cognitive components [17-19]. Besides, the latest advent of the Granger causal analysis [20] and multivariate autoregressive (MVAR) modeling [21] along with the source information toolbox (SIFT) [22] provided a novel framework to estimate and visualize the information flow within distributed brain networks based on time-frequency information [20]. In this study, we launched to combine EEG, eye-tracking and speech recording equipment to acquire multimodal physiological and behavioral data. By applying above-mentioned algorithms, it is promising to unveil the pattern of cortical networks during sentence reading and producing from a comprehensive spatiotemporal perspective.

2. Experiments

2.1. Stimuli and Participants

The visual stimuli consisted of 180 sentences with similar structures. Each of them is composed of 8 two-character Chinese words (16 characters/syllables per sentence).

Sixteen Mandarin speakers (age: 22.3 years, SD = 2.1) participated in this study. All the subjects reported normal or corrected-to-normal vision, right-handed [23], with normal hearing and speaking abilities. The ethical approval for this experiment was obtained from the Local Research Ethics Committee.

10.21437/Interspeech.2018-1908

2.2. Equipment and Data Acquisition

The experiment was conducted in an electromagnetically shielded room. Participants sat 60-65 cm away from a monitor screen (1360 x 768 pixels) and placed their forehead against a holder to prevent movements. Eye movement was recorded at 1000 Hz via a monocular pupil tracking system (Eyelink 1000, SR Research Ltd., Canada). Speech was recorded using a microphone (SONY ECM MS957) at 44100 Hz. Meanwhile, scalp-EEG signals were recorded with a 128-channel Quik-Cap (Neuroscan, USA) placed in accordance with the extended 10-5 system [24], see Figure 1. The sampling rate was 1000 Hz, and the channel impedance was maintained below 5k Ω throughout the acquisition.

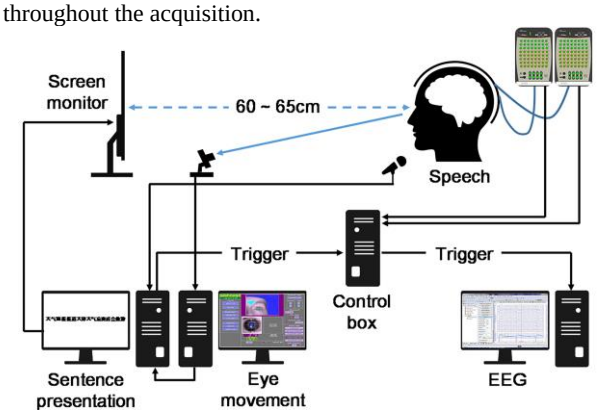

Figure 1: *Schematic diagram of multimodal acquisition for EEG, eye movement and speech data.*

2.3. Paradigm

For each trial, once the sentence was presented on the screen along a horizontal line, e.g. '天气预报报道天津天气凉爽适合旅游 (The weather forecast reports that the weather is cool in Tianjin and suitable for tourism)', the subject was asked to read and utter each word in a natural speed. A three-point (horizontally distributed) calibration [25] was adopted when the eye-tracking failed or shifted (Gaze accuracy deviation < 0.50º). After a 2000-ms resting period, a fixation cross appeared in the screen center for 1000 ms, followed by the presentation of a randomly selected sentence. When the subject's gazing point fell in either one of the 16 character fields, a trigger with the corresponding number would be marked on EEG signals. The trial ends with an ESC key press and a 1000 ms inter-trial-interval (ITI). All the subjects performed three experimental blocks with 60 sentences/trials for each block. The procedure varied from 52 to 88 minutes individually, with EEG, eye movement and speech data recorded in synchrony during the whole-range.

3. Methods and Results

3.1. Eye-tracking and speech segmentation

For an offline analysis, the locations and durations of eye movement were analyzed in MATLAB (MathWorks). Consistent with previous research [26], the subjects were found to process the characters in unit of two, forming into 8 two-character word processing units. The speech was segmented and aligned accordingly using the SPPAS software with human inspection [27]. Table 1 listed the averaged time latency of the eye and speech onset and offset from stimulus presentation for each of the 8 words in all the sentences. To separately analyze each cognitive stage, here we define the visual word processing period as the interval between each gazing onset and offset, articulation period as between the speech onset and offset, and speech planning period as from gazing onset till speech onset. Particularly, the speech planning period of the first word (lasting for 638 ms) was exempted from pre-word influence and was regarded as a reference for the following processes.

Table 1: *Averaged time latency (ms) of the gazing/speech onset /offset for 8 words.*

Word number	Gazing onset	Gazing offset	Speech onset	Speech offset
1	361	796	999	1409
2	796	1183	1409	1775
3	1183	1644	1775	2188
4	1644	2087	2188	2625
5	2087	2527	2625	3055
6	2527	3002	3055	3486
7	3002	3502	3486	3917
8	3502	3752	3917	4372

Figure 2 plots the superimposed trajectory of eye movement (blue solid line, with blue horizontal dotted lines as word boundaries) and speech spectrogram (grey background, with red vertical dot lines as speech boundary) of a given sentence. Generally, utterance for the 8-word sentences could be finished within 5s. The ordinate shows each word field for the reference of eye movement positions. As seen in the plot, before the subject pronounced the first word "天气", the gazing point has already reached the second word "预报". Similarly, in the following cases, uttering the former word was often overlapped with viewing the latter ones. In addition, immediately after utterance, auditory feedback could possibly join the reading and speaking time courses, implying a parallel and interlaced behavioral and cognitive integration in the ongoing oral reading process.

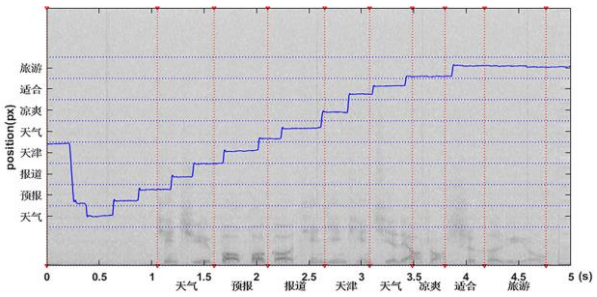

Figure 2: *Eye movement trajectory and speech segmentation.*

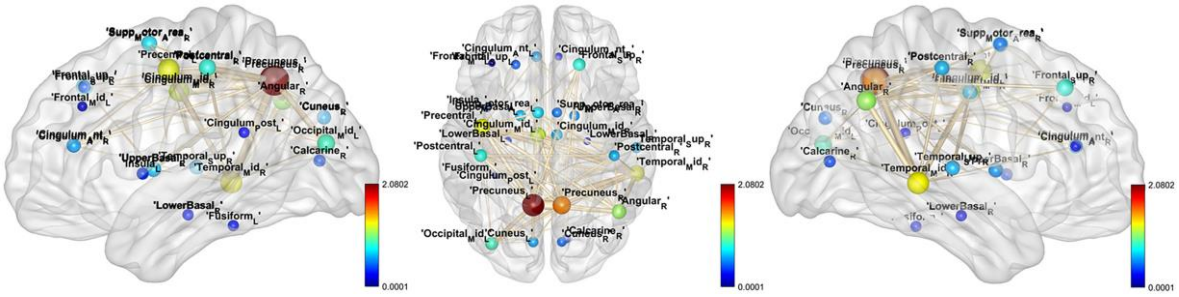

Figure 3: *Regions of interest (ROIs) and interconnections.*

3.2. EEG analysis

Preprocessing of EEG data was performed using the EEGLAB toolbox [28]. Individually, EEG signals were filtered at a bandwidth of 1-45 Hz and down-sampled to 250 Hz. Bad channels with over 10% of abnormal fluctuations were removed before re-referencing the data to average. Then the continuous data were segmented into 180 epochs ranging from-1000 ms to 5000 ms around each sentence presentation onset (0 ms), with the 1000ms pre-onset period as a baseline. We applied the Informax ICA algorithm [29] to transform the scalp-EEG data from a channel basis to a component basis, and separated out those maximally independent cortical sources from biological artifacts (from the eyes, muscles, and heart) and noise components [9]. An equivalent current dipole (ECD) model of each brain component was computed using boundary element model (BEM) to localize dipoles on the cortex [10]. Based on the dipole features, those physiologically plausible dipoles were selected and clustered across subjects to define the regions of interest (ROIs) [30]. We then applied routines from SIFT [22] to model spatio-temporal multivariate causal interactions between these ROI time-series. A linear vector (multivariate) autoregressive (VAR) model of order 10 was then fit to the multi-trial ensemble with 500 ms sliding window and a step size of 30 ms, using the Vieira-Morf lattice algorithm. Following the model fitting and tests of stability and residual whiteness, the Short-time Direct Directed Transfer Function (SdDTF) was estimated from the VAR coefficients to quantify time-varying connectivity. Figure 3 shows the nodes of ROIs and their interconnections for the whole range across all the subjects. The larger the node, the more active the region was and the tighter the connection was with other nodes. We can find that both hemispheres participated in speech production, where the hub regions were located in the left visual cortex (Occ), left posterior middle and inferior temporal gyrus (lpMTG/lpITG), inferior frontal gyrus (IFG), bilateral parietal (PL), prefrontal lobe (PFL), anterior insula (AI), anterior cingulum (ACC), right angular (rAG) and supramarginal syrus (rSMA). Table 2 additionally listed the anatomical locations for the plotted ROIs, followed by the most consistent functions that have been assigned to them according to previous research [7]. The results also suggest functional lateralization with left-biased lexical semantic, orthography and phonology representation, right-biased prosody, timing and motor control, along with bilateral integration of multisensory information.

For a whole-range observation, Figure 4 illustrates the activated neural networks of one subject along a 5-s time scale. At the onset of sentence presentation (0 ms), the brain was basically at a resting state, with activation in Occ, PL, AI and ACC for general visual response and attention allocation. At 361 ms when the gazing point fell into the first word, activity in the Occ outflowed in two directions: (i) to the left posterior occipital-temporal conjunction (lpOT, orthography-phonology mapping); (ii) to the lpMTG/lpITG (semantic processing).

Table 2: *Anatomical locations and functions related to speech processing (l-left; r-right; a-anterior; p-posterior; v-ventral; d-dorsal).*

Abbr.	Location	Functions
Occ	Occipital	visual word processing
OT	Occipital-Temporal	visual feature extraction (lp), orthography-phonology link (l)
ATC	Anterior temporal	semantic retrieval,
STG/ STS	Superior temporal gyrus/sulcus	auditory, acoustic (l&r), prosody, tone (rp)
MTG	Middle temporal	semantics, orthography-semantic-phonology (lp)
ITG	Inferior temporal	semantics (lp)
SMG	Supramarginal gyrus	orthography-phonology link, timing of speech, motor command, precise phonological decision (r)
AG	Angular gyrus	visual word forms (l), crossmodal integration spatial attention (r)
PFL	Prefrontal lobe	word retrieval (l) difficult word selection (r)
SFG	Superior frontal gyrus/sulcus	Syntax processing (l), Semantic retrieval (d)
MFG	Middle frontal	word semantic retrieval (l)
IFG	Inferior frontal	orthography-phonology mapping (lv), early syntactic processing (r)
PM	Premotor	articulatory/syllable coding, motor sequencing (l&r)
MC	Primary motor	speech production
PL	Parietal lobe	multisensory integration, visual spatial processing
AI	Anterior insula	articulatory planning (a) audio-visual integration, response selection
ACC	Anterior cingulated cortex	attention alloction, performance monitoring

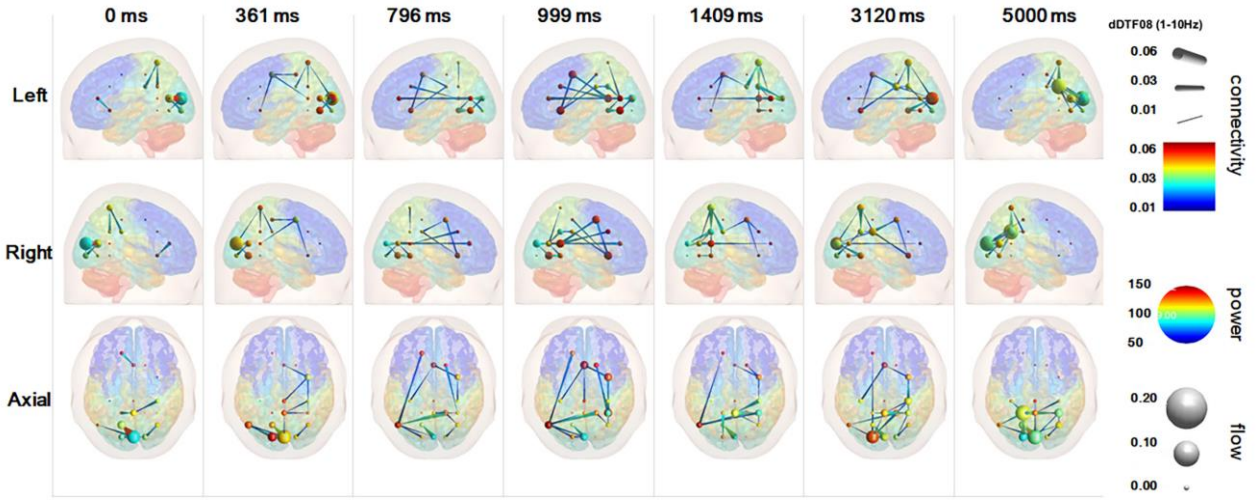

Figure 4: *Spatiotemporal brain dynamic networks of one subject during sentence recitation (perspective, left-right difference could be found from the Axial view).*

In addition, rPFL (difficult word selection) came into connection with AI (response selection) and rAG (visual word forms). PL was associated with Occ, rAG and sensorimotor gyrus, constructing a multisensory integration sub-network.

The gazing offset of the first word lasted till 796 ms, during which, responses in the Occ gradually decreases, while a lexical network, centered around the lpMTG/lpIMG, was getting to be connected with PFL (word retrieval), middle frontal gyrus (MFG, word retrieval) and right posterior superior temporal gyrus (rpSTG, Chinese tone/prosody modulation). In the meanwhile, a speech planning network connecting IFG, premotor (PM) and primary motor cortex (MC) was observed.

By the moment of articulation at 999 ms, activity strength peaks in IFG, PM, and MC as speech planning and execution networks, connecting with the lexical network that involved lpMTG/lpIMG, rpSTG, OT, rAG (cross-modal integration) and rSMA (timing of speech, motor command, and precise phonological decision).

When reaching the offset of articulation for the first word at 1409 ms, the firing within the speech planning and execution networks and lexical semantic, orthography, and prosody networks slowly died out, leaving the PL interacting with the Occ and SMA for multisensory integration.

In the subsequent processes, activity networks were widely activated for primary visual processing in Occ, orthography-phonology mapping in lpOT AG, SMG, IFG, word retrieval and semantic accessing in ACC, MFG, lpMTG/lpITG, prosody processing in rSTG, speech planning in IFG, PM, AI, articulation in MC, and selective auditory feedback control in bilateral STG (as reflected at 3120 ms). Activation after 5000 ms was only remained in Occ for basic visual processing and AG for general spatial attention.

4. Discussion

This study highlighted a speech production neuromechanism of bilateral cooperation, where lexical semantic, orthographic and phonological processing was exclusively left-lateralized; prosody and motor sequencing were right-preferred; and auditory response, spatial attention and multisensory integration were bi-hemispheric. The results are consistent with a growing body of emerging research on bilateral cooperation for language processing [2, 7, 9-12, 14, 16, 31].

Compared with the conventional static function localization, the current spatiotemporal dynamic model unfolded the temporal procedure for speech processing that initiated in the Occ for visual feature extraction, mediated in the pOT AG, SMG, and IFG for orthographic-phonetic mapping, interpreted within the pMTG/pITG semantic corpus, syllabified and phonetically encoded in the IFG, PM, AI, and executed in the MC for articulation. It is worth noting that in our illustration of the single subject (Figure 4), auditory response was weakly observed. This might be explained by individual differences if the subject suppresses auditory feedback control to make concession for other cognitive processes instead of conscious error correction sometimes.

5. Conclusions

By introducing a novel combination of a multimodal (EEG, eye movement and speech) data acquisition scheme and advanced algorithms including blind source separation and multivariate autoregressive based dynamic Granger causal modeling methods, the present study examined how perception, cognition and production emerge through bilateral cooperation of functionally distinct brain networks in an oral reading task. The observed spatiotemporal brain dynamics reflects that in the processes of visual word recognition, orthography-phoneme mapping, semantic accessing, speech planning, articulation, and self-monitoring, current flow could go along multiple streams in a parallel fashion, connecting distributed brain areas into functionally specific sub-networks and serve the fully operational system altogether. This result is supposed to extend our understanding of the speech neurological mechanisms from a spatiotemporal and dynamic casual view.

6. Acknowledgement

This study is supported by JSPS KAKENHI Grant (16K00297), the National Natural Science Foundation of China (No. 61503278, No. 61771333 and No. U1736219), and the Peiyang Scholar Program of Tianjin University (No. 2018XRG-0037).

7. References

1. Brumberg, J.S., et al., *Spatio-Temporal Progression of Cortical Activity Related to Continuous Overt and Covert Speech Production in a Reading Task.* Plos One, 2016. **11**(11): p. e0166872.
2. Hickok, G. and D. Poeppel, *The cortical organization of speech processing.* Nat. Reviews. Neurosci., 2007. **8**: p. 393-402.
3. Dronkers, N.F., M.V. Ivanova, and J.V. Baldo, *What Do Language Disorders Reveal about Brain-Language Relationships? From Classic Models to Network Approaches.* Journal of the International Neuropsychological Society Jins, 2017. **23**(9-10): p. 741.
4. Levelt, W.J.M. *A theory of lexical access in speech production*. in *Conference on Computational Linguistics*. 1996.
5. Zhao, B., J. Dang, and G. Zhang, *EEG Evidence for a Three-Phase Recurrent Process during Spoken Word Processing*, in *10th ISCSLP*. 2016: Tianjin, China.
6. Indefrey, P., *The Spatial and Temporal Signatures of Word Production Components: A Critical Update.* Front Psychol, 2011. **2**(255): p. 255.
7. Price, C.J., *A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading.* Neuroimage, 2012. **62**(2): p. 816.
8. Wohlert, A.B., *Event-related brain potentials preceding speech and nonspeech oral movements of varying complexity.* Journal of Speech & Hearing Research, 1993. **36**(5): p. 897.
9. Hickok, G., et al., *Bilateral capacity for speech sound processing in auditory comprehension: evidence from Wada procedures.* Brain & Language, 2008. **107**(3): p. 179-84.
10. Hartwigsen, G., et al., *Phonological decisions require both the left and right supramarginal gyri.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(38): p. 16494-9.
11. Vigneau, M., et al., *What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis.* Neuroimage, 2011. **54**(1): p. 577-593.
12. Alexandrou, A.M., et al., *The right hemisphere is highlighted in connected natural speech production and perception.* Neuroimage, 2017. **152**: p. 628-638.
13. Zhao, B., J. Dang, and G. Zhang, *EEG Source Reconstruction Evidence for the Noun-Verb Neural Dissociation along Semantic Dimensions.* Neuroscience, 2017. **359**.
14. Jia, S., et al., *Right hemisphere advantage in processing Cantonese level and contour tones: Evidence from dichotic listening.* Neuroscience Letters, 2013. **556**: p. 135–139.
15. Sammler, D., et al., *Dorsal and Ventral Pathways for Prosody.* Current Biology Cb, 2015. **25**(23): p. 3079.
16. Si, X., W. Zhou, and B. Hong, *Cooperative cortical network for categorical processing of Chinese lexical tone.* Proc Natl Acad Sci U S A, 2017. **114**(46): p. 12303-12308.
17. Jung, T.P., et al., *Independent Component Analysis of Electroencephalographic and Event-Related Potential Data.* Advances in Neural Information Processing Systems, 1996. **8**(8): p. 1548-1551 vol.2.
18. De, V.M., et al., *Removal of muscle artifacts from EEG recordings of spoken language production.* Neuroinformatics, 2010. **8**(2): p. 135-150.
19. Zhao, B., J. Dang, and G. Zhang. *A Neuro-Experimental Evidence for the Motor Theory of Speech Perception*. in *INTERSPEECH*. 2017.
20. Seth, A.K., A.B. Barrett, and L. Barnett, *Granger causality analysis in neuroscience and neuroimaging.* Journal of Neuroscience the Official Journal of the Society for Neuroscience, 2015. **35**(8): p. 3293-7.
21. O'Neill, G.C., et al., *Dynamics of large-scale electrophysiological networks: A technical review.* Neuroimage, 2017.
22. Mullen, T., et al., *An Electrophysiological Information Flow Toolbox for EEGLAB.* Biological Cybernetics, 2010.
23. Oldfield, R.C., *The assessment and analysis of handedness: the Edinburgh inventory.* Neuropsychologia, 1971. **9**(1): p. 97-113.
24. Oostenveld, R. and P. Praamstra, *The five percent electrode system for high-resolution EEG and ERP measurements.* Clin Neurophysiol, 2001. **112**(4): p. 713-9.
25. Schuster, S., et al., *Words in Context: The Effects of Length, Frequency, and Predictability on Brain Responses During Natural Reading.* Cereb Cortex, 2016. **26**(10): p. 3889-3904.
26. Huang, J., D. Zhou, and J. Dang. *Estimation of Speech-planning mechanism based on eye movement*. in *International Seminar on Speech Production*. 2017. Tianjin.
27. Bigi, B. *SPPAS: a tool for the phonetic segmentations of Speech*. in *Language Resources and Evaluation Conference*. 2012.
28. Delorme, A. and S. Makeig, *EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis.* Journal of Neuroscience Methods, 2004. **134**(1): p. 9.
29. Bell, A.J. and T.J. Sejnowski, *An information-maximization approach to blind separation and blind deconvolution.* Neural Comput, 1995. **7**(6): p. 1129-59.
30. Delorme, A., et al., *EEGLAB, SIFT, NFT, BCILAB, and ERICA: New Tools for Advanced EEG Processing.* Computational Intelligence & Neuroscience, 2011. **2011**(1687-5265): p. 10.
31. Cohen, L., et al., *The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients.* Brain A Journal of Neurology, 2000. **123 (Pt 2)**(2): p. 291-307.