A STUDY ON THE MECHANISMS OF SENSORIMOTOR INTERACTIONS IN PRODUCTION AND PERCEPTION OF SPEECH PHONEMES

Takemi Mochida

Graduate School of Systems Information Science

Future University-Hakodate

March 2011

Abstract

In this thesis, aiming at elucidating the functional basis of phoneme production and perception processes enabling inter-human speech communication, interdependence underlying speech motor control and auditory processing systems were studied based on physiological and psychophysical experiments. The dynamic behavior of speech organs during speech production has been found to be affected by self-produced auditory feedback signals. However, the functional characteristics of auditory feedback-based control of speech movements have not been fully elucidated. The involvement of speech motor control system in speech perception has also been evident. However, there have been controversial arguments regarding how incoming auditory information is processed and interpreted by the motor system and triggers a specific phoneme perception. From the above viewpoints, this thesis investigated the interaction mechanisms involved in the production and perception of speech phonemes, based on the following two experimental approaches: (1) as for speech production, evaluating changes in articulatory lip movement during speaking the bilabial plosive ([p]) repetitively, which occurred when a sudden alteration in timing and context of auditory feedback was introduced, and (2) as for speech perception, evaluating changes in phoneme intelligibility for bilabial, alveolar and velar plosives ([p], [t], and [k], respectively), which occurred when hearing them simultaneously with whispering (silently articulating) incongruent phonemes, in comparison with viewing a mouth motion producing incongruent phonemes.

The critical result of the first experiment was that the articulatory lip movement quickened immediately when the auditory feedback virtually preceded the expected timing by 50 ms. Such articulatory change was not observed when the feedback was presented more than 50 ms earlier or later than the actual timing, or when the feedback syllable was replaced by other one. These results suggest that errors between the internally predicted and actually provided auditory information detected in a temporally asymmetric window contribute to the compensation for the inter-articulatory timing in the syllable repetition task.

The critical result of the second experiment was that the silent articulation affected the phoneme perception in a different way from the well-known visual effects (ex., McGurk effect). Viewing lip motion ([p]) degraded hearing of phonemes produced by the tongue ([t] and [k]), and viewing tongue motions degraded hearing of phoneme produced by the lips, replicating the previous studies. On the contrary, articulating phoneme with the

lips ([p]) did not affect hearing of tongue-related phonemes ([t] and [k]), and articulating phonemes with the tongue did not affect hearing of phoneme produced by the lips. More interestingly, articulating phonemes with the tongue ([t] and [k]) degraded hearing of the other tongue-related phonemes ([k] and [t], respectively). These results suggest for the first time that motor interferential effect on speech perception is mediated by a different mechanism from the visual one: the auditory-visual integration occurs across different speech organs, whereas the auditory-articulatory integration occurs within the same organ.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor Prof. Dr. Nobuhiro Miki for his continuous support, guidance and patience throughout this work. My deepest appreciation goes to Prof. Dr. Masaaki Honda of Waseda University and Prof. Dr. Tokihiko Kaburagi of Kyushu University for their encouragement throughout the years. I would also like to thank Prof. Dr. Yuichi Fujino, Prof. Dr. Osamu Takahashi, and Prof. Dr. Yoshio Uwano of Future University-Hakodate for their insightful suggestions.

This work could not have been completed without the help of Dr. Hiroaki Gomi, Dr. Makio Kashino, Dr. Tadahisa kondo, Dr. Toshitaka Kimura, Dr. Sadao Hiroya, and Dr. Norimichi Kitagawa of NTT Communication Science Laboratories. Especially, I have greatly benefited from the constructive advice of Dr. Gomi and the ideas of Dr. Kimura.

Special thanks to Prof. Dr. Katsuhiko Shirai and Prof. Dr. Tetsunori Kobayashi of Waseda University who gave me the first chance to start the research experience.

Finally, I would like to thank my mother and brother for always encouraging me.

This thesis is dedicated to the memory of my father.

Co-authorship Statement

In Chapter 2, Takemi Mochida conceived and designed the experiments, and analyzed the data. Takemi Mochida and Hiroaki Gomi contributed to materials, and analysis tools. Takemi Mochida, Hiroaki Gomi, and Makio Kashino prepared the manuscript.

In Chapter 3, Takemi Mochida and Toshitaka Kimura contributed equally to conceiving and designing the experiments, and analyzing the data. Takemi Mochida, Toshitaka Kimura, Sadao Hiroya, Norimichi Kitagawa, Hiroaki Gomi, and Tadahisa Kondo prepared the manuscript.

Contents

List of Figures

List of Tables

1. Introduction and orientation

1.1 Motivation of the thesis

Understanding the mechanisms governing speech production and perception skills is essential to ensure satisfactory quality of remote communication systems. Speech communication over a remote network such as using an audio-visual conferencing system is vulnerable to transmission delay inherent to its processing. In order to guarantee successful speech communication on a system, the quantitative knowledge of dynamic characteristics of speech motor control and auditory processing is extremely important. Even a small amount of delay in the audio and visual transmission in a future system may still induce some undesirable effects in speech communication.

Echo is one of major known interferential sources, where the talker's speech transmitted and reflected at the listener end returns to the talker end, resulting in a delayed auditory feedback (DAF) of his/her own speech. It has been well known that speaking with exposure to DAF leads to various types of speech disfluencies, e.g., increased articulatory error, lengthened duration, augmented volume, and increased fundamental frequency [1-5]. Such disfluencies may occur as a result of several different types of voluntary and involuntary responses to DAF. The Lombard effect (or Lombard reflex) is well cited as an example of auditory-induced automatic motor response to the change in background noise level, where speakers involuntarily increase their vocal intensity according to increasing noise level [6, 7]. A reflexive adjustment of voice pitch based on auditory information is also evident to some extent [8, 9]. Although these sorts of reflexive mechanisms can be considered as the potential sources for such speech disfluencies induced by DAF, the precise mechanisms of how the delayed auditory input of self-produced speech can adversely affect the speech motor control has not been fully elucidated yet.

Asynchronicity of audio-visual transmission, which probably has not yet been taken so seriously as the echo in the current audio-visual conferencing systems, can be a potential source of another undesirable effect in speech communication. If sufficiently high time and spatial resolution of visual transmission is available, even a small amount of time lag between the auditory and visual information may occasionally affect speech perception. The McGurk effect is well cited as an example of audio-visual multisensory integration process in phoneme perception, where an illusory phoneme perception is produced when auditory information is accompanied by visual lip movement that does not match the auditory information. For example, an auditory [ba] combined with a visual [ga] is typically heard as [da] [10]. A study also has demonstrated that the perception of auditory phonemes can be disturbed when

the listener silently articulates incongruent phonemes, instead of seeing the visual motion of mouth [11]. Furthermore, a neurological study has demonstrated that seeing speech-related lip movements modulates the excitability of cortical representation of the lip muscles in a similar manner as hearing speech sounds [12]. These lines of evidence naturally lead to an interesting question of whether both the visual input and the self articulatory movement associated with an identical phoneme act in a same manner or even on an identical pathway within the neural circuit responsible for speech perception. However, little has been studied about how the self articulatory movements of each speech organ can affect speech perception [13].

All the above issues concerning the specific version of speech communication strongly suggest the existence of sensorimotor interactions related to speech production and perception. In the following subsection, accumulated evidence from neurophysiological and neuroimaging studies showing the link between speech production and perception will be briefly described.

1.2 Sensorimotor nature of speech production and perception

Speech articulators are controlled by neuronal activity in the inferior frontal motor cortex, which is in turn controlled by inferior frontal premotor and prefrontal circuits. At the same time, speech sounds elicit activity in the auditory system, mainly in superior temporal primary auditory, auditory belt and parabelt areas. Importantly, connections between these sites enable the cortex to strengthen these links and thereby store correlations between inferior frontal neurons that contribute to articulatory actions and superior temporal neurons that are involved in auditory perception. As neurons that frequently fire together strengthen their mutual connections, early articulations lead to the emergence of action–perception circuits for phonemes. The babbling sounds of infants during their first to second years of life become increasingly similar to the types of speech sounds that they hear frequently, suggesting that acoustic–phonological tuning to language-specific sounds occurs at a very early stage of development [14, 15]. Synaptic strengthening due to coactivation also suggests that somatosensory neurons might be involved in these action–perception circuits [16-19].

The neural basis of sensorimotor interactions during vocal production has been studied in both human and non-human primates [20-22]. As for speech production, a brain imaging study has shown that cerebral blood flow (CBR) in the auditory areas increase when syllables are whispered with the auditory input masked by noise, suggesting motor-induced changes in auditory activity during speaking [23]. Another study using magnetoencephalographic (MEG) recording has revealed that auditory cortical response to self-produced speech during speaking is attenuated compared with that during listening to a tape-playback, suggesting auditory activity modulation as a function of the expected acoustic feedback during speech production [24].

Studies of the neural basis of sensorimotor interactions for speech perception have been triggered by the discovery of sensorimotor neurons that are active during action execution and observation of corresponding action (mirror neurons, [25]), and even during audition of corresponding action-related sound (audiovisual mirror neurons, [26]), suggesting action–perception integration at the neuronal level. Consistent with theories postulating that speech motor control mechanisms are important in speech perception[27, 28], a functional magnetic resonance imaging (fMRI) study has shown that motor systems involved in speech production are critically involved in perceiving speech sounds [29]. Studies involving the application of transcranial magnetic stimulation (TMS) to motor areas have shown changes in the perception of speech presented in noise, indicating some role of the motor system at least under restricted conditions [30, 31].

There have been controversial studies claiming that the activation of motor speech areas during speech perception is weak, suggesting the restricted role of speech motor system [32, 33]. Patients with lesions in the left inferior frontal regions causing severe motor speech deficits (Broca's aphasia) still have intact speech perception [34]. Such evidence of a functional dissociation between production and perception questions the existence of sensorimotor interactions. Other studies have supported the involvement of speech motor system at the stages of semantic and lexical processing rather than phonological processing. For example, reading action words activates motor cortex in a somatotopic fashion [35], and stimulation of motor cortex via TMS affects response time in a lexical decision task [36]. All the above investigations, however, do not deprive the sensorimotor system of its critical role in motor control and auditory processing related to speech task, especially at the acoustic-phonological level. A more specific function of sensorimotor interactions in speech production and perception should be explored.

1.3 Plan of the thesis

The purpose of this thesis is to clarifying the sensorimotor functions in both production (chapter 2) and perception (chapter 3) of speech phonemes, independent from lexical, semantic or syntactic processing.

In chapter 2, the temporal aspect of auditory feedback-based articulatory control of the lips was studied using an auditory feedback alteration system. The highlight of the experiment was that speakers were exposed to sudden virtual advance or delay in auditory feedback timing during the repetitive and rhythmical production of isolated syllables. Behavioral changes in articulatory movements of the lips immediately after feedback alteration was introduced were evaluated using a 3D motion capture system. How sensory errors between the internally simulated and actually provided auditory information affect the associated speech motor control was statistically investigated in terms of timing and phoneme context.

In chapter 3, the direct effect of articulatory movements on speech perception was studied by evaluating phoneme intelligibility changes occurred when perceiving auditory phonemes with silently articulating the concordant/discordant phonemes. The highlight of the experiment was that listeners were exposed to auditory phonemes which are differently articulated but associated with the same articulator (i.e. [t] and [k], both tongue-related phonemes) via earphones while whispering those phonemes with their voices blocked by masking noise. The audio-visual integration effects were also observed for the same listeners for comparing with the auditory-articulatory integration effects. How auditory-articulatory integration in phoneme perception occurs was statistically investigated in terms of effector specificity.

In chapter 4 the summary of the thesis and future works were described.

1.4 References

1. Lee, B.S., *Effects of delayed speech feedback.* Journal of the Acoustical Society of America, 1950. 22(6): p. 824-826.
2. Tiffany, W.R. and C.N. Hanley, *Delayed speech feedback as a test for auditory malingering.* Science, 1952. 115(2977): p. 59-60.
3. Fairbanks, G., *Selective vocal effects of delayed auditory feedback.* Journal of speech and hearing disorders, 1955. 20(4): p. 333-346.
4. Zanini, S., et al., *Speaking speed effects on delayed auditory feedback disruption of speech fluency.* Percept Mot Skills, 1999. 89(3 Pt 2): p. 1095-109.
5. Stuart, A., et al., *Effect of delayed auditory feedback on normal speakers at two speech rates.* Journal of the Acoustical Society of America, 2002. 111(5): p. 2237-2241.
6. Lombard, E., *Le signe de l'elevation de la voix.* Annales maladies oreille larynx nez pharynx 1911. 37: p. 101-119.
7. Garnier, M., N. Henrich, and D. Dubois, *Influence of sound immersion and communicative interaction on the Lombard effect.* J Speech Lang Hear Res, 2010. 53(3): p. 588-608.
8. Burnett, T.A., et al., *Voice F0 responses to manipulations in pitch feedback.* J Acoust Soc Am, 1998. 103(6): p. 3153-61.
9. Burnett, T.A. and C.R. Larson, *Early pitch-shift response is active in both steady and dynamic voice pitch control.* J Acoust Soc Am, 2002. 112(3 Pt 1): p. 1058-63.
10. McGurk, H. and J. MacDonald, *Hearing lips and seeing voices.* Nature, 1976. 264(5588): p. 746-8.
11. Sams, M., R. Mottonen, and T. Sihvonen, *Seeing and hearing others and oneself talk.* Brain Res Cogn Brain Res, 2005. 23(2-3): p. 429-35.
12. Watkins, K.E., A.P. Strafella, and T. Paus, *Seeing and hearing speech excites the motor system involved in speech production.* Neuropsychologia, 2003. 41(8): p. 989-94.
13. Ito, T., M. Tiede, and D.J. Ostry, *Somatosensory function in speech perception.* Proc Natl Acad Sci U S A, 2009. 106(4): p. 1245-8.
14. Werker, J.F. and R.C. Tees, *Influences on infant speech processing: toward a new synthesis.* Annu Rev Psychol, 1999. 50: p. 509-35.
15. Dietrich, C., D. Swingley, and J.F. Werker, *Native language governs interpretation of salient speech sound differences at 18 months.* Proc Natl Acad Sci U S A, 2007. 104(41): p. 16027-31.
16. Westermann, G. and E. Reck Miranda, *A new model of sensorimotor coupling in the development of speech.* Brain Lang, 2004. 89(2): p. 393-400.
17. Wennekers, T., M. Garagnani, and F. Pulvermuller, *Language models based on Hebbian cell assemblies.* J Physiol Paris, 2006. 100(1-3): p. 16-30.

18. Garagnani, M., T. Wennekers, and F. Pulvermuller, *A neuroanatomically grounded Hebbian-learning model of attention-language interactions in the human brain.* Eur J Neurosci, 2008. 27(2): p. 492-513.
19. Garagnani, M., T. Wennekers, and F. Pulvermuller, *Recruitment and Consolidation of Cell Assemblies for Words by Way of Hebbian Learning and Competition in a Multi-Layer Neural Network.* Cognit Comput, 2009. 1(2): p. 160-176.
20. Curio, G., et al., *Speaking modifies voice-evoked activity in the human auditory cortex.* Hum Brain Mapp, 2000. 9(4): p. 183-91.
21. Eliades, S.J. and X. Wang, *Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations.* J Neurophysiol, 2003. 89(4): p. 2194-207.
22. Heinks-Maldonado, T.H., et al., *Fine-tuning of auditory cortex during speech production.* Psychophysiology, 2005. 42(2): p. 180-90.
23. Paus, T., et al., *Modulation of cerebral blood flow in the human auditory cortex during speech: role of motor-to-sensory discharges.* Eur J Neurosci, 1996. 8(11): p. 2236-46.
24. Houde, J.F., et al., *Modulation of the auditory cortex during speech: an MEG study.* J Cogn Neurosci, 2002. 14(8): p. 1125-38.
25. Rizzolatti, G., et al., *Premotor cortex and the recognition of motor actions.* Brain Res Cogn Brain Res, 1996. 3(2): p. 131-41.
26. Kohler, E., et al., *Hearing sounds, understanding actions: action representation in mirror neurons.* Science, 2002. 297(5582): p. 846-8.
27. Liberman, A.M., et al., *Perception of the speech code.* Psychol Rev, 1967. 74(6): p. 431-61.
28. Liberman, A.M. and I.G. Mattingly, *The motor theory of speech perception revised.* Cognition, 1985. 21(1): p. 1-36.
29. Wilson, S.M., et al., *Listening to speech activates motor areas involved in speech production.* Nat Neurosci, 2004. 7(7): p. 701-2.
30. Meister, I.G., et al., *The essential role of premotor cortex in speech perception.* Curr Biol, 2007. 17(19): p. 1692-6.
31. D'Ausilio, A., et al., *The motor somatotopy of speech perception.* Curr Biol, 2009. 19(5): p. 381-5.
32. Lotto, A.J., G.S. Hickok, and L.L. Holt, *Reflections on mirror neurons and speech perception.* Trends Cogn Sci, 2009. 13(3): p. 110-4.
33. Scott, S.K., C. McGettigan, and F. Eisner, *A little more conversation, a little less action--candidate roles for the motor cortex in speech perception.* Nat Rev Neurosci, 2009. 10(4): p. 295-302.
34. Moineau, S., N.F. Dronkers, and E. Bates, *Exploring the processing continuum of single-word comprehension in aphasia.* J Speech Lang Hear Res,

2005. 48(4): p. 884-96.
35. Hauk, O., I. Johnsrude, and F. Pulvermuller, *Somatotopic representation of action words in human motor and premotor cortex.* Neuron, 2004. 41(2): p. 301-7.
36. Pulvermuller, F., et al., *Functional links between motor and language systems.* Eur J Neurosci, 2005. 21(3): p. 793-7.

2. Physiological study of the auditory feedback control of articulatory lip movement

2.1 Background

During the development of speech production, different sorts of sensory feedback help to coordinate the movements of the respiratory, laryngeal, velopharyngeal, and articulatory subsystems. Cutaneous and/or somatosensory information on the status of multiple articulators and auditory information related to produced speech constitute important sources of feedback for speech motor control [1]. Various studies employing auditory feedback alteration have suggested that acoustic information is critical as regards learning and maintaining vowel production [2, 3] and voice pitch control [4, 5]. Evidence has also been obtained from humans and non-human primates showing that neural activity in the auditory cortex is modulated by self-produced vocalization [6-9]. In concert with these studies, theoretical models of speech acquisition and production have been proposed, which hypothesize that speech targets represented in auditory space are achieved using an articulatory-to-auditory map trained on self-produced auditory feedback [10, 11]. However, the debate continues as to whether such neural mechanisms also help to ensure stability in rapid and complex speech motor control [12, 13], aside from the well-studied reflexive adjustment of voice volume or pitch based on auditory information [5, 14-18]. Certain aspects of the effects of auditory feedback on speech articulation have been examined using the delayed auditory feedback (DAF) paradigm [19-23] where various types of speech disfluencies are induced, e.g., increased articulatory error, lengthened duration, augmented volume, and increased fundamental frequency. Similarly, a vocal duration reduction with an accelerated auditory feedback delay has also been reported [24]. However, the mechanisms that underlie these effects elicited by constant exposure to unusual feedback delay remain unclear. Auditory feedback may serve as an immediate source for the dynamic control of speech articulation, analogous to the well-known rapid adjustment of labial constriction based on cutaneous and/or somatosensory information [25-29].

In this study, the online control mechanism for articulatory lip movement was examined by suddenly shifting the auditory feedback timing in the ahead-of-time or delayed direction, and/or replacing the feedback syllable by other syllables, during the repetition of bilabial plosives /pa/. Labial distance trajectories under altered and normal feedback conditions were compared within a single cycle of lip closing/opening movement subsequent to the auditory alteration. Statistical analysis revealed that a quickened lip

closing/opening movement was clearly elicited when the auditory feedback preceded the real production by 50 ms. On the other hand, such change was not significant when the feedback was provided more than 50 ms before the real production or was delayed, and/or when the feedback syllable was replaced by /Φa/ or /pi/. These results suggest (1) an underlying mechanism that detects errors between anticipated and actually provided auditory consequences for the rapid modification of subsequent movements, and (2) a temporally asymmetric window for detecting auditory errors in which acoustic features of the syllable to be produced may be coded.

2.2 Materials and methods

2.2.1 Ethics Statement

All participants gave their written informed consent to participating in this study, which was approved by the Research Ethics Board of NTT Communication Science Laboratories.

2.2.2 Participants

Ten adults (seven males and three females) aged from 21 to 39 participated in the experiments. All the participants were native speakers of Japanese and exhibited no obvious speech difficulties as judged by the experimenters.

2.2.3 Apparatus

Figure 2.1A is a schematic diagram of the auditory feedback alteration system. The speech sounds produced by a participant are converted into voltage signals by an electret condenser microphone (Sony ECM-G3M driven by an Earthworks Microphone Preamp 1021). The signals are then filtered (NF 48 dB/oct filter P-85 in the phase-linear low pass mode) with a cutoff frequency of 6 kHz, and digitized at a sampling frequency of 16 kHz (Systems Design Service DASBOX Model-16/100). A custom made program for altering the input speech signals with a buffer size corresponding to 10 ms is run on a workstation. The processed signals are then converted to voltage signals (Systems Design Service DASBOX-16) and filtered (NF 48 dB/oct filter P-85 in the phase-linear low pass mode) with a cutoff frequency of 6 kHz. Finally, the voltage signals are converted into acoustic sounds and fed back to the participant bilaterally using earphones (Etymotic Research earphones ER-4S driven by Sony audio mixer SRP-X6004).

In the experiment, the participants sat on a chair and were asked to insert the earphones as deeply as possible in the ear canal. A microphone mounted in a floor stand was located close to the left ears of the participants who were asked to keep their heads in a fixed position throughout the experiments. The participants heard their own unaltered speech picked up by the microphone through the earphones while vocalizing an /a/ sound in their natural way. They were then asked to adjust the gain of the microphone so that they heard their own speech sounds most naturally. The participants were also asked to adjust the sound level of the pink noise they heard through the earphones, which was produced by a noise generator (Bruel & Kjaer Type 1405), while vocalizing an /a/ sound in their natural way, so that, as far as possible, they did not perceive their own bone-conducted auditory feedback, but without experiencing stress. The sound level of the noise chosen by the ten participants in the experiments was 61.5±2.75 dBSPL as measured by a probe microphone

(Etymotic Research Probe Microphone ER-7C).

The in-the-ear transducer was chosen with a view to eliminating the participants' own air-conducted auditory feedback most effectively. However, the occlusion effect caused by the in-ear earphone can influence the bone conduction threshold. The occlusion effect is the result of the acoustic energy created by the vibration of the walls of the external ear canal in response to a bone conducted signal trapped in the ear. When the tip of the earphone is fitted deeper in the ear canal, there is less opportunity for vibrations to occur and the occlusion effect is reduced [30]. This is why the participants were asked to insert the earphones as deeply as possible in the ear canal.

The three-dimensional motion of the upper and lower lips was measured with an optical motion capture system (Qualisys Qqus) at a sampling frequency of 250 Hz. Six low mass, retro-reflective markers with a diameter of 4 mm were placed on the vermilion borders of the upper and lower lips in the midsagittal section, the bridge and the tip of the nose, and the left and right side of the forehead, as shown in Fig. 2.1B. Two digital cameras placed on the left and right in front of the participant emitted infrared light that was reflected from the markers and back to the cameras. The position data of the four markers other than those on the upper and lower lips were used to calculate the relative positions of the lips with respect to the participant's head.

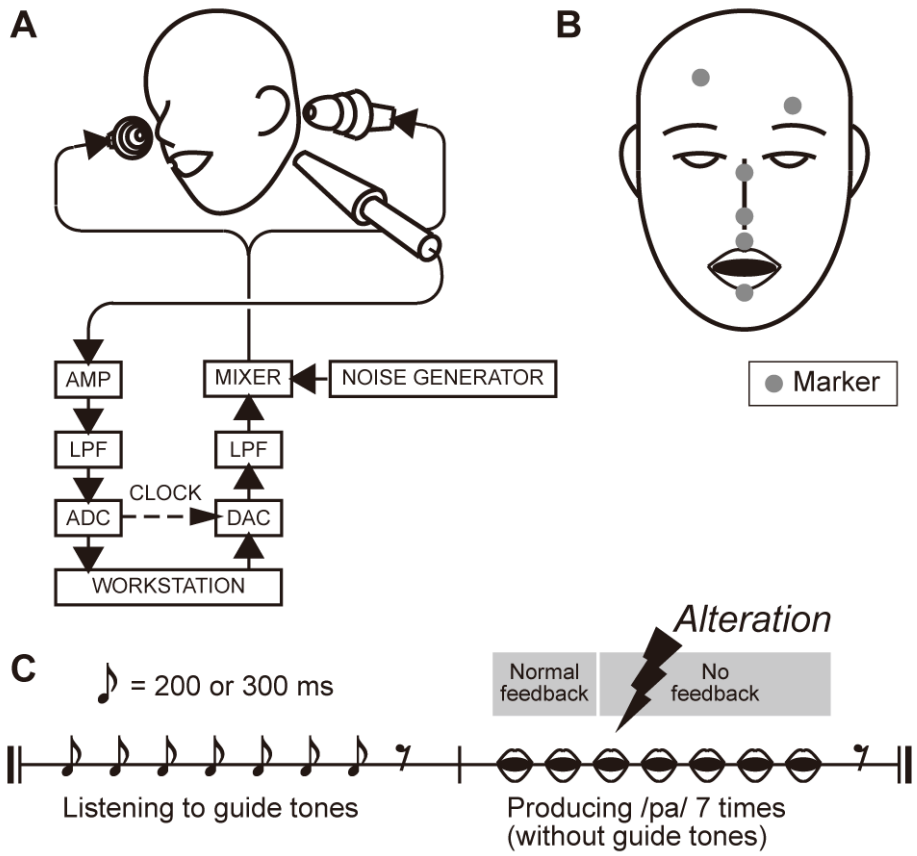

Fig. 2.1. Experimental equipment and protocol.
(A) Schematic diagram of auditory feedback alteration system. See text for details. **(B)** Placement of markers for measuring the three-dimensional motion of the upper and lower lips. Six markers were placed on the vermilion borders of the upper and lower lips in the midsagittal section, the bridge and tip of the nose, and the left and right side of the forehead. **(C)** Schematic diagram of experimental protocol. At the beginning of the trial, the participants heard a sequence of seven click tones with an interval of 200 or 300 ms through earphones. After hearing the final (seventh) click tone, the participants produced syllables at a rate identical to that indicated by the click tone sequence. No click tone was presented during the production period. Participants heard the unaltered speech feedback during the first two repetitions. The normal speech feedback was blocked after the second repetition, and /pa/, /Φa/, or /pi/ sound was presented once at –150, -100, -50, 0, +50, +100, or + 150 ms from the predicted third repetition onset.

2.2.4 Experimental procedures

In each trial in this experiment, the participants were asked to produce an isolated syllable /pa/ seven times while maintaining a constant speech rate. For each trial, the auditory feedback corresponding to the third repetition of /pa/ was altered by shifting the timing and/or replacing the type of syllable, while the subsequent feedback was blocked. A comparison of the articulatory lip movement under each altered condition with that under a normal condition enabled us to evaluate the effect of auditory feedback alteration on speech motor control more precisely than previous studies based on DAF. As for speech errors produced when employing DAF, their speech rate dependence can also be disputed in the light of certain controversial results [22, 23]. Therefore, two speaking rates (200 and 300 ms per syllable) were employed in our experiment in order to examine the speed dependence of the effect.

The experiment consisted of five test blocks and one control block. Each test block consisted of forty-six trials, where twenty-three different feedback conditions were employed for two different repetition rates (200 and 300 ms per syllable). Of the twenty-three feedback conditions, twenty-one were altered conditions where one of three syllables (/pa/, /Φa/, or /pi/) was presented at seven different timings (–150, –100, –50, 0, +50, +100, or +150 ms in relation to the onset of the third repetition), one was a blocked condition (no feedback after the second repetition), and one was unaltered. The control block consisted of twenty trials with unaltered feedback conditions, half of which were conducted at 200 ms per syllable and half at 300 ms per syllable.

In the experiment, the control block was introduced first, which took about 5 minutes, followed by five test blocks, each of which took about 10 minutes. There was a short break between each block. During the test blocks, the order of the feedback conditions applied to each participant was shuffled block by block. In the control block, the two syllable rates were alternated trial by trial.

2.2.5 Tasks

Figure 2.1C depicts the trial protocol. At the beginning of the trial, the participants heard a sequence of seven guide click tones with a fixed interval of 200 or 300 ms through their earphones. After hearing the final (seventh) click tone, the participants were asked to produce syllables at a syllable rate identical to that indicated by the click tone sequence. No click tone was presented during the production period. As illustrated in Fig. 2.1C, the participants heard unaltered speech feedback while producing the first two repetitions. The burst onset timing of the first two repetitions was detected by thresholding the segmental power of the signals calculated every 4 ms. The burst onset timing of

the third repetition was predicted before it was produced, based on the interval between those of the first two repetitions. The normal speech feedback was blocked after the second repetition, and the sound /pa/, /Φa/, or /pi/, spoken by the corresponding participant, was presented once either at –150, –100, –50, 0, +50, +100, or +150 ms from the predicted third repetition onset. These sound stimuli /pa/, /Φa/, and /pi/ were recorded by the participants just before they undertook this task. Note that this method enabled us to investigate not only the effect of speech sound alteration, but also the effect of the early feedback of speech sound, which was impossible to examine using the previously employed online signal modification methods [17-20].

When preparing these stimuli, the participants repeated /pa/, /Φa/, and /pi/ in their most natural way. While the participants were producing these syllables, the burst onset timing of one syllable was detected in the same way as in the experiments, and 200 ms of the signals from the detected onset were stored for each of the three syllables, while preserving the amplitude ratio among the syllables. Examples of the stored syllables for a participant are shown in Fig. 2.2A. When these pre-recorded syllables were presented in the experiments, the sound pressure level was adjusted by the computer program in every trial, based on that of the second repetition, so that the inter-syllabic ratio of the sound pressure level for /pa/, /Φa/, and /pi/ was maintained correctly as each participant produced these syllables in his or her natural way.

Figure 2.2B shows examples of auditory feedback signals presented to a participant under three different conditions during the experiments, while repeating /pa/ seven times at a rate of 300 ms per syllable. In Figs. 2.2Bi-iii, the participant's speech signals are shown in the upper panel, where the thick vertical line indicates the predicted onset of the third repetition. The corresponding auditory feedback signals are shown in the lower panel, where the thick vertical line indicates the onset of the altered auditory feedback signal. The auditory stimuli presented in Figs. 2.2Bi-iii were /pa/ at –100 ms, /Φa/ at 0 ms and /pi/ at +50 ms from the predicted onset of the third repetition, respectively. The prediction error of the onset timing of the third repetition was at most 20 ms in the posthoc analyses of the results of trials performed under the unaltered auditory feedback condition.

2.2.6 Data analysis

The time varying three-dimensional labial distance (LD) was calculated from the marker position data. For each participant, the LD trajectories of all trials were temporally aligned at the predicted third repetition onset by referring to the simultaneously recorded acoustic signals. The mean LD trajectory of five trials was obtained for each of forty-six different conditions in the five test

blocks (twenty-three feedback types, two speech rates). The mean trajectory of ten trials from the control (normal feedback) block was also obtained for the two speech rates.

The auditorily induced change in the labial movement was represented by a lag that provided the maximum cross-correlation between the LD trajectories under the altered and control conditions within the post-stimulus period. Note that this method was more stable and consistent than that using the displacement error or the velocity error, maybe because of the inter-participant variability in the time course of lip opening-closing cycle. In Fig. 2.3, the solid and dotted curves in the figure indicate the mean LD trajectories under the altered and control conditions, respectively. (The bottoms of curves within an opening-closing cycle correspond to the instant of bilabial closure.) The thick vertical line indicates the onset timing of the auditory stimulus, while the dotted vertical line indicates the predicted third repetition onset. The beginning of the post-stimulus period was set at 120 ms after the stimulus onset, based on the fact that the short latency auditory-vocal response has a latency ranging from 100 to 150 ms [18]. A 200 (300) ms period was chosen for a speech rate of 200 (300) ms per syllable. The cross-correlation function $R_{post}(m)$ of the lag m was represented by

$$R_{post}(m) = \sum_{n=0}^{N-1-m} LD_{ctrl}(n) \cdot LD_{alt}(n+m) / (N - |m|),$$

where $LD_{ctrl}(n)$ and $LD_{alt}(n)$ were the LDs at n under the control and altered conditions, respectively. Each LD trajectory was unbiased and windowed by a Blackman window to reduce the boundary effects. The lag that provided the maximum cross-correlation was represented as $\arg\max R_{post}(m)$. An ahead-of-time shift of the movement caused by an altered auditory feedback resulted in a minus lag value m, and vice versa.

To adjust for the phase difference between the trajectories of the altered and control conditions before alteration onset, the lag within the pre-stimulus period $\arg\max R_{pre}(m)$ (also shown in Fig. 2.3) was calculated and subtracted from $\arg\max R_{post}(m)$. The pre-stimulus period was set at the same length as the post-stimulus period. The cross-correlation function $R_{pre}(m)$ was calculated in

the same way as $R_{post}(m)$.

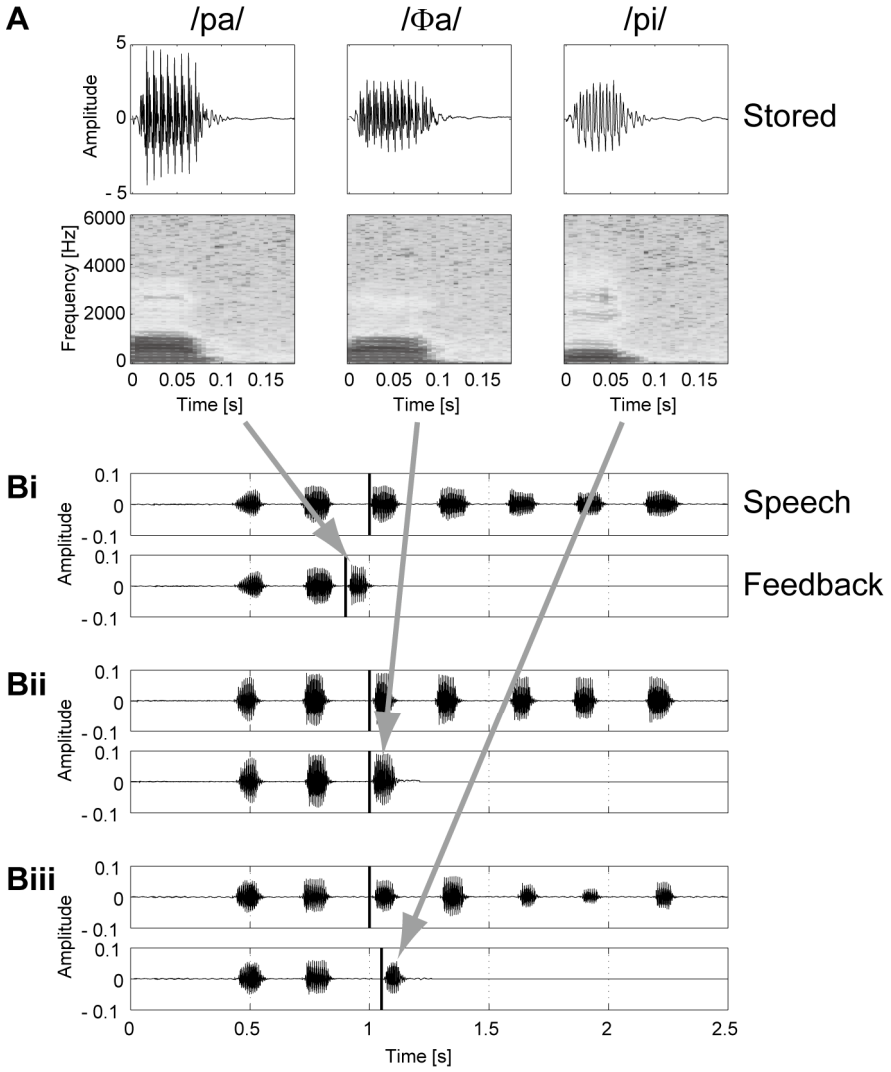

Fig. 2.2. Examples of acoustic signals.
(A) Examples of syllables stored for a participant when preparing sound stimuli. While the participant was producing /pa/, /Φa/, or /pi/ repeatedly, the burst onset timing of one syllable was detected in the same way as in the experiments, and 200 ms of the signals from the detected onset were stored while preserving the amplitude ratio among the syllables. When presenting these pre-recorded syllables in the experiments, the sound pressure level was adjusted by the computer program for every trial, based on that of the second repetition, so that the inter-syllabic ratio of the sound pressure level for /pa/, /Φa/, and /pi/ was maintained correctly as the participant produced the syllables in his or her natural way. **(B)** Examples of auditory feedback signals presented to a participant under three different conditions during the experiments, while he or she produced /pa/ seven times at a rate of 300 ms per syllable. In each pair of panels, Bi to Biii, the participant's speech signals are illustrated at the top, with

the thick vertical line indicating the predicted onset of the third repetition. The corresponding auditory feedback signals are in the lower panels in Bi – Biii, with the thick vertical line indicating the onset of the altered auditory feedback signal. The auditory stimuli presented in Bi, Bii and Biii were /pa/ at –100 ms, /Φa/ at 0 ms and /pi/ at +50 ms from the predicted onset of the third repetition, respectively.

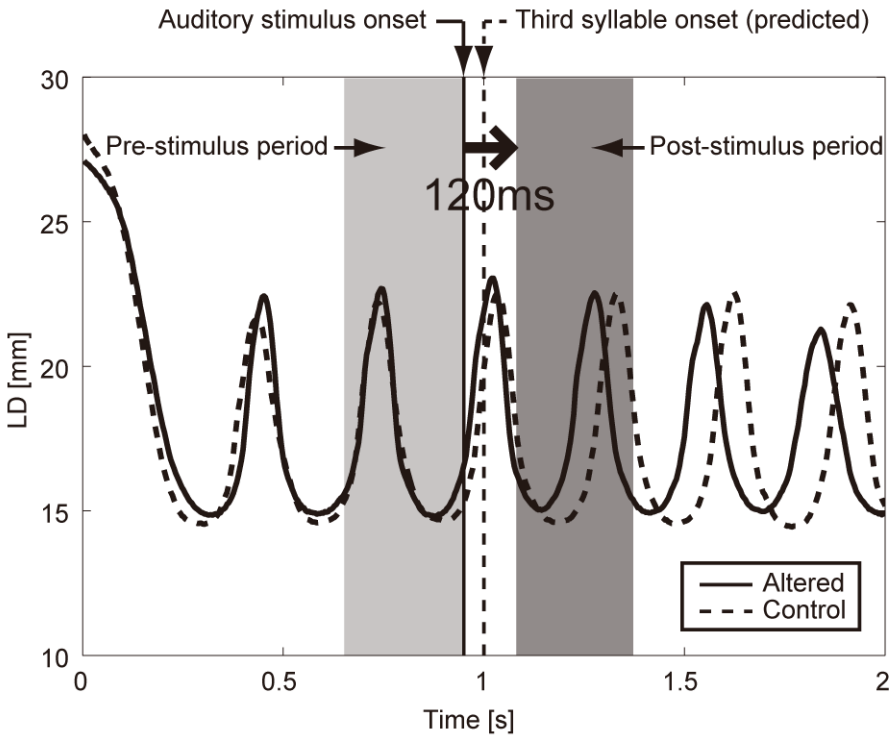

Fig. 2.3. Definition of pre- and post-stimulus periods.
The solid and dotted curves indicate the labial distance (LD) trajectories under the altered and control conditions, respectively. The thick vertical line indicates the onset timing of the auditory stimulus, while the dotted vertical line indicates the predicted third repetition onset. The bottom of curves within an opening-closing cycle corresponds to the instant of bilabial closure. The pre- and post-stimulus periods are highlighted by the light and dark gray rectangles, respectively. The lengths of the pre- and post-stimulus periods were identical to the syllable interval, i.e., 200 ms for a speech rate of 200 ms per syllable, and 300 ms for a speech rate of 300 ms per syllable. The top of the post-stimulus period was set at 120 ms after the onset timing of the auditory stimulus. The differences between two LD trajectories in each of the pre- and post-stimulus periods were calculated as the lags that provided the maximum cross-correlation between the two trajectories. The minus (plus) value of the lag corresponded to the ahead-of-time (delayed) shift of the trajectory caused by the auditory feedback alteration. See text for details.

2.3 Results

2.3.1 Labial distance trajectory

Figure 2.4 shows sample LD trajectory data during the production of /pa/ at a speech rate of 300 ms per syllable. The auditory feedback conditions shown from the top to bottom panels were as follows: pre-recorded /pa/ was presented once at –150, –100, –50, 0, 50, 100, 150 ms from the predicted third repetition onset. The solid vertical line in each panel indicates the onset timing of the auditory stimulus, while the dotted vertical line indicates the predicted third repetition onset. The solid curve in each panel shows the mean LD trajectory for five trials over the test blocks. The mean trajectory for ten trials in the control (normal feedback condition) block is shown as a dotted curve.

By comparing the two trajectories in each panel, the mouth opening movement subsequent to the auditory stimulus onset appeared generally to occur sooner for the –50 ms stimulus presentation. While a similar hasty movement was also observed for the -150 and -100 ms conditions, the effect seemed to be weaker. The deviation between the trajectories under each of the delayed feedback (50, 100, 150 ms) and control conditions was much smaller. Similar results were obtained for all ten participants.

In Fig. 2.4, the open and filled horizontal bars in each panel indicate the pre- and post-stimulus periods, respectively, for calculating the lag of the maximum cross-correlation between the LD trajectories under the altered and control conditions. The lag value may not necessarily reflect the exact amount of time shift, but will at least help to indicate which of the two series is leading the other, irrespective of which component of the amplitude, period, or phase of the LD trajectories was dominant in the difference. As observed in the top three panels in Fig. 2.4, the difference between the LD trajectories in the altered and control conditions tended to increase with time after the auditory alternation onset. Since such differences may be produced by a progressive accumulation of voluntary and involuntary effects, it would be difficult to specify the direct causal effect of auditory alteration on the LD trajectory. Therefore, this study focused on the LD trajectory during a short period after the auditory alteration. The following subsection presents a statistical evaluation of the differences between LD trajectories under each of altered and control conditions.

2.3.2 Auditorily induced rapid change in articulatory movement

Figure 2.5 shows the lag corresponding to the maximum cross-correlation (N = 10; error bar: standard error) between the LD trajectories under the altered and control conditions within the post-stimulus period, obtained by subtracting those within the pre-stimulus period. The minus value of the lag reflects an

ahead-of-time shift of the articulatory lip movement compared with the control, and vice versa. The top and bottom panels show the results obtained when the speech rates were 200 and 300 ms per syllable, respectively. Each color indicates the syllable presented as a stimulus. "No" indicates a condition where no feedback was presented after the production of the second repetition. The condition indicated as "normal" refers to a comparison of the normal feedback trials during the test blocks and those in the control block, which reflects the variance in the baseline speech rate of each participant throughout the experiment.

For 22 altered conditions at each speech rate, the statistical significance of the difference from the "normal" condition was evaluated with a two-sided paired t-test (dF = 9 for all comparisons, with the Bonferroni adjustment). A statistically significant change ($p < 0.05$) was found only when syllable /pa/ was presented 50 ms prior to the onset of syllable production for a rate of 300 ms per syllable. Under this condition, the auditory feedback alteration resulted in an ahead-of-time shift of the articulatory lip movement according to Fig. 2.5 (a minus lag value). A comparable large negative mean value was also found in Fig. 2.5 with a 50 ms preceding presentation of syllable /Φa/ at a rate of 300 ms per syllable. However, the difference from the normal condition was not statistically significant ($p > 0.05$) owing to the variation across subjects. Also from Fig. 2.5, the maximum positive mean values were found for a 50 ms delayed presentation of syllables /pa/ and /Φa/ at a rate of 300 ms per syllable. However, these were also statistically insignificant ($p > 0.05$). For a speech rate of 200 ms per syllable, the effects of auditory feedback alteration on the articulatory lip movement were found to be insignificant under all the conditions tested ($p > 0.05$).

From the experimental results, it was concluded that an ahead-of-time shift in the articulatory lip movement emerged rapidly when the auditory feedback preceded the real syllable production by 50 ms. However, too early a manipulation (–150 and –100 ms) of the auditory feedback did not significantly affect the subsequent articulatory lip movement. The delayed feedback also produced no significant change. Syllables that were not identical to those of the speech task (/Φa/ and /pi/) had no significant effect even when they were fed back 50 ms prior to the real syllable production. Finally, the articulatory changes were not significant for the faster speech rate (200 ms per syllable) under any of the alteration conditions tested.

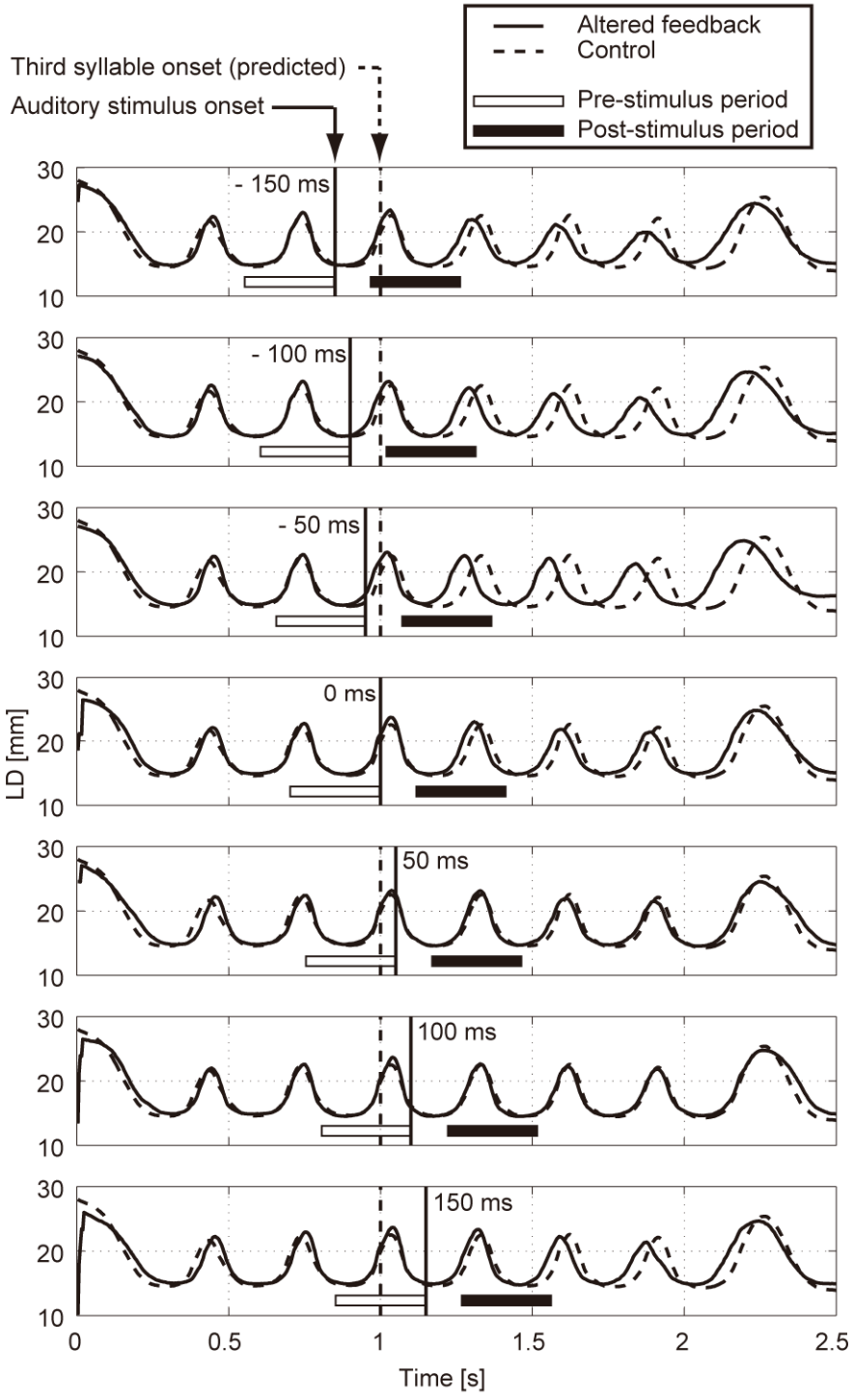

Fig. 2.4. Labial distance (LD) trajectories of a participant while producing /pa/ at 300 ms per syllable.

The auditory feedback conditions in each panel from the top to bottom were as follows: pre-recorded /pa/ was presented once at -150, -100, -50, 0, 50, 100, 150 ms from the predicted third repetition onset. The thick vertical line in each

panel indicates the onset timing of the auditory stimulus, while the dotted vertical line indicates the predicted third repetition onset. The solid curve in each panel shows the mean LD trajectory of five trials over the test blocks. The mean trajectory of ten trials in the control (normal feedback condition) block is shown as a dotted curve. The white and black horizontal bars in each panel indicate the pre- and post-stimulus periods, respectively, for calculating the lag of the maximum cross-correlation between the LD trajectories under the altered and control conditions.

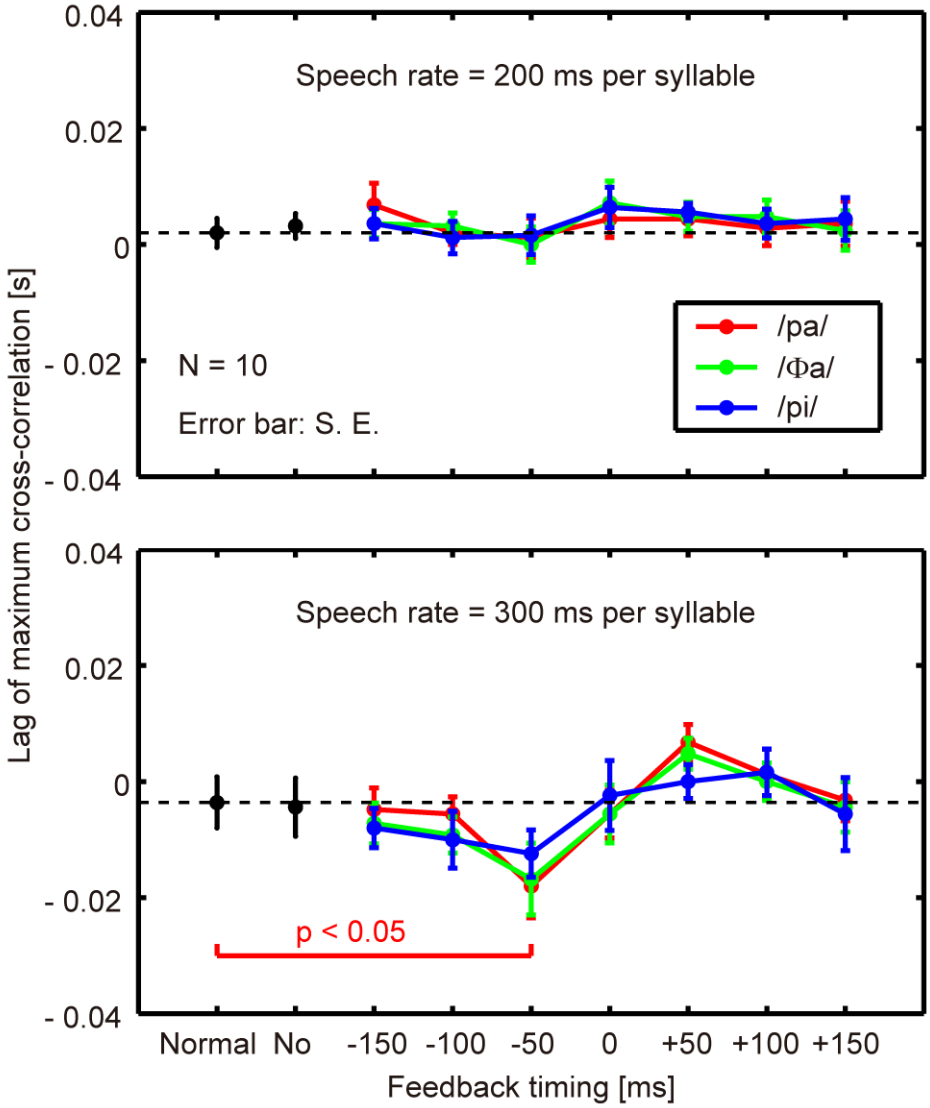

Fig. 2.5. Lag of maximum cross-correlation (N = 10; error bar: standard error). Top: speech rate of 200 ms per syllable, bottom: speech rate of 300 ms per syllable. "Normal": comparison of the normal feedback trials during the test blocks and those in the control block. "No": the auditory feedback was interrupted after producing the second repetition. Other indices from "–150" through "+150" indicate the onset timing of the auditory stimulus against the predicted third repetition onset. The legends /pa/, /Φa/, and /pi/ show the syllable presented auditorily to the participants. The statistical difference between the values obtained under each altered feedback condition and those obtained under a "normal" condition was evaluated with a two-sided paired t-test.

2.4 Discussion

2.4.1 Time-asymmetric effect of auditory feedback alteration

The experimental results obtained in the current study showed that the ahead-of-time and delayed auditory feedback affected the articulatory lip movement in a time-asymmetric manner during repetitive syllable production. Significantly hastened articulation at around 120 ms from the auditory alteration onset occurred when the auditory stimulus was presented 50 ms prior to the onset of syllable production. Taken together with the hypothetical feedforward and feedback mechanisms of speech motor control [31], the hastened articulation could be regarded as a sort of rapid compensatory articulation in the time domain, which was induced by a sensory error caused by the progressive auditory input. However, the feedback alteration effect was not significant when the feedback timing was earlier (–150 and –100 ms). This fact seemed to suggest that an internal simulation of the auditory consequences of speech motor commands is not completed 100 ms prior to the initiation of the articulatory lip movement.

More interestingly, our experimental result revealed that no delayed feedback had a significant effect on the subsequent lip movement. One possible explanation for this result may be an imperfect masking of the air- and bone-conducted auditory feedback. In our experiment, an in-ear earphone was used to realize the effective isolation of the air-conducted feedback of the participants' own speech output. In addition, a masking noise was delivered to their ears to disturb the sensation and/or perception of the air- and bone-conducted feedback to a certain degree. However, even a small amount of natural feedback might still reduce the effect of sensory error on the motor control compared with ahead-of-time feedback alteration. This might result in the insufficient effect of the delayed auditory feedback.

Another possible mechanism for the temporally asymmetric effect could be related to the response attenuation in the auditory cortex resulting from self-produced vocalization [6-9]. The precise temporal processing properties of such auditory attenuation on the time course of speech production, however, are less well understood. Further experimental and theoretical investigations are required to clarify the precise mechanisms underlying the time-asymmetric effect of auditory feedback alteration on the speech articulatory movement obtained in our experiment.

2.4.2 Context dependence of auditorily-induced response

The experimental results showed that the auditory feedback of /Φa/ and /pi/ did not significantly change the subsequent lip movement, irrespective of the timing of the feedback. Taking this fact together with the hypothetical

feedback-feedforward error correction mechanism [31], articulatory compensation in the time domain might be considered rather insensitive to an auditory input whose acoustic feature is not identical to that of the internal prediction.

The results also revealed of the effect of /Φa/ had a larger mean value than that of /pi/ being fed back 50 ms prior to /pa/ production at a rate of 300 ms per syllable, though both were statistically insignificant. One suspected cause is that /pi/ has a smaller relative acoustical power than /Φa/. In the experiment, the auditory feedback amplitude of each syllable was dynamically adjusted so that its syllabic power ratio to the syllable /pa/ to be produced by each participant was matched with that in his/her natural production. (See the Task subsection for details.) Figure 2.6 shows the relationship between the relative syllabic power of the auditory feedback and the difference in the magnitude of auditorily-induced articulatory change on a participant-by-participant basis (N = 10). If the magnitude of the articulatory change were dependent on the power of the auditory feedback, the data in Fig. 2.6 would exhibit a negative correlation. However, the correlation coefficient for ten participants was found to be low ($r = 0.54$, $p = 0.11$, $dF = 8$). Therefore, the smaller mean value of the effect of /pi/ feedback did not appear to result from its smaller amplitude.

Another possible cause of the smaller mean effect of the /pi/ feedback could be related to a larger acoustic deviation of /pi/ from /pa/ compared with that of /Φa/, in the light of the evidence showing that the auditory cortex responded differently to self-produced and externally produced speech sounds during speech production [9]. The auditory input of /pi/ while producing /pa/ might not be processed as a self-produced sound because of the large difference in vowel quality between /a/ and /i/ such as the location of the formants, despite the invariant feature of the initial /p/ independent of the following vowel [32].

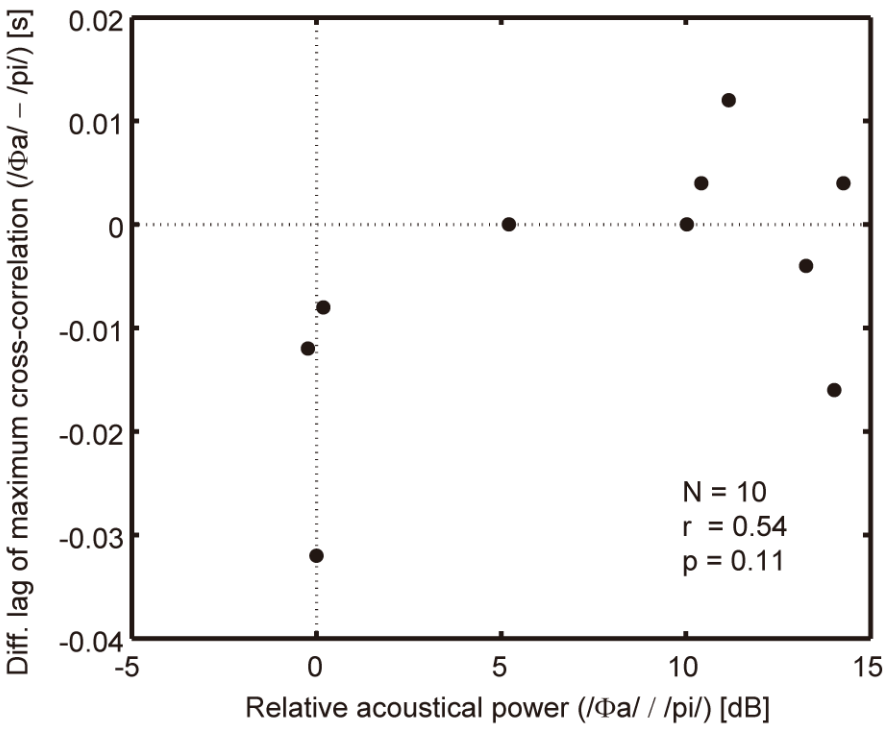

Fig. 2.6. Magnitude difference in auditory-induced articulatory change against relative acoustical power of auditory feedback.
The abscissa is the relative acoustical power between /Φa/ and /pi/. The ordinate is the difference between the mean lag shown in Fig. 2.5 for /Φa/ and /pi/ feedback 50 ms prior to the production onset under a rate of 300 ms per syllable. The correlation coefficient for ten participants was r = 0.54 (p = 0.11, dF = 8).

2.4.3 Speech rate dependency of response

The experimental result showed that none of the altered auditory feedback tested under the faster speech condition (200 ms per syllable) induced significant articulatory changes. So far little has been reported about the dependence of the auditory alteration effect on speech rate. There have been conflicting results regarding the speech-rate dependence of DAF-induced disfluencies, where speech errors were found to decrease [23] or increase [22] as the speaking rate increased. Most of the speech errors both the above studies involved various suprasyllabic features, which may not be a direct consequence of the short-latency auditory-motor response as obtained in our experiment. Further investigation is required to untangle the sources of the complex speech errors induced by DAF, and to understand the mechanism underlying the speech-rate dependence of the auditory-motor response.

A study on the accuracy with which speakers repeat a monosyllable in time with an external rhythm suggested two underlying processes depending on the repetition rate [33]. At a rate of 1 to 3 times per second, speakers could compensate for a discrepancy in timing between a syllable and the external guide tone in an adjacent or neighboring utterance ("ongoing processing"), while at a rate of 4 to 6 times per second, such one-by-one processing did not work ("holistic processing"). Considering our experimental condition in the light of Hibi's work, a rate of 200 ms per syllable is classified as holistic processing where the one-by-one adjustment of utterances was impossible. On the other hand, a rate of 300 ms per syllable (equivalent to 3.3 times per second) can be classified as either ongoing or holistic processing. Such a difference in the underlying processing strategy might have caused the speech rate dependence of the auditory-motor response obtained in our experiment. However, the speech task used in our experiment was very different from that used in Hibi's work in that the participants were required to repeat the syllable in a self-paced manner with no external rhythm provided while speaking. Another processing mechanism may be involved in the self-paced rhythmic production.

2.4.4 Language dependency of response

From the viewpoint of rhythmic properties, languages are considered to be classified as stress-, syllable-, or mora-timed, although a quantitative measure of speech rhythm has not been established. While the results of the current study were obtained from Japanese speakers, it would also be interesting to consider whether the same behavior occurs in other language speakers. Language-specific aspects of temporal organization of the kinematics of lower lip-jaw articulation have been compared between English, French, and Japanese,

which are assumed to be examples of stress-, syllable-, and mora-timed languages, respectively[34]. They have used reiterant speech task in which speakers were required to replace each syllable of a target phrase with a single syllable such as /ba/ or /ma/, while trying to maintain the rhythmic character of the original[35]. They have found highly linear relation between peak velocity and displacement in lower lip movement for all three languages, and concluded that the dynamics could be modeled as a universal second-order system with language-specific parameter settings. It is therefore inferred that, as far as the repetitive syllable production task is concerned, the auditory-motor effect observed in the current study would be expected to occur also in speakers other than Japanese.

2.5 References

1. Guenther, F.H., *Cortical interactions underlying the production of speech sounds.* J Commun Disord, 2006. **39**(5): p. 350-65.
2. Houde, J.F. and M.I. Jordan, *Sensorimotor adaptation in speech production.* Science, 1998. **279**(5354): p. 1213-1216.
3. Villacorta, V.M., J.S. Perkell, and F.H. Guenther, *Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception.* J Acoust Soc Am, 2007. **122**(4): p. 2306-19.
4. Jones, J.A. and K.G. Munhall, *Remapping auditory-motor representations in voice production.* Curr Biol, 2005. **15**(19): p. 1768-72.
5. Burnett, T.A. and C.R. Larson, *Early pitch-shift response is active in both steady and dynamic voice pitch control.* J Acoust Soc Am, 2002. **112**(3 Pt 1): p. 1058-63.
6. Curio, G., et al., *Speaking modifies voice-evoked activity in the human auditory cortex.* Hum Brain Mapp, 2000. **9**(4): p. 183-91.
7. Eliades, S.J. and X. Wang, *Dynamics of auditory-vocal interaction in monkey auditory cortex.* Cereb Cortex, 2005. **15**(10): p. 1510-23.
8. Heinks-Maldonado, T.H., et al., *Fine-tuning of auditory cortex during speech production.* Psychophysiology, 2005. **42**(2): p. 180-90.
9. Houde, J.F., et al., *Modulation of the auditory cortex during speech: an MEG study.* J Cogn Neurosci, 2002. **14**(8): p. 1125-38.
10. Callan, D.E., et al., *An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system.* J Speech Lang Hear Res, 2000. **43**(3): p. 721-36.
11. Guenther, F.H., M. Hampson, and D. Johnson, *A theoretical investigation of reference frames for the planning of speech movements.* Psychol Rev, 1998. **105**(4): p. 611-33.
12. Borden, G.J., *An interpretation of research of feedback interruption in speech.* Brain Lang, 1979. **7**(3): p. 307-19.
13. Postma, A., *Detection of errors during speech production: a review of speech monitoring models.* Cognition, 2000. **77**(2): p. 97-132.
14. Lombard, E., *Le signe de l'elevation de la voix.* Annales maladies oreille larynx nez pharynx 1911. **37**: p. 101-119.
15. Nonaka, S., et al., *Lombard reflex during PAG-induced vocalization in decerebrate cats.* Neuroscience Research, 1997. **29**(4): p. 283-289.
16. Sapir, S., M.D. McClean, and C.R. Larson, *Human laryngeal responses to auditory stimulation.* J Acoust Soc Am, 1983. **73**(1): p. 315-21.
17. Kawahara, H. and J.C. Williams, *Effects of auditory feedback on voice pitch trajectories: characteristic responses to pitch perturbations*, in *Vocal Fold Physiology: Controlling Complexity and Chaos*, P.J. Davis and N.H. Fletcher,

Editors. 1996, Singular Publishing Group, Inc.: San Diego. p. 263-278.
18. Burnett, T.A., et al., *Voice F0 responses to manipulations in pitch feedback.* J Acoust Soc Am, 1998. **103**(6): p. 3153-61.
19. Lee, B.S., *Effects of delayed speech feedback.* Journal of the Acoustical Society of America, 1950. **22**(6): p. 824-826.
20. Tiffany, W.R. and C.N. Hanley, *Delayed speech feedback as a test for auditory malingering.* Science, 1952. **115**(2977): p. 59-60.
21. Fairbanks, G., *Selective vocal effects of delayed auditory feedback.* Journal of speech and hearing disorders, 1955. **20**(4): p. 333-346.
22. Stuart, A., et al., *Effect of delayed auditory feedback on normal speakers at two speech rates.* Journal of the Acoustical Society of America, 2002. **111**(5): p. 2237-2241.
23. Zanini, S., et al., *Speaking speed effects on delayed auditory feedback disruption of speech fluency.* Percept Mot Skills, 1999. **89**(3 Pt 2): p. 1095-109.
24. Peters, R.W., *The effect of changes in side-tone delay and level upon rate of oral reading of normal speakers.* J Speech Hear Disord, 1954. **19**(4): p. 483-90.
25. Abbs, J.H. and V.L. Gracco, *Control of complex motor gestures: orofacial muscle responses to load perturbations of lip during speech.* Journal of Neurophysiology, 1984. **51**(4): p. 705-723.
26. Gracco, V.L. and J.H. Abbs, *Dynamic control of the perioral system during speech: kinematic analyses of autogenic and nonautogenic sensorimotor processes.* Journal of Neurophysiology, 1985. **54**(2): p. 418-432.
27. Gracco, V.L. and J.H. Abbs, *Central patterning of speech movements.* Exp Brain Res, 1988. **71**(3): p. 515-26.
28. Saltzman, E., et al., *Dynamics of intergestural timing: a perturbation study of lip-larynx coordination.* Exp Brain Res, 1998. **123**(4): p. 412-24.
29. Kelso, J.A., et al., *Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures.* J Exp Psychol Hum Percept Perform, 1984. **10**(6): p. 812-32.
30. Mueller, H.G., *CIC Hearing Aids: What Is Their Impact On The Occlusion Effect.* The Hearing Journal, 1994. **47**(11): p. 29-35.
31. Tourville, J.A., K.J. Reilly, and F.H. Guenther, *Neural mechanisms underlying auditory feedback control of speech.* Neuroimage, 2008. **39**(3): p. 1429-43.
32. Blumstein, S.E. and K.N. Stevens, *Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants.* J Acoust Soc Am, 1979. **66**(4): p. 1001-17.
33. Hibi, S., *Rhythm perception in repetitive sound sequence.* J Acoust Soc Jpn, 1983. **(E)4**(2): p. 83-95.
34. Vatikiotis-Bateson, E. and J.A.S. Kelso, *Rhythm type and articulatory*

dynamics in English, French and Japanese. Journal of Phonetics, 1993. **21**: p. 231-265.

35. Kelso, J.A., et al., *A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modeling.* J Acoust Soc Am, 1985. **77**(1): p. 266-80.

3. Psychophysical study of the effect of articulatory movement information on phoneme perception

3.1 Background

From both theoretical and experimental points of view, speech perception has long been thought to be linked with the speech motor control [1-3]. Recent neurophysiological and neuroimaging studies have reported that phoneme perception activates motor-related neural circuits in the brain, which are invoked with the production of the same phoneme [4-6]. These studies have suggested that motor areas somatotopically related to the individual speech organs, such as the lips and the tongue, can be coactivated in a phoneme-dependent manner during speech perception [7].

Speech perception is also thought to involve audio-visual multisensory integration process. It is well known that an illusory phoneme perception is produced when auditory information is accompanied by simultaneous visual lip movement that does not match the auditory information (e.g., the McGurk effect [8]). For example, an auditory [ba] combined with a visual [ga] is typically heard as [da]. Recently, Sams et al. have demonstrated that the perception of auditory phonemes can also be disturbed when the listener silently articulates incongruent phonemes [9]. These lines of evidence naturally lead to an interesting question of whether both the visual speech motion and the self articulatory movement act in a same manner, or even on an identical pathway, within the neural circuits responsible for speech perception. Indeed, it has been demonstrated that seeing speech-related lip movements as well as hearing speech sounds modulated the excitability of cortical representation of the lip muscles [10]. However, little has been studied about how the self articulatory movements of each speech organ can affect speech perception [11].

In the current study, the author hypothesized that phoneme perception depends on the listener's own articulatory movement of each speech organ (the lips and the tongue) as well as on the visually presented speech motion. To test this hypothesis, the author examined whether phoneme intelligibility for [p], [t], and [k] is affected by whispering concordant/discordant phonemes and by watching videos of a model speaker producing concordant/discordant phonemes. The crucial organ for articulating the phoneme [p] is the lips, whereas the tongue is crucial to the articulation of the remaining two phonemes ([t] and [k]). The correct perception rate for each phoneme revealed the following results. Viewing lip motion ([p]) degraded hearing of phonemes produced by the tongue ([t] and [k]), and viewing the tongue motions degraded hearing of the phoneme produced by the lips, replicating the previous studies

[9]. The critical finding of the present study is that the silent articulation affected the phoneme perception in a different way from the visual effects. While articulating phoneme with the lips did not affect hearing of the tongue-related phonemes and articulating phonemes with the tongue did not affect hearing of the lip related phoneme, more interestingly, articulating phonemes with the tongue degraded hearing of the different phonemes produced by the tongue. Namely, articulating [t] degraded hearing of [k] and vice versa. These results suggest that motor interferential effect on speech perception is mediated by a different mechanism from the visual one: the auditory-visual integration occurs across different speech organs, whereas the auditory-articulatory integration occurs within the same organ.

3.2 Materials and methods

3.2.1 Ethics Statement

All the participants gave their written informed consent to participating in this study, which was approved by the Research Ethics Board of NTT Communication Science Laboratories.

3.2.2 Participants

Ten healthy adults (four males) aged 18 to 40 years (mean age 26.3 +/- 7.5 years) participated in the experiment. All the participants were native speakers of Japanese and exhibited no obvious speech difficulties as judged by the experimenters.

3.2.3 Tasks

In each trial in the experiment, participants were auditorily presented with one of seven Japanese single syllables, [pa], [ta], [ka], [ba], [da], [ga] and [a], all spoken by a model speaker, and asked to identify the syllable they heard by pressing one of seven labeled buttons. Three different conditions were administered to each participant in separate blocks: (1) hearing syllables with seeing a blank screen (auditory condition), (2) hearing syllables with whispering concordant/discordant syllables (motor condition), and (3) hearing syllables with seeing videos of a model speaker's face producing concordant/discordant syllables (visual condition).

The auditory stimuli were presented to participants via a headphone (Sennheizer HD280Pro) at a level of 60 dBSPL, being embedded in white noise in order to exclude the possibility of participants hearing their own whispered voice in the motor condition. The signal-to-noise ratio was 5 dB. The beginning and the end of the noise were faded in and out by 0.5 s, respectively. The auditory stimuli were preceded by four clicks (interclick interval of 0.67 s) as shown in Fig. 3.1, which indicated to participants the timing to whisper a syllable in the motor condition. In each trial in the motor condition, one of the four syllables, [pa], [ta], [ka], or [a], was visually presented in Japanese characters at the onset of the white noise and disappeared at the second click. Participants were asked to whisper the indicated syllable three times while seeing a blank screen, in time with the third and fourth clicks and the onset of the auditory stimulus, as shown in Fig. 3.1. In each trial in the visual condition, a video of a model speaker's face producing one of the three syllables, [pa], [ta], or [ka], was presented synchronously with the auditory stimulus. The onset of the auditory stimulus was aligned to the onset of the syllable in the audio track associated with the video. (The audio track was not presented to participants.) Prior to the video presentation, the initial frame of the video was presented at

the onset of the white noise and remained until the onset of the auditory stimulus, as shown in Fig. 3.1.

3.2.4 Experimental procedures

The experimental session for each participant consisted of two blocks of familiarization followed by 11 blocks of test, as shown in Fig. 3.2. In the familiarization phase, all participants performed one auditory condition block and then one motor condition block. In the test phase, five sets of one motor condition block and one visual condition block were performed, with the order of two blocks within each set randomized and counterbalanced for each participant. One auditory condition block was performed at the end of the test phase for all participants. During the experimental session, participants took short breaks between blocks. Each trial was initiated when participants entered their response.

Each five motor condition block consisted of 84 trials in which each of the 28 different trials (7 stimuli x 4 subtasks) were performed three times. Each five visual condition block consisted of 63 trials in which each of the 21 different trials (7 stimuli x 3 subtasks) were performed three times. The auditory condition block consisted of 105 trials in which each of the seven stimuli were presented 15 times. (See Table 3.1 for an overview.) The order of the trials in each block was randomized for each participant.

3.2.5 Data analysis

For each auditory stimulus under each subtask condition, 15 responses were collected per participant from which the correct response rates for [pa], [ta], and [ka] were calculated. Adopting the alternative choice of the seven auditory stimuli ([pa], [ta], [ka], [ba], [da], [ga], and [a]) rather than of the three ([pa], [ta], and [ka]) would allow for fair perception of the three phonemes to be examined.

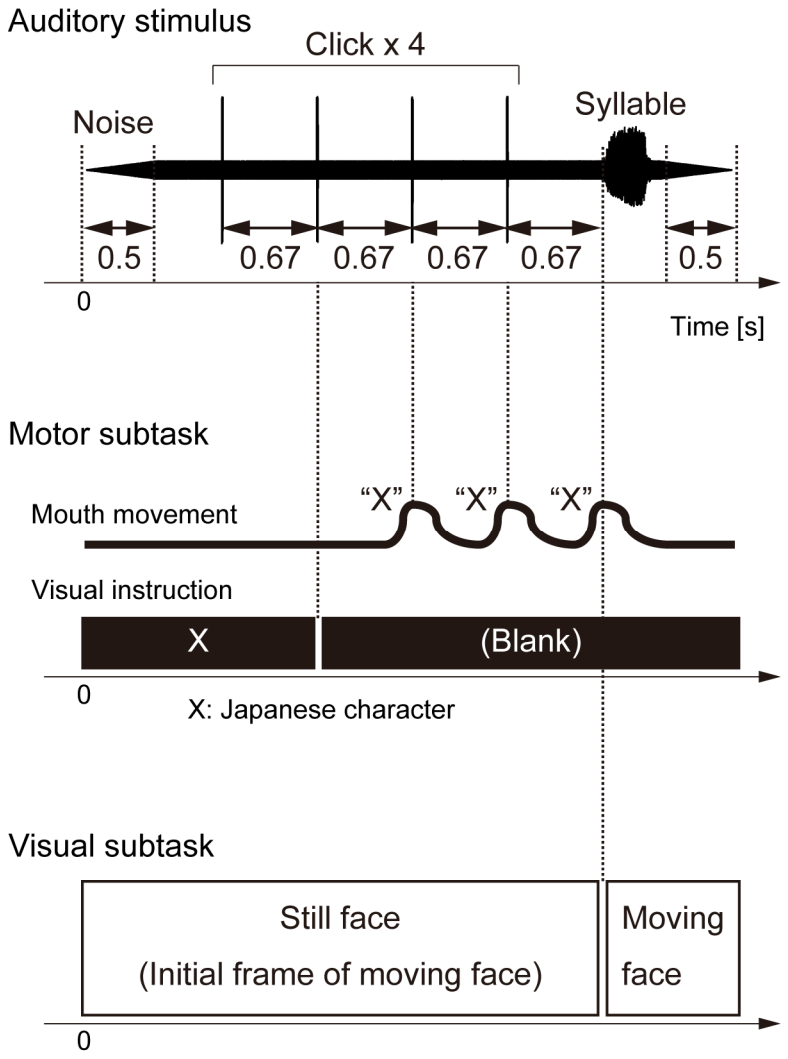

Fig. 3.1. Auditory stimulus and subtasks.
Auditory stimulus was embedded in white noise in order to exclude possibility of participants hearing own whispered voice in the motor condition. The signal-to-noise ratio was set to 5 dB. The noise was faded in and out linearly over 0.5 s. The stimulus was preceded by four clicks spaced by 0.67 s, which indicated to participants the timing to whisper a syllable in the motor condition. In the motor condition, syllables to be whispered by participants were visually presented in Japanese characters first and disappeared at the second click. Participants whispered syllables three times in time with the third and fourth clicks and the onset of the stimulus. In the visual condition, videos of a model speaker's face producing syllables were presented, synchronized with the auditory stimuli. The initial frame of each video was presented from the noise onset to the stimulus onset.

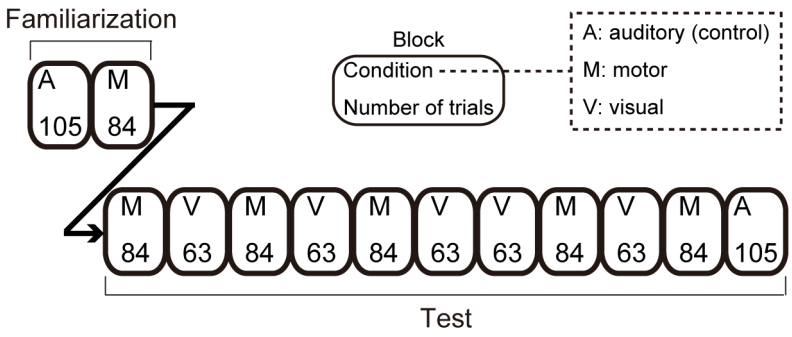

Fig. 3.2. Experimental sequence.
Participants performed two familiarization blocks prior to test phase, one under auditory condition and one under motor condition. In the test phase, five pairs of one motor and one visual condition blocks were performed where the order of two blocks within each pair were randomized and counterbalanced. One auditory condition block was performed at the end the test phase. Participants took a short break between blocks. See Table 3.1 for details of trials performed in each block.

Condition	Auditory stimulus type	Subtask type	Number of trials in a block	Num. of blocks in a session
Auditory (control)	7 ([pa][ta][ka] [ba][da][ga] [a])	n.a.	105 (7 stimuli x 15 trials)	1
Motor		4 ([pa][ta][ka][a])	84 (7 stimuli x 4 subtasks x 3 trials)	5
Visual		3 ([pa][ta][ka])	63 (7 stimuli x 3 subtasks x 3 trials)	5

Table 3.1. Overview of blocks in the experiment.
In the auditory (control) condition block, the seven stimuli were presented 15 times in a randomized order. In each five motor condition blocks, 28 different combinations of stimulus and subtask (whispering) were performed 3 times in a randomized order. In each five visual condition blocks, 21 different combinations of stimulus and subtask (seeing video) were performed 3 times in a randomized order. The order of blocks in the experiment was shown in Fig. 3.2.

3.3 Results

3.3.1 Concordant case

Influence of the participants' own articulatory movements and the visually presented model speaker's mouth motion on their perception of congruent phonemes was assessed. Correct response rates for [p], [t], and [k] under auditory (control), motor, and visual conditions were shown in Fig. 3.3. The mean rates under control condition indicated that perception of the auditory stimuli was not perfect because of the background noise (0.720, 0.993, and 0.747 for [p], [t], and [k], respectively). In the case of [p] and [k], the mean correct response rates under the motor and visual conditions were both higher as compared to the control. Although that was not the case for [t] because the rate under the control condition was close to the ceiling, the rates under the motor and visual conditions were at least not inferior to the control. A one-way repeated measures ANOVA was conducted for each auditory stimulus to examine whether the correct response rates were statistically different across the control, motor, and visual conditions. As shown in Table 3.2, the effect of condition was significant only for the perception of phoneme [p]. A post-hoc analysis using paired t-test with Bonferroni correction revealed a significant difference ($p<.01$) only between the visual and control conditions (see Fig. 3.3). From these results, it could be summarized that articulatory movements and visual motion matching with auditory stimuli either improved or did not affect the phoneme perception.

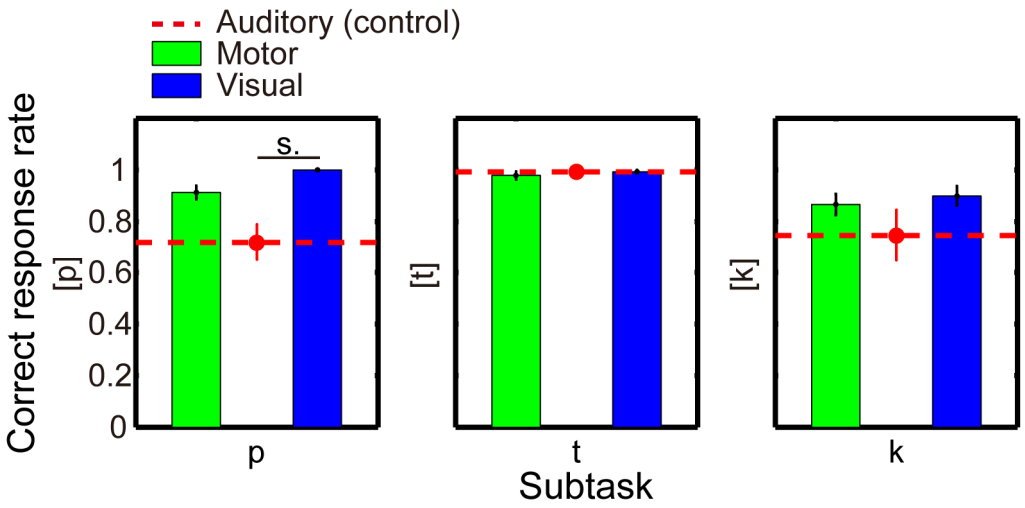

Fig. 3. 3. Phoneme intelligibility under concordant subtasks.
Mean and standard error (N = 10) of correct response rates for each phoneme [p], [t], and [k] when whispering (motor) or seeing (visual) concordant phonemes were shown. A one-way repeated measures ANOVA was performed for each phoneme with condition as factor (see Table 3.2). A significant main effect was found only for the phoneme [p], where post-hoc analysis revealed a significant difference only between visual and control conditions ($p<.01$ by paired t-test with Bonferroni correction).

Auditory stimulus type	F(2,18)	p
[pa]	11.968	< .005
[ta]	1.000	> .05
[ka]	2.723	> .05

Table 3.2. Effect of concordant subtask on phoneme intelligibility.
For each stimulus under concordant subtask conditions, statistical difference in correct response rate was analyzed using a one-way repeated measures ANOVA with condition (auditory/motor/visual) as factor. A significant main effect was observed only for [pa].

Discordant case

The participants' perception of phonemes while they silently articulated incongruent phonemes or saw the model speaker's mouth producing incongruent phonemes was assessed. Correct response rates for [p], [t], and [k] were shown in the top, middle, and bottom rows of Fig. 3.4, respectively, where two different types of discordant motor/visual subtask were assigned to each column. In all panels, the mean correct response rates under the motor and visual conditions did not exceed the control level. For each discordant combination of auditory stimulus and motor/visual subtask type, a one-way repeated measures ANOVA was conducted to examine whether the correct response rates were statistically different across the control, motor, and visual conditions. As shown in Table 3.3, the effect of condition was significant for all stimulus-subtask combinations. A post-hoc analysis using paired t-test with Bonferroni correction revealed the following significant differences ($p<.01$): a reduction of the rate for [p] by the visual subtask [t] and [k] (see top rows of Fig. 3.4), a reduction of the rate for [t] by the visual subtask [p] and the motor subtask [k] (see middle rows of Fig. 3.4), and a reduction of the rate for [k] by the visual subtask [p] and the motor subtask [t] (see bottom rows of Fig. 3.4).

Considering that the crucial articulator for the production of [p] is the lips, whereas the crucial articulator for the production of [t] and [k] is the tongue, the above results could be restated as follows: (1) Whereas the perception of lip-related phoneme ([p]) was degraded by the visual tongue motion ([t] and [k]), it was not by the articulatory tongue movement. (2) Whereas the perception of tongue-related phonemes ([t] and [k]) was degraded by the visual lip motion ([p]), it was not by the articulatory lip movement. (3) The perception of each of tongue-related phoneme ([t] and [k]) was degraded by the other articulatory tongue movements ([k] and [t], respectively), whereas it was not by the visual tongue motions. The dissimilarity between the interferential effect of discordant visual and motor subtasks on phoneme perception could be summarized as follows: the auditory-visual integration occurred across different speech organs, whereas the auditory-articulatory integration occurred within a same organ, as illustrated in Fig. 3.5

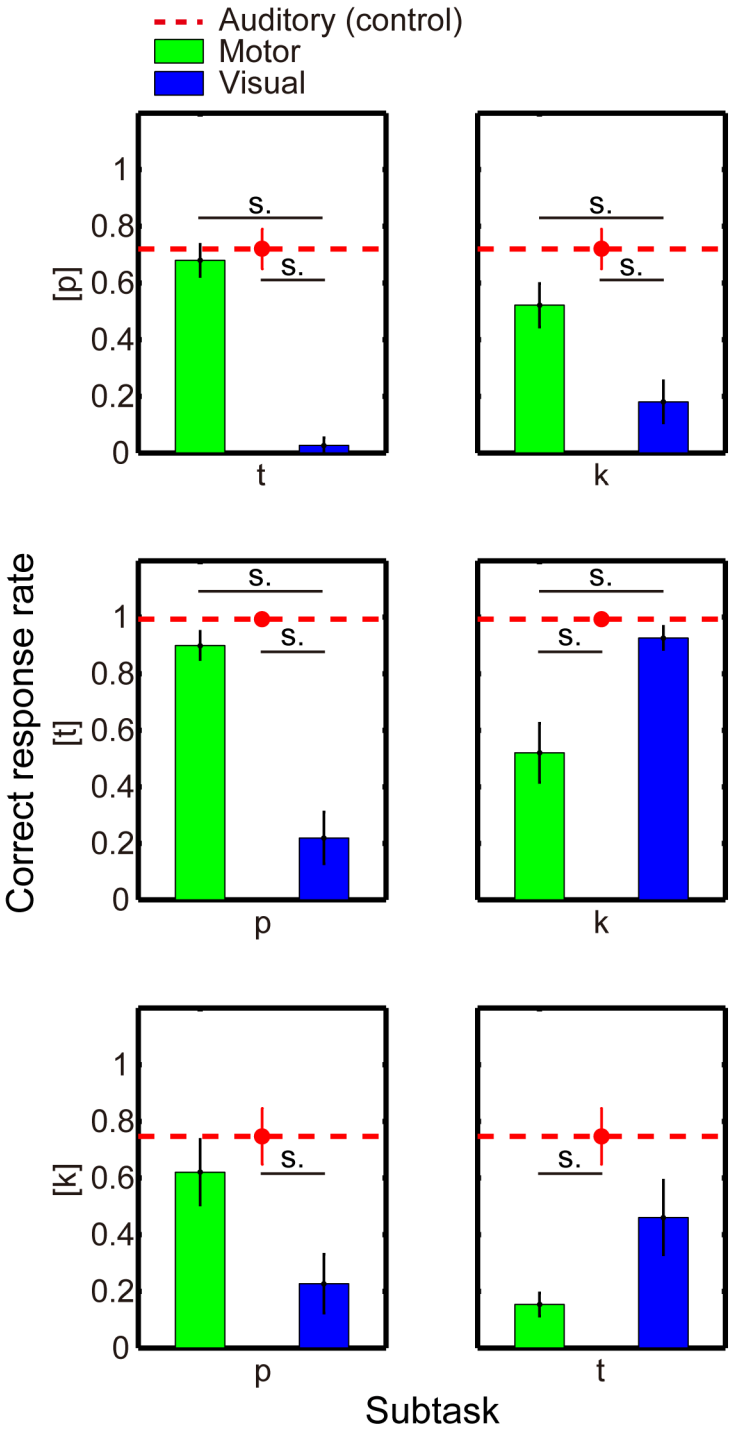

Fig. 3.4. Phoneme intelligibility under discordant subtasks.
Mean and standard error (N = 10) of correct response rates for each phoneme [p], [t], and [k] when whispering (motor) or seeing (visual) discordant phonemes were shown. A one-way repeated measures ANOVA was performed for each discordant pair of stimulus and subtask, with condition as factor (see Table 3.3). A significant main effect was found for all pairs. Post-hoc analysis

revealed a significant difference between motor and control conditions ($p<.01$ by paired t-test with Bonferroni correction) only for the stimulus [t] when whispering [k] and for the stimulus [k] when whispering [t]. See text for details.

Auditory stimulus type	Subtask type (Motor/Visual)	F(2,18)	p
[pa]	ta	54.727	< .005
	ka	16.689	< .005
[ta]	pa	40.529	< .005
	ka	20.253	< .005
[ka]	pa	12.700	< .005
	ta	17.967	< .005

Table 3.3. Effect of discordant subtask on phoneme intelligibility.
For each combination of stimulus and discordant subtask conditions, statistical difference in correct response rate was analyzed using a one-way repeated measures ANOVA with condition (auditory/motor/visual) as factor. A significant main effect was observed for all combinations.

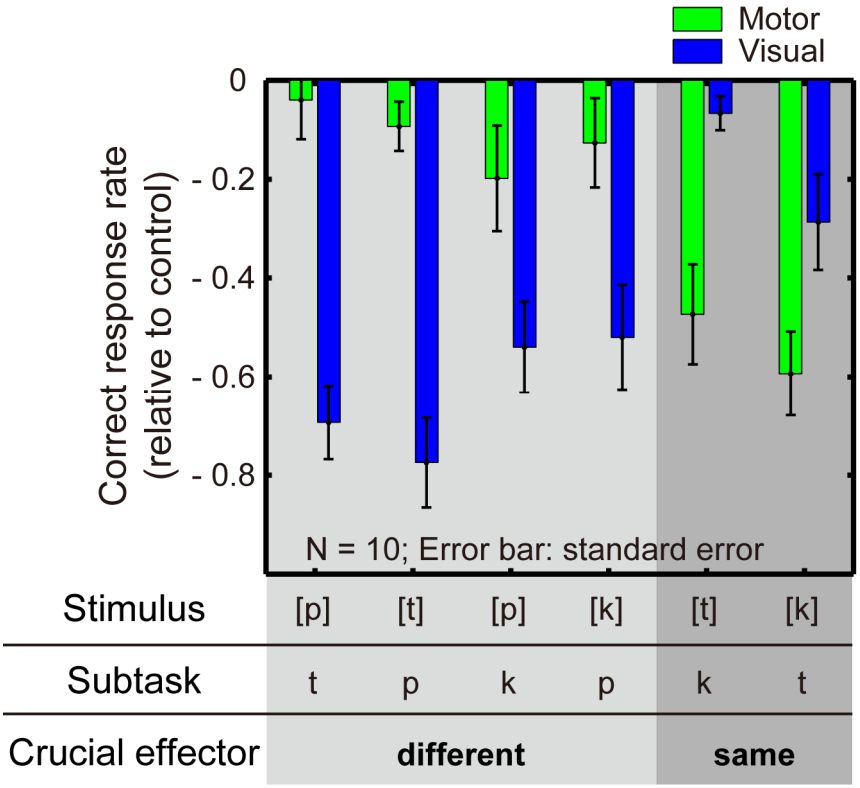

Fig. 3.5. Effect of discordant subtask on phoneme intelligibility in terms of speech organs (effectors).

Change in correct response rate for each stimulus caused by each discordant subtask was evaluated by subtracting from each control level. Considering that [p] is articulated by the lips whereas both [t] and [k] are articulated by the tongue, the effect of discordant subtask on phoneme perception could be summarized as follows: auditory-visual integration occurred across different speech organs, whereas auditory-articulatory integration occurred within a same organ.

3.4 Discussion

3.4.1 Articulatory vs. visual interferential effects

The three phonemes [p], [t], and [k] examined in the current study are all plosives (oral stop), for which the air flow in the vocal tract is blocked and released by specific movements of the lips or the tongue. Phonetically these phonemes are classified into labial (p), alveolar (t), and velar (k) plosives according to the place of articulation, i.e., where in the vocal tract (front, central, and back, respectively) the blockage is formed. In a typical example of the McGurk effect, auditory sound [p] presented simultaneously with visual motion of mouth pronouncing [k] leads to the perception of [t] (fusion effect) [8]. This effect may reflect the underlying properties of audio-visual integration where a mismatch between the place of articulation associated with auditory sound (front) and visual motion (back) can elicit the perception of another place of articulation in between the two (central).

On the contrary, the current study revealed that the self articulatory movements affected the perception of auditory phonemes in a motor-dependent manner. Among the three places of articulation, only labial is associated with the movements of the lips, whereas the remaining two (alveolar and velar) are associated with the tongue movements. And, in fact, the articulatory movement for pronouncing alveolar [t] disturbed the auditory perception of velar [k], and vice versa, whereas the articulatory movement for labial [p] did not affect the perception of either [t] or [k]. Also the articulatory movements for both [t] and [k] did not affect the perception of [p]. The motor-dependent manner of the auditory-articulatory interferential effect observed in the current study may be a reflection of somatotopic linkage between the neural networks for speech production and perception suggested by a series of recent studies [4-6, 10]

3.4.2 Which aspect of articulatory movement affected perception?

Although the involvement of motor system in speech perception has been conceptually well described [12, 13] and also experimentally evident [14-18], there have been controversial arguments regarding how incoming auditory information is processed and interpreted by the motor nervous system and triggers a specific phoneme perception. One type of evidence suggests that speech motor control process does not directly contribute to speech perception, whereas it is closely related to action word/concept processing, semantic and syntactic processing , or even to lexical decision [19-21]. However, the effector-specific manner of auditory-articulatory interaction observed in the current study strongly supports the possibility of direct linkage between the processes for speech motor control and phoneme perception.

The remaining issue to be further investigated is which stage of speech motor control process such as planning, execution, and proprioceptive consequences played an essential role in modulating phoneme perception in the current study. The phoneme intelligibility changes observed in the current study may have reflected several different levels of integration between the neural representation of auditory input and of articulatory movements. There has been an evidence that covert speech can affect speech perception in an articulatory-constrained manner [22]. In the current study, articulatory imagery elicited during the preparation of silent articulation might have a certain effect on participants' perceptual response. It will be important to further examine whether articulatory imagery itself can have a precise somatotopic representation, which may cause an effector-specific effect as observed in the current study.

3.5 References

1. Liberman, A.M., et al., *Perception of the speech code.* Psychol Rev, 1967. **74**(6): p. 431-61.
2. Devlin, J.T. and K.E. Watkins, *Stimulating language: insights from TMS.* Brain, 2007. **130**(Pt 3): p. 610-22.
3. Pulvermuller, F. and L. Fadiga, *Active perception: sensorimotor circuits as a cortical basis for language.* Nat Rev Neurosci, 2010. **11**(5): p. 351-60.
4. Fadiga, L., et al., *Speech listening specifically modulates the excitability of tongue muscles: a TMS study.* Eur J Neurosci, 2002. **15**(2): p. 399-402.
5. Wilson, S.M., et al., *Listening to speech activates motor areas involved in speech production.* Nat Neurosci, 2004. **7**(7): p. 701-2.
6. Pulvermuller, F., et al., *Motor cortex maps articulatory features of speech sounds.* Proc Natl Acad Sci U S A, 2006. **103**(20): p. 7865-70.
7. D'Ausilio, A., et al., *The motor somatotopy of speech perception.* Curr Biol, 2009. **19**(5): p. 381-5.
8. McGurk, H. and J. MacDonald, *Hearing lips and seeing voices.* Nature, 1976. **264**(5588): p. 746-8.
9. Sams, M., R. Mottonen, and T. Sihvonen, *Seeing and hearing others and oneself talk.* Brain Res Cogn Brain Res, 2005. **23**(2-3): p. 429-35.
10. Watkins, K.E., A.P. Strafella, and T. Paus, *Seeing and hearing speech excites the motor system involved in speech production.* Neuropsychologia, 2003. **41**(8): p. 989-94.
11. Ito, T., M. Tiede, and D.J. Ostry, *Somatosensory function in speech perception.* Proc Natl Acad Sci U S A, 2009. **106**(4): p. 1245-8.
12. Liberman, A.M. and I.G. Mattingly, *The motor theory of speech perception revised.* Cognition, 1985. **21**(1): p. 1-36.
13. Galantucci, B., C.A. Fowler, and M.T. Turvey, *The motor theory of speech perception reviewed.* Psychon Bull Rev, 2006. **13**(3): p. 361-77.
14. Meister, I.G., et al., *The essential role of premotor cortex in speech perception.* Curr Biol, 2007. **17**(19): p. 1692-6.
15. Mottonen, R. and K.E. Watkins, *Motor representations of articulators contribute to categorical perception of speech sounds.* J Neurosci, 2009. **29**(31): p. 9819-25.
16. Watkins, K. and T. Paus, *Modulation of motor excitability during speech perception: the role of Broca's area.* J Cogn Neurosci, 2004. **16**(6): p. 978-87.
17. Wilson, S.M. and M. Iacoboni, *Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception.* Neuroimage, 2006. **33**(1): p. 316-25.
18. Zheng, Z.Z., K.G. Munhall, and I.S. Johnsrude, *Functional overlap between regions involved in speech perception and in monitoring one's own voice during*

speech production. J Cogn Neurosci. **22**(8): p. 1770-81.
19. Scott, S.K., C. McGettigan, and F. Eisner, *A little more conversation, a little less action--candidate roles for the motor cortex in speech perception.* Nat Rev Neurosci, 2009. **10**(4): p. 295-302.
20. Roy, A.C., et al., *Phonological and lexical motor facilitation during speech listening: a transcranial magnetic stimulation study.* J Physiol Paris, 2008. **102**(1-3): p. 101-5.
21. Sato, M., P. Tremblay, and V.L. Gracco, *A mediating role of the premotor cortex in phoneme segmentation.* Brain Lang, 2009. **111**(1): p. 1-7.
22. Sato, M., et al., *Multistable syllables as enacted percepts: a source of an asymmetric bias in the verbal transformation effect.* Percept Psychophys, 2006. **68**(3): p. 458-74.

4. Conclusion

4.1 Summary of the thesis

The experimental studies described in this thesis elucidated some functional aspects of the sensorimotor systems involved in speech production and perception at the acoustic-phonological level, independent from lexical, semantic or syntactic processing.

In chapter 2, a rapid auditorily induced change in articulatory lip movement was found when auditory feedback preceded real syllable production by 50 ms when isolated syllables were spoken repeatedly at a rate of 300 ms per syllable. The change was not significantly induced when the feedback occurred earlier than 50 ms or was delayed, and/or the feedback syllable was replaced by other syllables. The results suggested that a compensatory mechanism detected sensory errors between the internally predicted and actually provided auditory information associated with the self-produced speech, by using a temporally asymmetric window in which acoustic features of the syllable to be produced may be coded. This study provides evidence that the temporal dynamics of articulatory lip movement must be correctly maintained not only with somatosensory feedback resulting from peripheral motor activation but also with auditory feedback of self-produced speech.

In chapter 3, the hearing of speech sounds was found to be affected by the listeners' own speech movements: the perception of auditory phonemes was disturbed when the listener articulated incongruent phonemes. This motor interferential effect on speech perception showed a different pattern from the visual one known as the McGurk effect [1]. Auditory-visual integration can occur across different speech organs: observing lip motion affects our hearing of phonemes produced by the tongue, and vice versa. On the other hand, auditory-articulatory integration can occur within the same speech organ: articulating a phoneme with the tongue affects our hearing of the other phonemes produced by the tongue, whereas articulating a phoneme with the lips does not affect our hearing of the tongue-related phonemes. Recent brain imaging studies have reported that phoneme perception activates specific motor-related neural circuits, which are invoked with the production of the same phoneme [2-4]. The perception of phonemes produced by specific articulatory movements of individual speech organs such as the lips and tongue, which are somatotopically mapped onto different motor brain areas, activates those areas differently in a phoneme-dependent manner [5]. Our findings may be associated with this somatotopic activation of motor brain regions during phoneme perception.

The quantitative knowledge of dynamic characteristics of speech motor control and auditory processing can be used to help ascertain the cause of discomfort associated with conventional remote conferencing systems. Existing signal processing techniques such as echo cancellation and voice activity detection (VAD) seem to work fairly well in reducing an unacceptable amount of speech disturbance due to transmission delay [6]. However, transmission loss at the onset of each utterance in a conversation somewhat unavoidable in VAD sometimes causes undesirable speech overlap between speakers at different sites. In order to enable higher interactivity in future multi-modal communication systems, fluent conversation and reliable communication should be guaranteed while allowing natural overlap of utterances. The experimental studies described in chapter 2 and 3 have indicated that an improperly designed audio-visual conferencing system may lead to misarticulation (phoneme production error) and misperception (phoneme perception error) during a conversation, which stem from hearing delayed feedback of own speech and ill-timed others' utterances while speaking, as well as from hearing others' utterances with seeing their incongruent visual motion of the mouth. A desirable multi-modal communication system should therefore be achieved, within the available transmission capacity of the network used, by optimally balancing the 1) round-trip delay in each modality and 2) inter-modal lag, in terms of the functional characteristics of sensorimotor interactions associated with human speech production and perception.

4.2 Future work

The experimental results presented in this thesis have strongly suggested that the neural mechanisms of phoneme production and perception involve multisensory-motor integration. However, the inter-sensory and inter-sensorimotor interactions between multiple sensory (i.e., auditory, somatosensory, and visual) modalities remain unknown.

The acquisition of exquisite control of the respiratory, laryngeal, velopharyngeal, and articulatory subsystems required for intelligible speech production is inseparable from auditory and somatosensory feedback during learning [7, 8]. On the contrary, post-lingually deaf patients can produce intelligible speech to some extent [9]. Some studies have even claimed that the somatosensory-motor loop is essential to maintain articulatory movements for both normal-hearing and deaf individuals [10, 11]. More specific nature of the relationship between auditory-motor and somatosensory-motor systems should be clarified.

There have been some controversial views on the direct involvement of motor system in speech perception, claiming that coactivation of motor areas during the perception of phonemes does not reflect a primary, perceptual route to the comprehension of speech [12, 13]. However, accumulated evidence from neurophysiological and neuroimaging studies has shown that the auditory and/or visual perception of others' actions activates the observer's corresponding motor representations in the brain [14-17]. The experimental results presented in Chapter 3 of this thesis have revealed, for the first time, that auditory-motor integration occurs in a different manner from audio-visual one. Another study has examined auditory-somatosensory interaction in speech perception by using passive facial skin stretching device [18]. All these evidence support multisensory-motor nature of sublexical phonological representation in speech perception. The neural mechanisms related to higher-level linguistic information processing such as lexicon, semantics and syntax should also be examined in terms of inter-sensory and inter-sensorimotor interactions.

4.3 References

1. McGurk, H. and J. MacDonald, *Hearing lips and seeing voices.* Nature, 1976. **264**(5588): p. 746-8.
2. Fadiga, L., et al., *Speech listening specifically modulates the excitability of tongue muscles: a TMS study.* Eur J Neurosci, 2002. **15**(2): p. 399-402.
3. Wilson, S.M., et al., *Listening to speech activates motor areas involved in speech production.* Nat Neurosci, 2004. **7**(7): p. 701-2.
4. Pulvermuller, F., et al., *Motor cortex maps articulatory features of speech sounds.* Proc Natl Acad Sci U S A, 2006. **103**(20): p. 7865-70.
5. D'Ausilio, A., et al., *The motor somatotopy of speech perception.* Curr Biol, 2009. **19**(5): p. 381-5.
6. Venkatesha Prasad, R., et al. *Comparison of voice activity detection algorithms for VoIP*. in *Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on*. 2002.
7. Guenther, F.H., S.S. Ghosh, and J.A. Tourville, *Neural modeling and imaging of the cortical interactions underlying syllable production.* Brain Lang, 2006. **96**(3): p. 280-301.
8. Shiller, D.M., et al., *Perceptual recalibration of speech sounds following speech motor learning.* J Acoust Soc Am, 2009. **125**(2): p. 1103-13.
9. Lane, H. and J.W. Webster, *Speech deterioration in postlingually deafened adults.* J Acoust Soc Am, 1991. **89**(2): p. 859-66.
10. Tremblay, S., D.M. Shiller, and D.J. Ostry, *Somatosensory basis of speech production.* Nature, 2003. **423**(6942): p. 866-9.
11. Nasir, S.M. and D.J. Ostry, *Speech motor learning in profoundly deaf adults.* Nat Neurosci, 2008. **11**(10): p. 1217-22.
12. Lotto, A.J., G.S. Hickok, and L.L. Holt, *Reflections on mirror neurons and speech perception.* Trends Cogn Sci, 2009. **13**(3): p. 110-4.
13. Scott, S.K., C. McGettigan, and F. Eisner, *A little more conversation, a little less action--candidate roles for the motor cortex in speech perception.* Nat Rev Neurosci, 2009. **10**(4): p. 295-302.
14. Nishitani, N. and R. Hari, *Viewing lip forms: cortical dynamics.* Neuron, 2002. **36**(6): p. 1211-20.
15. Skipper, J.I., et al., *Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception.* Cereb Cortex, 2007. **17**(10): p. 2387-99.
16. Jarick, M. and J.A. Jones, *Observation of static gestures influences speech production.* Exp Brain Res, 2008. **189**(2): p. 221-8.
17. Watkins, K.E., A.P. Strafella, and T. Paus, *Seeing and hearing speech excites the motor system involved in speech production.* Neuropsychologia, 2003. **41**(8): p. 989-94.

18. Ito, T., M. Tiede, and D.J. Ostry, *Somatosensory function in speech perception*. Proc Natl Acad Sci U S A, 2009. **106**(4): p. 1245-8.

Appendix

Specifications of the instruments used in the experiments

Chapter 2

Auditory feedback alteration:	
Electret condenser microphone	ECM-G3M (Sony, Japan)
Microphone amplifier	ZDT 1021 (Earthworks, USA)
Low-pass filter (48 dB/oct)	P-85 (NF, Japan)
Analog-to-digital converter (16 bits), Digital-to-analog converter (16 bits) (controlled by custom-made software on workstation)	Dasbox (Comex Electronics, Japan)
Workstation	Sun Ultra 2 (Sun Microsystems, USA)
Noise generator	Type 1405 (Bruel & Kjaer, Denmark)
Audio mixer	SRP-X6004 (Sony, Japan)
In-ear earphones	ER-4S (Etymotic Research, USA)
Calibration probe microphone	ER-7C (Etymotic Research, USA)
Llip motion capture:	
3D motion capture system (controlled by Qualisys Track Manager on Windows PC)	Oqus motion capture cameras (Qualisys, Sweden)
Windows PC	Dimension 9150 (Dell, USA)

Auditory stimulus recording:	
Electret Condenser Microphone	ECM-330 (Sony, Japan)
Analog-to-digital converter	SE-U33GX (Onkyo, Japan)
Recording and Editing Tools (on Windows PC)	Audacity, Praat (Freeware)
Visual stimulus recording:	
Camcorder	HDR-HC3 (Sony, Japan)
Editing Tools (on Windows PC)	Premiere Elements 7 (Adobe Systems, USA)
Stimulus presentation:	
Headphone	HD280Pro (Sennheiser, Germany)
Calibration probe microphone	ER-7C (Etymotic Research, USA)
LCD Display (18.1 inches)	FlexScan L66 (Nanao, Japan)
Presentation Tools (on Windows PC)	Windows Media Player
Windows PC	ThinkPad X200 (Lenovo, USA)

Publication List (as the first author)

Related to the thesis

Full Paper:
Mochida T, Gomi H, and Kashino M. Rapid Change in Articulatory Lip Movement Induced by Preceding Auditory Feedback during Production of Bilabial Plosives. *PLoS ONE* 5: e13866, 2010.

Conference Paper:
Mochida T, Kimura T, Hiroya S, Kitagawa N, Gomi H, and Kondo T. Effector-specific effect of self-articulatory movement on speech perception. In: *The Society for Neuroscience 40th Annual Meeting*. San Diego: Society for Neuroscience, 2010.

Mochida T, Gomi H, and Kashino M. Involuntary and short-latency articulatory compensation induced by altered auditory feedback. In: *The Society for Neuroscience 35th Annual Meeting*. Washington, D.C.: Society for Neuroscience, 2005.

Other works

Full Paper:
Mochida T, and Honda M. Estimation of the vocal tract area function from the impulse response at the lips. *The Journal of the Acoustical Society of Japan* 55: 147-155, 1999 (in Japanese).

Conference Paper:
Mochida T, Hiroya S, Honda M, Nishikawa K, and Takanishi A. Articulatory control of talking robot by mimicking formant trajectories of human speech. In: *6th International Seminar on Speech Production*. Sydney: 2003, p. 173-178.

Mochida T, Honda M, Hayashi K, Kuwae T, Tanahashi K, Nishikawa K, and Takanishi A. Control system for talking robot to replicate articulatory movement of natural speech. In: *International Conference on Spoken Language Processing*. Denver: 2002, p. 1533-1536.

Mochida T, and Honda M. Estimation of vocal-tract area function from lip impulse response based on lossy vocal-tract digital filter model. *The Journal of the Acoustical Society of America* 110: 2776-2776, 2001.

Mochida T, and Honda M. An experimental study on acoustical measurement of vocal-tract area function. In: *CREST Workshop on Speech Motor Control and Modeling*. Japan: 2001, p. 40-40.

Mochida T, and Honda M. A study on estimation of vocal tract area function from impulse response at the lips. *Technical Report of IEICE* SP98-124: 49-56, 1999 (in Japanese).

Mochida T, and Honda M. Acoustical measurement of the vocal-tract area function: Improvements on acoustical measurement method and numerical method. In: *Hokkaido Workshop on Speech Production*. Japan: 1998, p. 24-25.

Mochida T, and Honda M. Evaluation of an acoustical measurement method for measuring vocal tract area. *Technical Report of IEICE* SP97-46: 15-22, 1997 (in Japanese).

Mochida T, and Honda M. Acoustic measurement of vocal tract area function: An experimental study. *The Journal of the Acoustical Society of America* 100: 2658-2658, 1996.

Mochida T, Kobayashi T, and Shirai K. Speech synthesis of Japanese sentences using large waveform data-base. *Technical Report of IEICE* SP93-91: 13-18, 1993.

Article:
Mochida T, Kobayashi T, and Shirai K. Speech synthesis of Japanese sentences using large waveform data-base. *Waseda University Advanced Research Center for Science and Engineering* Technical Report No.93-17: 1993.

Mochida T. Speech synthesis of Japanese sentences using speech data-base. *Waseda University Bulletin of the Centre for Informatics* 17: 58-67, 1994 (in Japanese).