INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Neuro-Experimental Evidence for the Motor Theory of Speech Perception

Bin Zhao[1], *Jianwu Dang*[1, 2], *Gaoyan Zhang*[1]

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

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

Abstract

The somatotopic activation in the sensorimotor cortex during speech comprehension has been redundantly documented and largely explained by the notion of embodied semantics, which suggests that processing auditory words referring to body movements recruits the same somatotopic regions for that action execution. For this issue, the motor theory of speech perception provided another explanation, suggesting that the perception of speech sounds produced by a specific articulator movement may recruit the motor representation of that articulator in the precentral gyrus. To examine the latter theory, we used a set of Chinese synonyms with different articulatory features, involving lip gestures (LipR) or not (LipN), and recorded the electroencephalographic (EEG) signals while subjects passively listened to them. It was found that at about 200 ms post-onset, the event-related potential of LipR and LipN showed a significant polarity reversal near the precentral lip motor areas. EEG source reconstruction results also showed more obvious somatotopic activation in the lip region for the LipR than the LipN. Our results provide a positive support for the effect of articulatory simulation on speech comprehension and basically agree with the motor theory of speech perception.

Index Terms: motor theory of speech perception, articulatory gestures, somatotopic activation, synonym, EEG source reconstruction

1. Introduction

Neurobiological and psycholinguistic studies have long postulated that knowledge about articulatory features of individual phonemes has an important role in speech perception and comprehension [1-3]. One of the most intriguing and highly cited theories is the motor theory of speech perception [4-6], which claims that the listener perceives speech by simulating "intended articulatory gestures" of the speaker. This perception-production circuit was demonstrated by a neurobiological evidence showing that passively listening to phonemes and syllables tends to activate the motor and premotor cortex [7]. Interestingly, these activations were somatotopically organized according to the articulatory effector that recruited in the production of these phonemes [8]. As reported in an fMRI study by Pulvermüller [2], distinct motor regions in the left precentral gyrus governing articulatory movements of the lips were also differentially activated when subjects listened to the lip-related phonemes. Researchers also used repetitive transcranial magnetic stimulation (TMS) to temporarily disrupt the lip representation area in the left primary motor cortex, and found this TMS-induced disruption impaired the categorical perception of phonemes and syllables that involved lip movement in their articulation [9-12]. However, regarding the higher level semantic process of spoken words, it still remains to be determined: questions here are whether the perception-articulation link keeps its contribution to speech perception and comprehension and to what extent the articulatory information of phonemes and syllables interacts with semantics during word recognition and understanding [12]. Lack of investigation on these topics is probably due to the semantic interference. In high-level word or sentence comprehension, the somatotopic motor activation by a verb or the predicate part in a sentence tends to be mixed up and attributed to the motor-related semantic retrieval in the sensorimotor regions, where the execution of the motor acts is also represented somatotopically [13]. This notion is known as the theory of embodied semantics [14].

To disentangle phonological motor effects from semantic ones, this study examined whether the gesture information would be reflected in the perception and comprehension of a spoken word when maximally minimizing the semantic influence. To this end, we employed sets of Chinese synonyms that have near-synonymous or identical meanings but recruit different articulatory gestures. Considering that verbs with action meaning tend to activate the frontal motor regions and cause confusions [15], all the synonyms used in this experiment were the concrete nouns only referring to static objects. For comparison with previous studies, we also distinguished the synonyms by the recruited shape of the lips. The test word groups consisted of synonyms, in each pair of which one required lip gestures (e.g., the syllable includes a labial consonant like /b/, /p/ or a vowel like /u/), and the other group require no lip gestures (e.g., the syllable includes a consonant like /k/, /g/ or a vowel like /i/). The control groups are also synonymous pairs, yet neither of the two items required lip gestures.

To find a temporal clue for the articulatory motor effect, we utilized high resolution electroencephalograph (EEG) to detect instantaneous brain responses to the time-varying acoustic signals and used event-related potential (ERP) analysis to compare group differences. An EEG source reconstruction technique was employed to trace back to the sensorimotor cortex to find out whether the lip motor regions respond differently to speech sounds with different lip gestures.

2. Materials and Methods

2.1. Materials

2.1.1. Subjects.

The subjects of this study were 22 (12 females and 10 males) native speakers of Mandarin with a mean age of 22.3 years (SD = 2.1). They were all right-handed [16], with normal hearing and normal or corrected-normal vision, and reported no diagnosed history of psychiatric disorders or neurological

http://dx.doi.org/10.21437/Interspeech.2017-1741

deficits. Ethical approval for this experiment was obtained from the Local Research Ethics Committee.

2.1.2. *Stimuli*

The auditory stimuli included four types of synonymous pairs (20 pairs or 40 items for each type) and 160 white noise segments (hissing sounds with the same frequency of the word stimuli). The synonyms were all two-character Chinese words, lasting for 800 ms (400 ms per character), and they were evaluated on their familiarity, concreteness, object-relatedness and emotional features by another 20 native Mandarin speakers, which is to make sure that each pair of the synonyms have no significant differences in the above possible influential factors ($ps > 0.05$). Table 1 gives out an illustration of the four types of synonyms. For type (1), the first characters of the two items in this pair are different, one of them require lip gestures (LipR) and the other one require no lip gestures (LipN), and the second characters of this pair are the same (_S). Type (2) is a negative example of type (1), in which the first characters of this pair are different, yet neither of them need lip gestures (LipN), and their second characters are also the same. For type (3), the first characters are the same (S_), while the second ones differ in their lip gesture requirement (the former) or not (the latter). Type (4) is a control group of type (3), in which the first characters are the same and neither of their second characters need lip gestures. The third column of Table 1 gives one example of the synonyms for each type in Chinese. As the periods of difference (PODs) for type (1) and (2) are on the first character (0-400 ms), and for type (3) and (4) are on the second character (400-800 ms), we analyzed the (1) vs (2) contrast and (3) vs (4) contrast in these two PODs respectively.

Table 1: *Illustration of the four types of synonyms.*

PODs	Types	Examples
0-400 ms	(1) LipR_S vs LipN_S	书本 vs 课本
	(2) LipN_S vs LipN_S	夕阳 vs 斜阳
400-800 ms	(3) S_LipR vs S_LipN	窗户 vs 窗口
	(4) S_LipN vs S_LipN	外衣 vs 外套

2.2. Methods

2.2.1. *Data acquisition*

The experiment was conducted in a soundproof, electromagnetically shielded laboratory. Each trial starts with a fixation cross in a computer screen center for around 400 ms, followed by a randomly selected 800-ms auditory stimulus from the 160 items of synonyms or 160 modulated white noise segments through headphones. A 1000-ms inter-trial-interval was set between two adjacent trials. Subjects were instructed to passively listen to them. The EEG signals were recorded from their scalps with a 128-channel Quik-Cap (Neuroscan, USA) that is placed in accordance with an extended 10-5 system [17]. The sampling rate was 1000 Hz, and the channel impedance was maintained below 5k Ω throughout the acquisition.

2.2.2. *ERP analysis*

The continuous EEG signals were first filtered at a bandwidth of 0.1-45 Hz and re-referenced to a common average reference. EEG artifacts induced by eye blinks, eye movements and muscle activities were detected and corrected using the independent component analysis (ICA). Then epochs during -200-800 ms around each stimulus onset were extracted, with the 200-ms pre-onset period as a baseline for correction. Invalid epochs with extreme values, abnormal trends/spectrum, and improbably /abnormally distributed data were rejected from the datasets. After that, valid epochs were averaged by stimulus conditions (totally 9 conditions: 4 synonym types × 2 items + 1 noise). Three electrodes in the precentral (FC3), central (C3) and postcentral (CP3) gyrus were selected on the basis of previous localization experiments for the motor-lip representation [9].

2.2.3. *EEG source reconstruction*

EEG source reconstruction is a procedure to locate the brain origins that generate the scalp-recorded EEG signals, which is employed in this study to recover the cortical activation patterns when the brain processes words involving lip gestures or not. As the signal conduction in the head of each subject varies in size and structure, we performed the boundary element method to generate realistic head models for each individual. These boundary element models were then co-registered with a standard MRI template for comparison. Considering the concurrency of multiple active sources and the ill-posed problem in almost all inverse algorithms, we performed the current density reconstruction (CDR) on the constraint of standardized Low Resolution Electromagnetic Tomography, which is capable of exhibiting cerebral dynamic sources on a 3D cortical map and providing the activation extent at each time point. In this experiment, a special attention was paid to the time periods when significant ERP differences ($p < 0.05$) were found between the synonymous pairs.

3. Results

3.1.1. *ERP results*

In Figure 1, ERP waves of the four synonym types (row1-4) and noise were compared on the electrodes of FC3, C3 and CP3 (column 1-3) in their corresponding PODs (0-400 ms for type (1) and (2); 400-800 ms for type (3) and (4)). In general, ERP wave of the noise showed less fluctuation than that in the word cases, especially during the 400-800 ms range. In the 0-400 ms period, noise stimuli elicited a negative peak shortly after 100 ms (N1 component) and a positive peak around 200 ms (P2 component) in all three electrode sites, as shown in the illustration of type (1) and (2). This could probably be explained by the electrocortical mapping from the near auditory-related cortical areas, namely the superior temporal gyrus, Broca's area, and Wernicke's area, where the acoustic-phonetic and phonological information is processed to discriminate noises from speech signals. After that, the process of noises was exempted from the higher-level analysis of semantic meaning in N400 and the later periods. In terms of the word conditions, ERP results of type (1) showed that compared to LipN_S, LipR_S elicited a more significant negative peak ($p < 0.05$) around 200 ms at FC3 and C3 electrode sites, as marked with black lines on the time axis. This negative peak was not significant at the PC3 electrode. For type (2), no significant ($P > 0.05$) ERP difference was found between the two LipN_S conditions during the whole POD. In the 400-800 ms range, a large negative peak was also found in the S_LipR condition of type (3) 200 ms after the onset of the second character (600 ms post-onset of the two-character word). This negative peak lasted

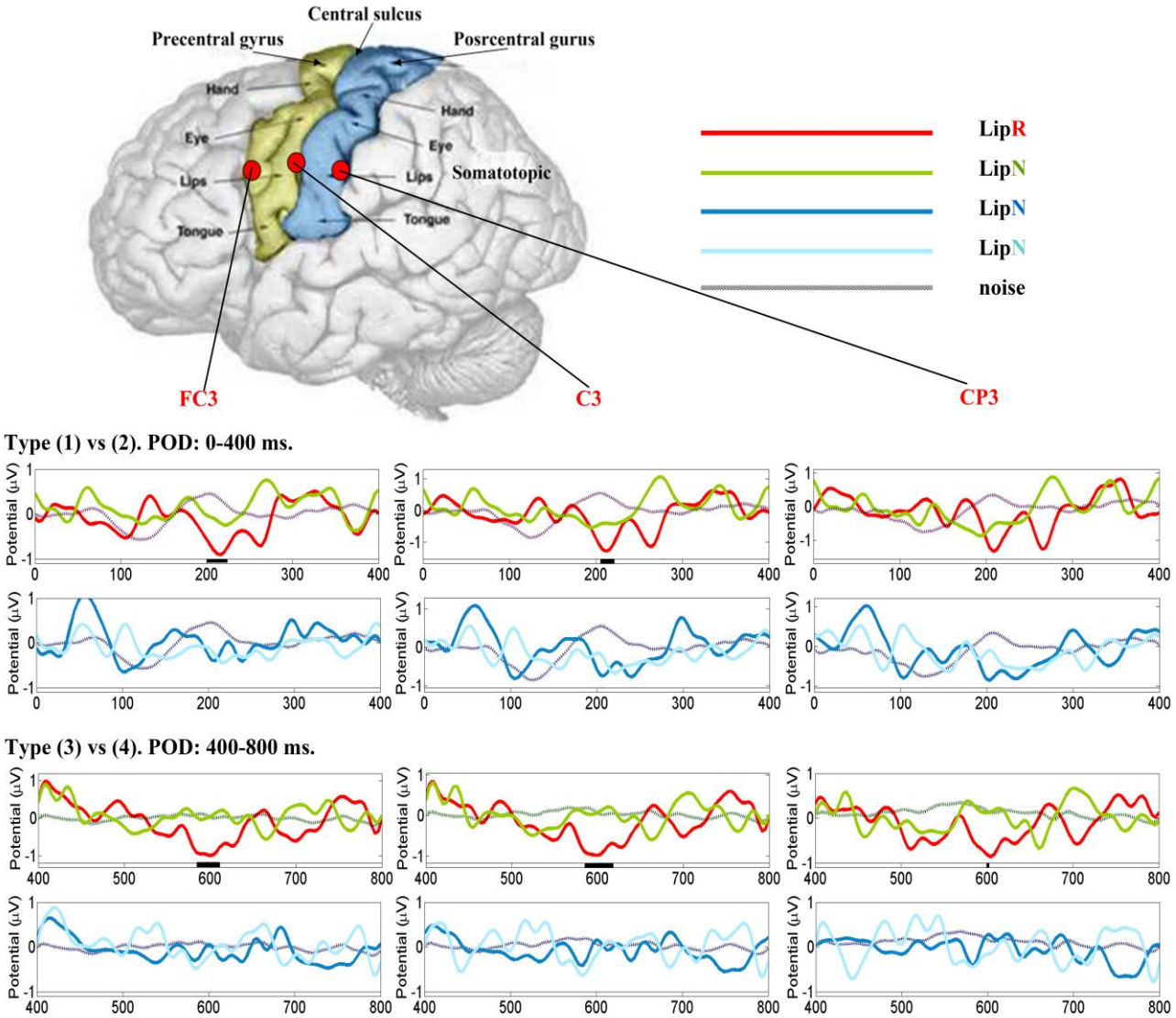

Figure 1: *ERP waveform comparison of the four types of synonyms and noise at FC3, C3 and CP3 electrode sites. The four rows of the ERP plots correspond to the four types of synonyms in order. The three columns of the ERP plots correspond to the three electrode sites at the precentral, central and postcentral lip-related regions.*

for the longest duration at FC3 and C3 electrodes, while it was quite transient at CP3 electrode. In type (4), the two S_LipN conditions again failed to show significant differences at 200 ms after onset of the test word. These results revealed that in passive speech perception tasks, words with lip-related features elicited stronger activations in the precentral lip motor regions than their synonyms lacking lip-related gestures with a 200-ms latency.

CDR results

The CDR results of type (1) and (3) were inspected during the whole range of their PODs for the comparison of the LipR-LipN contrasts. As shown in Figure 2, the intensity of the maximum activation was normalized as 100%, where the cortical regions with an activation intensity over 60% were highlighted referring to the color scale on the left. As expected, in type (1), at 200 ms after the onset of the first character, the inferior precentral and central gyrus, covering the lip-related motor and premotor areas were activated obviously by the LipR_S stimuli, which was absent in the lipN_S condition. Similarly, in type (3), when the second character was displayed for 200 ms (600 ms after onset), S_LipR also induced obvious stimulation near or over the lip somatotopic regions, which was also failed to be detected in the S_LipN condition. These findings are consistent with the ERP outcomes, providing neuro-experimental evidence for an involvement of the articulatory motor control in the lip-related regions in response to speech sounds involving lip-related articulatory information.

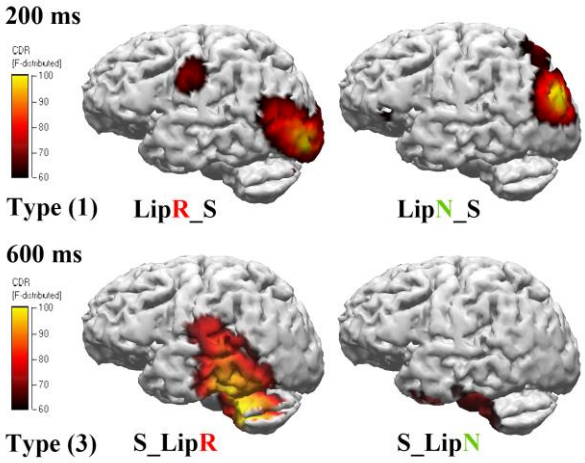

Figure 2: *CDR results on the cortical maps for type (1) and (3) at 200 ms and 600 ms post-stimulus onset, respectively. The intensity of the maximum activation was normalized as 100%, and the cortical regions with an activation intensity over 60% was highlighted referring to the color scale on the left.*

4. Discussion

In the current study, we localized and characterized the neural activity patterns in the lip-related sensorimotor system when the brain processes spoken words of synonyms involving different articulatory gestures of the lips. It was found that even at high level semantic process of spoken words, the articulatory information of phonemes and syllables still make a difference in the cognition and discrimination between synonyms. As the ERP and CDR results reflected, during speech perception, the articulatory features of the phonetic-distinctive speech sounds tend to recruit specific motor circuits in the precentral gyrus in a somatotopic fashion around 200 ms after the speech onset.

Similar results were also seen in a recent EEG and TMS study [3] reporting that the disruption of the lip-motor regions significantly suppressed the ERP response 166–188 ms after the onset of sound "ba" and 170–210 ms after the onset of sound "ga". These findings suggested a temporal clue for the articulatory motor effects on early auditory discrimination of phonological features, starting around 200 ms after speech onset, and also proved the spatial reference to the gestural information in a somatotopic fashion in speech perception tasks, which are compliant with the major thought of the motor theory of speech perception. Nevertheless, one critical point that needs to be clarified is that although this study provides support for the somatotopic activation in the articulatory motor regions, it should not be interpreted directly in the way that perceptual representations of phonemes and perceptual phonemic processes are localized in the precentral cortex, nor the speech perception process heavily depends on the articulatory simulation. As clinical evidence has pointed that damage in the premotor and motor regions does not necessarily cause speech perception problems. Instead, the motor recruitment in speech perception might be partly explained by a perception-production circuit, concerning with the correlation learning principle [18]. As explained, speaking results from motor activity of the articulators with auditory monitoring the self-produced sounds, which predicts synaptic strengthening and formation of a specific articulatory-auditory link. As a result, a correlated neuronal circuit connecting the auditory and the articulatory motor system was established. Therefore, whenever an auditory or an articulatory component is triggered, other nodes in this circuit tend to be evoked at the same time. In this sense, the articulatory motor involvement in speech perception is more likely to be attributed to the co-activation of perception and articulation circuits than a mandatory dependence.

Therefore, our results might be better interpreted in the sense that during speech perception, gesture information in the precentral motor and premotor regions tend to be evoked and coactivated by speech sounds in a somatotopic fashion, assisting the cognition of the phonemes and syllables, and the comprehension of the spoken words. In addition, considering the contribution of the semantic information and other dimensions of vocabulary formation during the process of lexical information [13], we boldly infer that spoken words could be perceived and cognized in multiple dimensions, including the phonetic features, articulatory gestures, semantic associations, and even some word-form transformations.

5. Conclusions

This study examines the motor contribution to speech perception at the word level, or more specifically, the possible role of the perception-articulation circuits in passive word listening tasks. It is found that during perception and comprehension of two synonyms, the one that recruits lip gestures for its articulation, comparing to the other that involves little lip movement, elicited a stronger neural response in the precentral lip motor cortex around 200 ms post-onset. This result suggests an automatic motor-somatotopic association when perceiving spoken words with specific articulatory features, even when the semantics are carefully controlled. Basically, our findings are in agreement with the motor theory of speech perception in the sense of the articulatory facilitation to comprehension. However, it's still arguable whether the motor circuits play a causal contribution to speech perception, which requires a step further investigation in future research.

6. Acknowledgements

This research is supported partially by the National Basic Research Program of China (No. 2013CB329301), the National Natural Science Foundation of China (No. 61233009 and 61503278), and is supported partially by JSPS KAKENHI Grant (16K00297).

7. References

[1] J. M. Correia, B. M. Jansma, and M. Bonte, "Decoding articulatory features from fmri responses in dorsal speech regions," *Journal of Neuroscience the Official Journal of the Society for Neuroscience*, vol. 45, no. 35, pp. 15015-15025, 2015.

[2] F. Pulvermüller, M. Huss, F. Kherif, P. M. F. Moscoso, O. Hauk, and Y. Shtyrov, "Motor cortex maps articulatory features of speech sounds," *Proc Natl Acad Sci U S A*, vol. 103, no. 20, pp. 7865-70, 2006.

[3] R. Möttönen, R. Dutton and K. E. Watkins, "Auditory-motor processing of speech sounds," *Cerebral Cortex*, vol. 23, no. 5, pp. 1190-7, 2013.

[4] A. M. Liberman, F. S. Cooper, D. P. Shankweiler and M. Studdert-Kennedy, "Perception of the speech code," *Psychol. Rev,* vol. 74, pp. 431–461, 1967.

[5] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition* vol. 21, pp. 1–36, 1985.

[6] B. Galantucci, C. A. Fowler and M. T. Turvey, “The motor theory of speech perception reviewed,” *Psychon. Bull. Rev,* vol. 13, pp. 361–377, 2006.
[7] S. M. Wilson, A. P. Saygin, M. I. Sereno and M. Iacoboni, “Listening to speech activates motor areas involved in speech production,” *Nat. Neurosci, v*ol. 7, pp. 701–702, 2004.
[8] A. D'Ausilio, F. Pulvermüller, P. Salmas, I. Bufalari, C. Begliomini and L. Fadiga, “The motor somatotopy of speech perception,” *Current Biology Cb,* vol. 19, no. 5, pp. 381-5, 2009.
[9] R. Möttönen, J. Rogers and K. E. Watkins, “Stimulating the lip motor cortex with transcranial magnetic stimulation,” *Journal of Visualized Experiments*, vol. e51665, no. 88, 2014.
[10] R. Möttönen and K. E. Watkins, “Motor representations of articulators contribute to categorical perception of speech sounds,” *The Journal of Neuroscience,* vol. 29, no. 31, pp. 9819–9825, 2009.
[11] I. G. Meister, S. M. Wilson, C. Deblieck, A. D Wu and M. Iacoboni, “The essential role of premotor cortex in speech perception,” *Current Biology Cb*, vol. 17, no. 19, pp. 1692, 2007.
[12] M. R. Schomers, E. Kirilina, A. Weigand, M. Bajbouj and F. Pulvermüller, “Causal influence of articulatory motor cortex on comprehending single spoken words: tms evidence,” *Cerebral Cortex,* vol. 25, no. 10, pp. 3894-3902, 2015.
[13] A. L. Arévalo, J. V. Baldo and N. F. Dronkers, “What do brain lesions tell us about theories of embodied semantics and the human mirror neuron system?” *Cortex*, vol. 48, no. 2, pp. 242, 2012.
[14] L. Aziz-Zadeh and A. Damasio, “Embodied semantics for actions: Findings from functional brain imaging,” *Journal of Physiology-Paris, vol.* 102, pp. 35-39, 2008.
[15] G. Vigliocco, D. P. Vinson, J. Druks, H. Barber and S. F. Cappa “Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies,” *Neuroscience & Biobehavioral Reviews, vol.* 35, pp. 407-426, 2011.
[16] R. C. Oldfield, “The assessment and analysis of handedness: The Edinburgh inventory,” *Neuropsychologia, vol.* 9, no. 1, pp. 97-113, 1971.
[17] R. Oostenveld and P. Praamstra, “The five percent electrode system for high-resolution EEG and ERP measurements,” *Clinical Neurophysiology* vol. 112, pp. 713-719, 2001.
[18] N. Caporale and Y. Dan, “Spike timing-dependent plasticity: a Hebbian learning rule,” *Annual Review of Neuroscience*, vol. 31, pp. 25–46, 2008.