Audiovisual integration in Dichotic listening

Kei Omata & Ken Mogi

Department of Computation Intelligence and Systems Science, Tokyo Institute of Technology
& Sony Computer Science Laboratories
omata@brn.dis.titech.ac.jp

Abstract

The presentation of visual stimuli of moving mouth is known to affect the phoneme perception of the auditory stimuli during natural speech, and provides important insights into the scientific and technological study of speech. In order to investigate the laterality of the audio-visual integration we studied the relationship between the McGurk effect (McGurk and MacDonald 1976) and Right Ear Advantage (REA) in dichotic listening (DL). When visual /ga/ was presented with the dichotic auditory stimuli, both the McGurk effect and REA was observed. On the other hand, when visual /ba/ was paired with the dichotic auditory stimuli, the visual /ba/ strongly affected the proportion of the phoneme perceived. In addition, the phoneme perception in the DL was influenced by the auditory syllables fed into the right ear more than those fed into the left ear. These results suggest that the cortical processes that underlie the McGurk effect is consistent with, and probably shares the same process with, those that lead to REA.

1. Introduction

Visual stimuli often help us to perceive the auditory syllables. In the presence of noise, the intelligibility of phoneme perception is improved by the visual observation of the speaker’s facial and lip movements [1]. An ability of the lip-reading is manifest without a special training. In a striking example, lip-reading information is involved with speech perception under incoherent audiovisual stimuli (the McGurk effect [2]). For example, the visual /ga/ paired with the auditory /ba/ usually causes the perception of /da/, suggesting the existence of an integration process between visual and auditory information.

In dichotic listening, two different words are delivered to the left and right ear simultaneously. Normal subjects with left-hemisphere language lateralization tend to report more frequently the words presented to the right ear (Right Ear Advantage (REA) [3, 4]). In contrast, a left ear advantage is observed for non-speech and musical sounds. On the basis of neuropsychological evidence, these results have been explained by the ‘structural theory’ [3], in which the contralateral pathway from the ear to the auditory cortex suppresses the ipsilateral one. A recent neuroimaging study using positron emission tomography suggests that linguistic dichotic stimuli evoke stronger cortical responses in the left temporal lobe, while non-linguistic dichotic stimuli evoke a stronger response in the right one [5]. Another MEG (magnetoencephalography) study found that in tasks involving non-verbal tones, neurophysiological inhibition was driven by the contralateral auditory pathway over the ipsilateral one during dichotic listening [6].

It is known that the effect of visual speech stimuli is stronger than that of REA in DL. When the visual information is concordant with the auditory input to one ear, the perception of that syllable increased strongly, irrespective of the ear of stimulation [7]. It is interesting to investigate how the audio-visual information is integrated in DL when there is a discrepancy between audio and visual information, possibly leading to the McGurk effect. Here we report that the McGurk effect is observed in DL condition.

2. METHODS

2.1. Subjects

Eleven healthy volunteers (3 females and 8 males) participated in the experiment. Subjects’ age ranged from 23 to 29. They were all native Japanese speakers with normal hearing and normal or corrected-to-normal vision. All the subjects were right-handed, except for one female.

2.2. Stimuli

All syllables were articulated by a male Japanese speaker. Three consonant-vowel (CV) syllables were used: /ba/, /da/, /ga/. Each syllable was recorded with a DV-video camera (Sony DCR-TRV30) and a soundcard (M-Audio Audiophile USB) with a microphone. These videos were edited by standard software (premire 6.0 and Cubase SL 2) on the PC (IBM Thinkpad X31).

In the V (visual) condition, the visual stimuli of /ba/, /da/, and /ga/ were presented. The video started and ended with the speaker’s lips closed. The duration of the video clip was 1 second, with the syllable articulation starting at 500 ms after stimuli onset. There were 3 different stimuli for each syllable. Video clips of /na/ syllables were used for control, because /na/ syllable is similar to /da/ syllable in the articulation. The results of the discrimination between them are expected to similar to each other, if subjects respond based by the visual stimuli.

In the A (auditory) condition, auditory stimuli were presented to both ears. There are 9 combinations: /ba-ba/, /ba-da/, /ba-ga/, /da-ba/, /da-da/, /da-ga/, /ga-ba/, /ga-da/, /ga-ga/ (left ear – right ear). In addition, there were three diotic condition where two different syllables were presented simultaneously: /ba+da/, /ba+ga/, /da+ga/. In order to increase the occurrence of the McGurk effect in Japanese subjects [8], Gaussian noise (signal-to-noise ratio 13.9dB) was added to all auditory stimuli. The duration of auditory stimuli was 1 sec, with the CV stimuli starting at 500ms after stimulus onset. The duration of formant transitions was about 300ms. The
intensity of each auditory sound was adjusted to be in the same range of ±1.1dB.

In the AV (audio-visual) condition, visual stimuli (visual /ba/, /ga/) and dichotic auditory stimuli (/ba-ga/, /ga-ba/) were combined to create 4 different audio-visual stimuli. In addition, diotic auditory /ba+ga/ was combined with two visual tokens (visual /ba/, /ga/). Three coherent audio-visual stimuli were used for control stimuli: audio-visual /ba/, /da/, and /ga/. Finally, the McGurk effect-inducing stimulus (visual /ga/ paired with auditory /ba/) was present. In summary, there were 10 types of stimuli totally in the AV condition.

Subjects were presented with each visual stimulus on the monitor and auditory stimulus through the headphone (AKG K271s). They were asked to gaze the monitor and listen to each utterance and were required to report what they heard ([Ba],[Da],[Ga]) by clicking on the buttons displayed on the monitor with a mouse.

3. Results

In the V condition, the visual /ba/ was discriminated easier than other syllables (97%). The level of discrimination of the visual /da/ and /ga/ were almost the same (the chi-square test, df = 1, N = 627, $p > 0.5$). The visual /ga/ was sometimes perceived as /da/ ($P = 0.5, z = 3.24, p < 0.001$). The proportion of the visual /na/ identifications as control was similar to the result of the visual /da/ (Figure 1).

In the A condition, three normal auditory syllables yielded a high intelligibility score of 98%. Fig. 2 shows the data for the diotic merged stimuli. When the acoustical /ba/ and /da/ were presented dichotically, subjects perceived /da/ than /ba/ regardless of the ear of presentation. The intelligibility score of /da/ identifications was significantly higher when it was presented to the right ear (df = 1, N = 646, $p < 0.05$).

The REA is defined by subtracting the proportion of the phoneme perception of a stimulus to the left ear from that of the stimulus to the right ear. The REA of /da/ and /ga/ perceptions were both 8%. When the /ba/ and /ga/ syllables were presented dichotically, subjects mainly perceived the stimuli input to the right ear (df = 1, N = 647, $p < 0.05$). The REA of /ba/ perception was 7% and that of /ga/ perception was 44% (Figure 3). When the /da/ and /ga/ stimuli were presented, the REA of /da/ perceptions was 34% and that of /ga/ perceptions was 33% (df = 1, N = 650, $p < 0.01$).

In the AV condition, three coherent stimuli (the audio-visual /ba/, /da/, /ga/) were distinguished well (99%). When the audio-visual stimulus inducing the McGurk effect was presented, the intelligibility score of /da/ perceptions was 76%. Most of subjects perceived /ba/ syllable under the diotic /ba-ga/ stimulus with the visual /ba/ (81%). Conversely, the diotic /ba+ga/ stimulus with the visual /ga/ was perceived as /da/ 69% of the time. In the dichotic condition, where visual /ba/ was combined with the dichotically presented /ba-ga/ or /ga-ba/, the intelligibility score of /ba/ perceptions was 81% or 88%. The REA of /ba/ perception was 7% and that of /ga/ perceptions was 5% (df = 1, N = 647, $p > 0.1$). In the case of visual /ga/ with dichotic auditory stimuli /ba-ga/ or /ga-ba/, subjects basically perceived the stimuli input to the right ear. The REA of /da/ perception was 21% and that of /ga/ perception was 24% (df = 1, N = 637, $p < 0.01$) (Figure 4).
Figure 4: The responses of the phoneme perception in the AV condition. VB: visual /ba/ VG: visual /ga/.

4. Discussion

The results reveal some interesting insights into the nature of audio-visual integration under dichotic listening condition.

In the AV condition, where the visual /ga/ was paired with the dichotically presented auditory /ga-ba/ stimulus, the subjects perceived /da/. In the presentation of auditory /ba-ga/ paired with the visual /ga/, subjects sometimes perceived /da/. These results demonstrate that the McGurk effect occurs in the DL. The presence of the McGurk effect in DL thus provides novel insights into the nature of audio-visual integration in DL.

McGurk and McDonald [2] reported that there are two modes of integration when the audio-visual stimuli are discordant. When the visual /ga/ is dubbed to the auditory /ba/, subjects perceived /da/. This type of audio-visual pair was called "fused pairs". On the other hand, the subject sometimes perceived /ba-ga/ or /ga-ba/ when presented with the visual /ba/ with the auditory /ga/. This type was called 'combination pairs'.

In this experiment, the dominance of visual /ba/ remained in the dichotic audio-visual stimuli irrespective of the ear of stimulation. The visual /ba/ combined with the dichotic auditory stimulus /ba-ga/ or /ga-ba/ could induce either the combination perception or /ba/ perception. The subjects likely reported the /ba/ perception as a "reasonable" choice. The presentation of visual /ga/ with the dichotic auditory stimulus /ba-ga/ or /ga-ba/ might result in the perception of /da/ (the McGurk effect) or /ga/. Then subjects thus must respond the both /da/ and /ga/ responses in a manner similar to REA. This was the novel audio-visual integration situation studied in our present study.

The dominance of the visual /ba/ in the AV condition might be related with the increase of the phoneme perception using the visual information in the presence of noise [1]. The auditory stimulus that is discordant with the visual /ba/ may become background sounds similar to noise. Conversely, the visual /ga/ is able to combine with both auditory /ba/ and /ga/. Thus, the REA may be effective when the visual /ga/ is paired with auditory /ba/ and /ga/.

In view alternative to the structural theory, the ‘attentional theory’ [9] allows that the attention to a particular type of auditory stimulus influence the ear advantage rather than to a structural advantage. However, although attention influences the REA, attention is initially biased to the right ear in the typical DL condition, where subjects concentrate to both ears for the consonant-vowel stimuli [10, 11]. Thus, in this experiment, we didn’t specially take account of ‘attention’ to the auditory stimuli.

There are many researches on the laterality of the language function in the left hemisphere. For instance, A. Shestakova et al. [12] reported that the mismatch negativity (MMNm) to the phoneme-category change over the left hemisphere rather than the right hemisphere appear when natural speech stimuli (/i/, /a/, /u/) constantly changing and varying in acoustic content are presented. There is a laterality of hearing natural speech sounds and the left hemisphere is important for the presentation of the phoneme perception.

The structural theory provides a possible explanation for induction of REA under the DL. Indeed, magnetic fields evoked by contralateral auditory stimulation are greater than those evoked by ipsilateral stimulation, suggesting that there is an interhemispheric asymmetry in magnetic auditory evoked fields [13]. Therefore, on MEG (magnetoecephalography) study, the ipsilateral inhibition appears only when the contralateral sound fundamental frequency is similar to the ipsilateral sound. The M100 amplitude but not M50 over the right auditory cortex increased gradually when sounds of increasing intensity are provided at the ipsilateral ear in the right hemisphere [7]. It supports the idea that the interaction between contralateral and ipsilateral pathways induces the inhibition that is due to the level of primary and secondary auditory cortex. Verbal dichotic stimuli induce stronger cortical responses in the left lobe [8].

Recent experiments using neuroimaging suggest that the audiovisual integration processing is due to activities around the posterior left occipito-temporal cortex. Sekiyama et al. [14] reported by means of FMRI and PET that the activation in response to the audio-visual stimulus that induces the McGurk effect was found in the left temporal cortex, extending more posteriorly toward the central region (the left superior temporal sulcus) in the AV condition with noise. The activity of the left superior temporal sulcus was induced by the semantically congruent audio-visual speech [15].

J. A. Jones et al. [16] reported that when the auditory /aba/ is presented with visual /ava/, the number of /b/ responses from the behavioral results is positively correlated with a cluster of activation in the left occipito-temporal junction in the brain, which region was thought to process moving stimuli (VS). That is, the presentation of the auditory stimuli allows auditory information to infect activity in visual processing cortex. On one hand, silent lip-reading and visual pseudospeech activate the auditory cortical sites, which include BA 41, 42, 22, suggesting that lip-reading modulates the perception of auditory speech at a prelexical level [17]. These suggest the both possibility of the auditory information processing in the visual area and the visual one in the auditory area.

Neuroimaging studies supported the idea that the left and right STS (superior temporal sulcus) region are responsible for the auditory-visual integration processing. The left STS is a especially intriguing region in the audiovisual processing in speech perception including the McGurk effect. On the other hand, REA is clearly related to the left auditory cortex dominance in language perception.
Our results suggest that the cortical processing resulting in the McGurk effect can co-exist with that inducing the REA. There are several possibilities. We could simply assume that the syllable sounds from both ears are delivered to the left primary auditory cortex, and then the processing induce the REA and the information of the phoneme perception is sent to the left STS and the more posterior part for interaction with the visual information. However, at present the time course of audio-visual integration for the McGurk effect in the brain is not known. Indeed, in congruent AV stimuli, audio-visual interaction in the auditory cortex is known to precede that in the multisensory STS region [18]. This result suggests the presence of early multisensory interactions. In order to clarify the exact nature of audio-visual integration in naturally and artificially occurring situations, including the DL, it is necessary to investigate in more detail the time course of cortical processing involved in the McGurk effect and REA.

Finally, the results reported in this study suggest that visual /ba/ has a strong influence to the syllable perception, with the subjects distinguishing it easier than other syllables. This result is consistent with the interpretation the mouth closing motion in the bilabial syllables provides salient visual clues.

In conclusion, we reported here that the McGurk effect is observed in the dichotic listening condition. Our results provide some constraints on the cortical mechanism of the robust audio-visual integration in a complex stimulus environment.

5. References