Neural processing of degraded speech using speaker's mouth movement

Tomomi Mizuochi-Endo[1], *Michiru Makuuchi*[1]

[1]Section of Neuropsychology, National Rehabilitation Center for Persons with Disabilities

mizuochi-tomomi@rehab.go.jp

Abstract

Previous studies reported that visual speech cues enhance speech perception ability, but the critical contribution of the brain areas to successful AV integration for degraded speech is still unclear. To clarify this, we performed an fMRI study on word perception, using noise vocoded speech with clips showing a speaker's face.

We recruited 17 right-handed healthy adults, who were presented short video clips in which a Japanese male read aloud Japanese 3-mora nouns. The sounds were noise vocoded at 16 and 32 bands. The experimental conditions were designed as a 2x2 factorial design, crossing Modality (audio-only, A / audio-visual, AV) and Intelligibility (16 / 32 -bands). In both AV and A conditions, the sound and the clip were presented, but in the A conditions the speaker's mouth was blurred. During fMRI scanning, the participants were instructed to choose one word from a forced-choice probe with four alternatives after they listened/watched the sound and clip. Trial time courses in each model were estimated in the posterior STS, the lip motor area, and the inferior frontal gyrus in the left hemisphere.

Behavioral data revealed that successful AV integration improved participants' performance, in line with previous studies. Imaging data showed that the brain network associated with speech processing activated differently in time-course depending on both the modality and intelligibility of speech perception.

Index Terms: AV integration, speech perception, fMRI

1. Introduction

In speech perception, we use not only auditory information but also visual information. The speaker's face can help speech perception in noisy condition [1]. To reveal the contribution of the speaker's face in speech perception, researchers used degraded sounds such as noise vocoded speech [2]. Noise vocoding is a technique to synthesize noise and other sounds such as speech. Temporal envelopes of original speech sounds were extracted from frequency bands and were used to modulate noises of the same bandwidths. The advantage of this technique is that the intelligibility of sounds can be controlled easily by changing the number of frequency bands. A smaller number of bands makes the sounds more degraded.

Brain imaging studies have revealed that the comprehension of degraded speech is associated with activity in the left superior temporal sulcus (STS) [3, 4]. The STS activation increased for more intelligible sounds. In addition, the motor area (M1) of lip movements was activated more during the processing of degraded speech than that of natural speech [5, 6]. The inferior frontal gyrus (IFG) including Broca's area is involved in degraded speech processing [7]. However, some studies did not take the participants' performance into account even though their performances were low [8 – 11]. For example, the accuracy of a word recognition task using 6 band vocoded sound [9, 10] and the percentage of correct words in a sentence repetition task using 8 band vocoded sound [8] were less than 40 %. Although previous studies revealed brain areas serving the processing of degraded sounds, these studies are unable to compare brain activity between correct trials and incorrect trials because of their low task accuracies.

The posterior STS (pSTS) is known to play a key role in audiovisual (AV) integration [12 – 15]. Nath et al [11] showed that the functional connectivity between the pSTS and the sensory area was biased toward the more reliable modality; the functional connectivity between the pSTS and auditory cortex increased when the sound stimuli carried less noisy information than the visual stimuli; in addition, the functional connectivity between the visual area and the pSTS increased when visual stimuli conveyed better signals relative to noisy sound stimuli. These results suggest that the pSTS may integrate AV information in a bottom-up manner. The M1 and IFG are also involved in AV integration of speech sounds and the speaker's face [16]. These areas seem to play roles in motor imagery for subvocal rehearsal during speech processing [17, 18].

Skipper and colleagues pointed to a model of speech perception in which visual contextual information is used to test hypotheses about the identity of speech sounds [17, 19, 20]. Their hypothesis-and-test or analysis-by-synthesis model was based on the analysis-by-synthesis theory [21]. In this model, a hypothesis is specified in terms of the motor commands that might elicit the hypothesized movements for the target speech sound. Skipper and colleagues specified the pSTS, supramarginal gyrus, somatosensory cortices, motor area, and the pars opercularis as the cortical areas that support the mechanisms underlying this model. The hypothesis generated in the pSTS is transformed into the representation of the motor goal of that speech movement in the IFG. This motor goal is in turn mapped to the motor commands in the motor area. These motor commands yield a prediction of the auditory (from the motor area to the pSTS) consequences of these commands as an efferent copy. The resulting predictions constrain speech processing by biasing a particular interpretation or hypothesis in the pSTS. If the hypothesis and the prediction differ from each other, the cycle involving the pSTS, IFG and motor area repeats until the difference is suppressed. This network provides the mechanisms for the improvement of speech comprehension accompanied by mouth movement presentations. It is unclear, however, whether these areas contribute to the successful AV

integration or rather the *attempt* to integrate auditory and visual information regardless of outcome.

In the current study, we performed an fMRI study of word perception using noise vocoded speech [2] accompanied with video clips of a speaker's face to test the critical brain areas contributing to successful AV integration. Especially, we focused on the pSTS, M1 and IFG that may be involved in degraded speech processing and AV integration. We presented single words as stimuli, using lexical information to exclude top-down processes, and employed more intelligible speech sounds than those used in previous studies to compare brain activity between correct and incorrect trials. If the hypothesis-and-test network correctly captures the improvement in speech comprehension accompanied by mouth movements, the activity in the pSTS, M1 and IFG may differ depending on the predictability of input sounds. Conditions with highly intelligible sounds or clear audio-visual presentations would facilitate/strengthen comprehenders' prediction, and sound processing may complete in one cycle of the hypothesis-and-test network. In contrast, conditions with less intelligible sounds or audio-only presentations would require more cycles looping over the network for (successful) sound processing.

Thus, we infer that this difference in sound processing is reflected by the time course of brain activation and investigate the differences using trial time course analysis [22].

2. Methods

2.1. Participants

Twenty Japanese healthy right-handed adults (10 females, aged 23.1 ± 3.7) participated in this study. All had no history of neuropsychiatric disorders, and had normal or corrected-to-normal vision. Three participants were excluded from the analysis; two for showing no performance improvement by visual speech cues, and the other one for technical problems.

Seventeen participants (8 female, aged 23.3 ± 3.8) were included in the analyses. All participants were fully informed of the methods and the techniques of the noninvasive fMRI recordings, and they signed a written consent form to participate in the study. The study protocol was approved by the ethical committee of the National Rehabilitation Center for Persons with Disabilities in Japan.

2.2. Stimuli

The stimuli were 250 three-mora Japanese nouns of high familiarity (more than 6.25 in 7-point scale) selected from the NTT database [23]. We recorded speech sounds of the stimuli read aloud by a Japanese male actor using a Linear PCM Recorder PCM-D100 (Sony, Tokyo, Japan) along with a digital video using a digital HD video camera recorder HDR-CX700V (Sony, Tokyo, Japan). Speech sounds were noise vocoded at 16 and 32 bands using MATLAB (http://sethares.engr.wisc.edu/vocoders/chanvocoder.m), because we can vary sound intelligibility by changing the number of bands. These bands are decided by the results of a pilot behavioral study. When the participants listened to only noise vocoded speech and chose one from 4 alternatives, the accuracy was 40 % for 16 bands, and 74 % for 32 band. We replaced the sounds in the video clips used in the fMRI experiment with these vocoded sounds. Each clip showing a speaker's face was extracted from 1 sec before to 1 sec after the speech motion by VideoStudio 9 (Corel Corporation, Ontario, Canada). The length of each clip was about 2.5 sec. These clips were presented as an audio-visual condition (AV). On the other hand, an audio-only condition (A) involved video clips in which the speaker's mouth was blurred along with 500ms white noise to remove information from the speaker's mouth, created by Adobe Premiere Elements 12 (Adobe Systems Inc, CA, USA).

The experimental conditions were designed as a 2x2 factorial design, with factors INTEGRATION (A / AV), and INTELLIGIBILITY (16 / 32). In the audio-visual conditions (AV16 / AV32), the videos that combined a speaker's face and a vocoded sound were presented on a 32-inch 1920 x 1080 LCD monitor (NNL-LCD, NordicNeuroLab, Bergen, Norway). In the audio-only conditions (A16 / A32), vocoded sounds were presented with the mosaic videos. The participants listened to the vocoded sounds with the mosaic video in these conditions. Additionally, we added a visual-only condition (V), where the videos were presented with the sounds replaced with white noise, yet we excluded this condition from the analysis. Each condition contained 50 stimuli and the number of total trials was 250.

2.3. Procedure

The stimuli were given in a pseudo-randomized manner with an event-related design. From the first to fifth trial, each condition was presented once in a random order. A null event with a fixation cross was presented every 5 trials. Such set (5 trials of experimental conditions + 1 null event) repeated 50 times with a 10-second inter-trial interval (ITI). The total time of the fMRI scan was about 50 minutes. The stimuli presentation was controlled by the presentation software (Neurobehavioral Systems, CA, USA). The video stimuli were displayed on a screen placed behind the scanner and the participants saw the videos via a mirror above their eyes. The sounds were presented through an MRI compatible earphone (Silent Scan, Avotec, FL, USA). Following the clip, a forced-choice probe with four alternatives was presented. The four alternatives consisted of 1 correct word, 2 incorrect words and "None". The words in the alternatives presented only one mora by Hiragana and the rest of 2 moras were replaced with asterisk, because it was too difficult to find words that had high familiarity and pronunciation to match the stimuli. In the three alternatives of words, Hiragana was shown in 1st, 2nd, and 3rd mora respectively. The participants were instructed to press the response buttons with their index or middle finger of both hands as quickly as possible.

2.4. Behavioral data analysis

Mean reaction times (RTs) and accuracy rates were calculated for each condition in each participant. RTs were analyzed by within-subject three-way ANOVAs with factors INTELLIGIBILITY (16 band / 32 band), INTEGRATION (A / AV) and CORRECTNESS (correct / incorrect). The accuracy rates were also analyzed by within-subject 2-way ANOVAs with the factors of INTELLIGIBILITY and MODALITY.

2.5. fMRI data acquisition

A 3T MRI scanner (MAGNETOM Skyra; Siemens, Germany) was used to obtain 1500 scans for each experimental session with a gradient-echo echo-planner imaging (EPI) sequence (repetition time [TR] = 2000 ms, echo time [TE] = 30 ms, flip angle = 90 degrees, field of view = 192 × 192 mm, matrix 64 × 64, 35 slices, slice thickness = 3 mm with 1 mm gap, 3 × 3 × 4 mm3 resolution, bandwidth = 1816 Hz/pixel). The slices

were aligned to the AC-PC plane, achieving the whole brain coverage. After the experimental sessions, T1-weighted MPRAGE images were obtained as high-resolution anatomical images used for preprocessing (inversion time = 900 ms, TR = 2300 ms, TE = 2.98, flip angle = 9 degree, field of view = 256 × 256 mm, matrix 256 × 256, sagittal 224 slices, 1 mm isotropic resolution, bandwidth = 240 Hz/pixel).

2.6. fMRI data analysis

The fMRI data were processed with the SPM12 software package (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, United Kingdom). The functional images were realigned to the mean image, and the difference of slice acquisition timing was corrected. Then, the functional images were coregistered to individuals' anatomical images that were used for the estimation of spatial normalization parameters. Subsequently, all functional images were spatially normalized to the Montreal Neurological Institute (MNI) space with voxel resampling at $3 \times 3 \times 3$ mm^3 before being smoothed with a 6-mm full width at half maximum (FWHM) Gaussian kernel. The condition effects in each voxel were estimated per participant by the general linear model. The sentence presentation periods (with 5.6 s duration) and the probe presentation periods (with 3 s duration) in each condition were modeled with box-car functions that were convolved with a canonical hemodynamic response function, so that the 10 regressors (correct or incorrect trials in five conditions) were included in the design matrix. The six realignment parameters were included as covariates of no interest in the model to account for movement-related variance. Low-frequency noise was removed using a high-pass filter with a cut-off period of 128 s. Signal increase relative to the baseline in each condition of each participant was estimated according to the general linear model. The resulting individual contrast images were normalized using the DARTEL parameters, smoothed with 6-mm full-width at half-maximum Gaussian kernel, and submitted to the second-level (group) analysis, a within-subject three-way ANOVAs with factors CORRECTNESS (correct / incorrect), MODALITY (A / AV) and INTELLIGIBILITY (16 band / 32 band), but cluster level inference is not available.

Since we had a critical interest in the pSTS, M1 and IFG, we also performed the 2x2x2 within-subject ANOVAs within the pSTS, M1 and IFG. In these analyses, the sensitivity of the statistical tests increases because the search volumes are much smaller than the whole brain. We created an averaged MRI of the participants to build masks for the pSTS, M1 and IFG. Since previous studies have demonstrated that pSTS, M1, and IFG have critical roles in audiovisual integration, a total of three VOIs (Volume of interests) were defined as the overlap areas between anatomical and functional masks. The anatomical masks within pSTS, M1, and IFG were defined as pSTS, precentral gyrus, and IFG. The functional mask within pSTS was defined as a sphere centered at the MNI coordinates in previous studies [3, 11, 12, 14, 15] for the left pSTS [-53, -51, 11] with the radius being the standard deviation of the Y values (5 mm). The functional masks within M1 and IFG were defined as 6 mm radius spheres centered at the MNI coordinates in previous studies for the M1 lip [-42, 24, 6], tongue [-54, -54, 3] [24], and IFG (BA44) [-46, 26, 20] [8].

The selection of local maxima was performed based on the participants' own statistical maps using the F-contrast for the effects of interest contrast with a threshold of $p < 0.001$ (uncorrected for multiple comparisons) (the null hypothesis for this contrast is no activation for all of the correct trials for the five experimental conditions). The VOI time series data of each condition were then extracted as eigenvariates from the voxels that were significant and adjusted ($p < 0.001$ uncorrected) for the effects of interest contrast.

2.7. Trial time course analysis (TTC)

According to Makuuchi et al. [22], TTCs were estimated for each participant's preprocessed (upsampled to have data points at every 0.51 sec by a piecewise cubic Hermite interpolation, high-pass filtered [128 s], and linear trend removed) time series data of each session. The TTCs for each condition were estimated as follows: The TTC of correct and incorrect trials for each condition in a VOI were modeled with 37 variables representing the BOLD (blood oxygen level dependent) signals every 1 sec from -4 to 16.0 sec after the sound onset. Then, we made a general linear model:

$$Y = X * \beta + e$$

where Y represents the preprocessed time series data of a VOI, X the design matrix, β the coefficients of explanatory variables, and e the error term. The coefficients (βs) were estimated at each time point, assuming a linear time invariant system. The estimated βs were then averaged across sessions. The estimated values reached the peak around 5-6 sec after sound onset for all conditions. To analyze the time course of sound perception process among conditions, we set 3 time windows: before peak (3-4 sec after sound onset), peak (5-6 sec after sound onset), and after peak (7-8 sec after sound onset). A within-subject three-way ANOVA with the factors CORRECTNESS (correct / incorrect), MODALITY (A / AV) and INTELLIGIBILITY (16 bands / 32 bands) was carried out mean value of 3 every 2 seconds from 1 sec to 10 sec after the sound onsets to inspect the condition effects within each time window.

3. Results

3.1 Behavioral Results

Results of Reaction Times (RTs) to the 4-alternative probes showed significant main effects of all 3 factors, without an interaction. (CORRECTNESS; $F1, 16 = 126.86$, $p < 0.001$, MODALITY; $F1, 16 = 13.36$, $p < 0.003$, INTELLIGIBILITY; $F1, 16 = 7.11$, $p < 0.03$) (Fig. 1). The RTs of the correct, AV, and 32-band conditions were significantly shorter than that of the incorrect, A, and 16-band conditions respectively. In terms of accuracy, an interaction was not significant but the main effects of all factors were significant (MODALITY; $F1, 16 = 119.45$, $p < 0.001$, INTELLIGIBILITY; $F1, 16 = 9.10$, $p < 0.01$). For the AV and 32-band conditions, the accuracy was significantly higher than that of the A and 16-band conditions respectively.

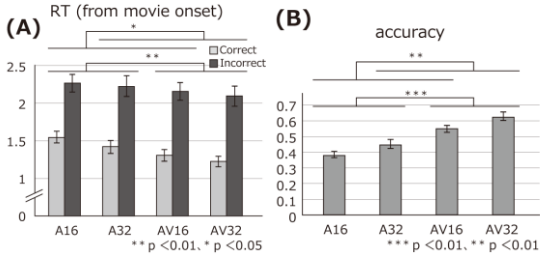

Figure 1: *Results of behavioral task*

3.2 Image Results

We excluded the M1 tongue area from analysis because the estimated TTCs did not show activation in all participants. In the other VOIs, we excluded some participants because no significantly activated voxels were found in the VOI. The numbers of participants for the analyses were 13 for the pSTS, 12 for the M1 lip area, and 15 for the IFG. ANOVAs did not show a three-way interaction for all time windows in all VOIs.

3.2.1 From 3 sec to 4 sec

In the pSTS, the interaction was not significant, but a main effect of INTELLIGIBILITY was significant ($F1, 12 = 4.37$, $p = 0.05$). The activations for the 16-band conditions were significantly larger than that for the 32-band conditions. In the IFG, an interaction was not significant but the main effects of CORRECTNESS and MODALITY were significant (CORRECTNESS: $F1, 14 = 14.23$, $p < 0.003$, MODALITY: $F1, 14 = 16.10$, $p < 0.003$). The activations for the correct trials and the A condition were significantly larger than that for the incorrect trials and the AV condition respectively. The M1 lip area showed no significant difference (Fig. 2).

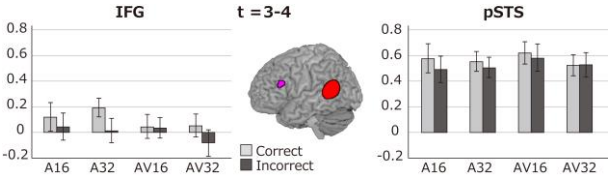

Figure 2: *Brain activity (t = 3-4)*

3.2.2 From 5 sec to 6 sec

In the pSTS, an interaction between CORRECTNESS and MODALITY was significant ($F1, 12 = 5.94$, $p < 0.05$). Analyses for the simple main effect revealed a significant larger activation for the AV condition than for A in the incorrect trials ($F1, 38 = 5.92$, $p = 0.07$). In the M1 lip area, an interaction between CORRECTNESS and INTELLIGIBILITY was significant ($F1, 11 = 4.79$, $p = 0.05$) but the simple main effects were not. The IFG showed no significant difference (Fig. 3).

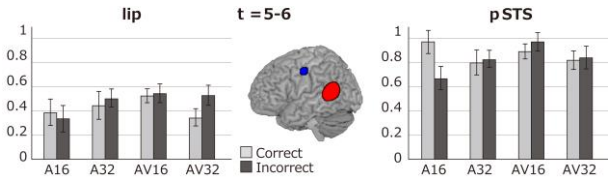

Figure 3: *Brain activity (t = 5-6)*

3.2.3 From 7sec to 8 sec

In the pSTS, there was no significant difference. In the M1 lip area, an interaction between CORRECTNESS and MODALITY was significant ($F1, 11 = 4.41$, $p = 0.06$) but the simple main effects were not. In the IFG, an interaction between CORRECTNESS and MODALITY was significant ($F1, 14 = 21.68$, $p < 0.0005$), as well as the simple main effects of both factors. The activation was significantly larger for the incorrect than correct conditions within the AV conditions ($F1, 44 = 26.14$, $p < 0.0001$), and larger for the A condition than the AV condition regardless of correctness (correct; $F1, 44 = 4.68$, $p < 0.05$, incorrect; $F1, 44 = 7.25$, $p < 0.03$) (Fig. 4).

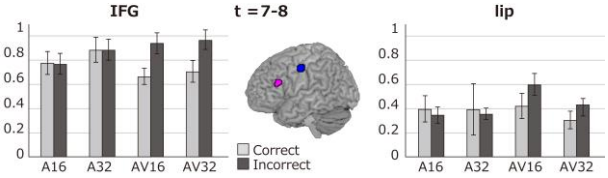

Figure 4: *Brain activity (t = 7-8)*

4. Discussion

In this study, we performed an fMRI study of improved perception of degraded noise-vocoded speech by integrating visual face motion. The results of the behavioral task demonstrated perception improvement for degraded speech sounds accompanied by facial videos, in line with Sumby & Pollack's study [1]. The results of fMRI showed that the dorsal pathway including pSTS, M1 and IFG were involved in the successful AV integration at different time points.

In the time window before the BOLD response peak (3-4s), a larger activation for the 16-band noise-vocoded sounds than the 32-band sounds was detected in the pSTS, known as voice area [25]. This result was opposite to previous studies [3, 4]. This discrepancy may result from the intelligibility of sound stimuli. The noise vocoded sounds used in previous studies (6 or 8 band) were more degraded than the sounds used in current study (16 or 32 bands) and the participants in the previous studies could not perceive most of the sounds as speech. The finding that more intelligible sounds were associated with greater pSTS activation suggests that the pSTS activation might reflect the ratio of sounds that participants could perceive as speech. On the other hand, the lowest accuracy of the current study (16 band- A condition) was about 40 % (the level of at-chance was 25%), indicating that the participants perceived more stimuli as speech than previous studies. The results of current study support pSTS selectivity for speech sounds, with the pSTS being more activated for lower intelligible speech.

In the IFG, the activity for the A condition was significantly larger than the AV condition. This larger activation for the A condition may reflect the processing in the ventral pathway. The ventral pathway consists of the inferior front-occipital fascicle connecting the pSTS and IFG, implied in semantic processing [26]. The activation of the IFG was also larger for the correct trials than incorrect trials. The IFG is known to play a role in selecting words by accessing and searching the mental lexicon [27]. Accordingly, the IFG activation may indicate that the IFG supports word finding in the mental lexicon through the ventral pathway for stimuli perceived as speech without using motor information.

At the peak time window (5-6 s), an interaction between INTEGRATION and CORRECTNESS was significant in the pSTS, driven by the smaller activation for A16 incorrect trials relative to the other conditions (Fig 3). This suggests that the participants could not perceive the stimuli as speech sounds in the A16 incorrect trials. In other words, the participants could perceive 16-band sounds as speech if the face video was accompanied. In this sense, visual information facilitates speech perception by AV integration mechanism in the pSTS. In the hypothesis-and-test model, perception occurs only after the integration of the predicted motor representation and input

sensory information [19]. In association with the RTs of the behavioral task that showed shorter response times for correct trials, the neural computations for correct trials finish earlier than incorrect trials. This is in line with Skippers' model [16] which suggests that the number of hypothesis-and-test cycles for processing strongly predicable sounds is smaller than that for weakly predicable sounds. Therefore, the activation of the pSTS at the peak time window represents the first matching process between the prediction from hypothesis and the sensory input for speech sounds.

In the M1 lip area, the activation for the incorrect trials with 32-band sounds increased whereas that for the correct trials decreased. The increase in activation for incorrect trials may reflect repeated mental simulation of articulation of a predicted word generated in the IFG. Since the participants were asked to pick the word from the 4 alternatives presented, they may regenerate other candidate words that matched to the one of the choices repeatedly until they find the target one. In other words, the larger activity in the M1 lip area represents the second hypothesis-and-test cycle [20] whereas the smaller activity for AV32 correct trials may indicate that sound processing is complete. This interpretation is supported by the shorter RTs for the AV32 correct trials.

In the post-peak time window (7-8 s), there was no significant difference in the pSTS, whereas an interaction between INTEGRATION and CORRECTNESS was significant in the M1 lip and IFG. This indicates that the pSTS activation in this time window was not strictly relevant to sound processing after the second matching process between the hypothesis and the prediction. As mentioned above, the participants should select the alternative that represented a part of the word in the task. If the first matching between the prediction and the sensory input failed, the matching process to accomplish the task may change to another matching process between the prediction from motor simulation and the prediction from the displayed alternative. The latter matching process may involve the IFG for word selection/searching in mental lexicon. Thus, the neural activation in this time window may be associated with the interaction between the M1 lip and the IFG. In both areas, the activation increased for the AV incorrect trials. This may reflect the prediction from mental simulation of articulation in the M1 lip, and a selection demand in the IFG [27]. In other words, IFG is engaged in the selection of word from mental lexicon to compare with the prediction generated by mental simulation of articulation.

A major limitation of this study is that our speech stimuli consisted of only one speaker, and therefore it may be difficult to generalize about the neural mechanisms of successful AV perception based on the results of current study. Future studies that use multiple speakers are needed.

5. Conclusions

Although we did not find critical regions that were associated with performance improvement in speech perception accompanied by facial videos as a main effect of CORRECTNESS, we identified the brain network associated with degraded speech in four different levels based on the hypothesis-and-test model [16]. At the first level, when the speech sound is intelligible enough, the IFG is engaged for retrieving an appropriate word from the mental lexicon stored in the temporal lobe via the ventral pathway. At the second level, when the sound is not perceived as speech, the pSTS supports the integration of the sound and the visual information. At the third level, when the participants perceive the sounds as speech without identifying the actual word, the dorsal pathway supports the processing of the hypothesis-and-test model [16]. The M1 lip area generates a hypothesis from the facial movements by motor simulation, and the pSTS is recruited for checking if the input sound and the hypothesized word match. At the fourth level, in cases where the hypothesis generated for the input sound does not match any of the alternative choices, the hypothesis generation and verification in respond to the input repeat until it finds to a solution.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP 14457703.

7. References

[1] W. H. Sumby and I. Pollack "Visual Contribution to Speech Intelligibility in Noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212-215, 1954.

[2] R. V.Shannon, F. G. Zeng, V.Kamath, J.Wygonski, and M. Ekelid "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303-304, 1995.

[3] C. McGettigan, A. Faulkner, I. Altarelli, J. Obleser, H. Baverstock, and S. K. Scott "Speech comprehension aided by multiple modalities: behavioural and neural interactions," *Neuropsychologia*, vol. 50, no. 5, pp. 762-776, 2012.

[4] J. Obleser, R. J. Wise, M. A. Dresner, and S. K. Scott "Functional integration across brain regions improves speech perception under adverse listening conditions," *J Neurosci*, vol. 27, no. 9, pp. 2283-2289, 2007.

[5] H. E. Nuttall, D. Kennedy-Higgins, J. Hogan, J. T. Devlin, and P. Adank "The effect of speech distortion on the excitability of articulatory motor cortex," *Neuroimage*, vol. 128, pp. 218-226, 2016.

[6] C. J. Wild, A. Yusuf, D. E. Wilson, J. E. Peelle, M. H. Davis, and I. S. Johnsrude "Effortful listening: the processing of degraded speech depends critically on attention," *J Neurosci*, vol. 32, no. 40, pp. 14010-14021, 2012.

[7] A. A. Zekveld, D. J. Heslenfeld, J. M. Festen, and R. Schoonhoven "Top-down and bottom-up processes in speech comprehension," *Neuroimage*, vol.32, no. 4, pp. 1826-1836, 2006.

[8] F. Eisner, C. McGettigan, A. Faulkner, S. Rosen and S. K. Scott "Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations," *J Neurosci*, vol. 30, no. 21, pp. 7179-7186, 2010.

[9] A. Hervais-Adelman, M. H. Davis, I. S. Johnsrude and R. P. Carlyon "Perceptual learning of noise vocoded words: effects of feedback and lexicality," *J Exp Psychol Hum Percept Perform*, vol. 34, no. 2, pp. 460-474, 2008.

[10] A. G. Hervais-Adelman, R. P. Carlyon, I. S. Johnsrude and M> H. Davis "Brain regions recruited for the effortful comprehension of noise-vocoded words," *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 1145-1166, 2012.

[11] A. R. Nath and M. S. Beauchamp "Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech," *J Neurosci*, vol. 31, no. 5, pp. 1704-1714, 2011.

[12] M. S. Beauchamp, K. E. Lee, B. D. Argall and A. Martin, A "Integration of auditory and visual information about objects in superior temporal sulcus," *Neuron*, vol. 41, no. 5, pp. 809-823, 2004.

[13] G. A. Calvert, R. Campbell and M. J. Brammer "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex," *Curr Biol*, vol. 10, no. 11, pp. 649-657, 2000.

[14] L. C. Erickson, B. A. Zielinski,J. E. Zielinski, G. Liu, P. E. Turkeltaub, A. M. Leaver and J. P. Rauschecker "Distinct cortical locations for integration of audiovisual speech and the McGurk effect," *Front Psychol*, vol. 5, 534, 2014

[15] H. Lee, and U. Noppeney "Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension," *J Neurosci*, vol. 31, no. 31, pp. 11338-11350, 2011.

[16] J. I. Skipper, V. van Wassenhove, H. C. Nusbaum and S. L. Small "Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception,". *Cereb Cortex*, vol. 17, no.10, pp. 2387-2399, 2007.

[17] L. Aziz-Zadeh, L. Cattaneo, M. Rochat, and G. Rizzolatti "Covert speech arrest induced by rTMS over both motor and nonmotor left hemisphere frontal sites," *J Cogn Neurosci*, vol. 17, no. 6, 928-938, 2005.

[18] P. Dechent, K. D. Merboldt, and J. Frahm "Is the human primary motor cortex involved in motor imagery?" *Brain Res Cogn Brain Res*, vol. 19, no. 2, pp. 138-144, 2004.

[19] J. I. Skipper, H. C. Nusbaum and S. L. Small "Listening to talking faces: motor cortical activation during speech perception," Neuroimage, vol. 25, no. 1, pp. 76-89, 2005.

[20] J. I. Skipper "Echoes of the spoken past: how auditory cortex hears context during speech perception" *Philos Trans R Soc Lond B Biol Sci*, vol. 369, 1651, 20130297, 2014.

[21] K. N. Stevens "Remarks on analysis by synthesis and distinctive features," *Models for the perception of speech and visual form*, 1967.

[22] M. Makuuchi, J. Bahlmann, and A. D. Friederici "An approach to separating the levels of hierarchical structure building in language and mathematics," *Philos Trans R Soc Lond B Biol Sci*, vol. 367, no. 1598, pp. 2033-2045, 2012.

[23] S. Amano S *NTT Databese Series. Nihongo-no Goitokusei (Lexical properties of Japanese),* Vol. 7. Tokyo, Japan: Sanseido, 2000

[24] M. R. Schomers, E. Kirilina, A. Weigand, M. Bajbouj and F. Pulvermuller "Causal Influence of Articulatory Motor Cortex on Comprehending Single Spoken Words: TMS Evidence," *Cereb Cortex*, vol. 25, no. 10, pp. 3894-3902, 2015.

[25] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike "Voice-selective areas in human auditory cortex," *Nature*, vol. 403, no.6767, pp. 309-312, 2000.

[26] S. M. Gierhan "Connections for auditory language in the human brain," *Brain Lang*, vol. 127, no. 2, pp. 205-221, 2013.

[27] P. Tremblay, and V. L. Gracco "On the selection of words and oral motor responses: evidence of a response-independent fronto-parietal network," *Cortex*, vol. 46, no. 1, pp. 15-28, 2010.