PROCEEDINGS OF THE ROYAL SOCIETY A

Proc. R. Soc. A (2008) **464**, 319–340
doi:10.1098/rspa.2007.1910
Published online 27 November 2007

Fusion and combination in audio-visual integration

By Kei Omata[1,2] and Ken Mogi[2,1,*]

[1]*Department of Computational Intelligence and System Science, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8502, Japan*
[2]*Sony Computer Science Laboratories, Inc., 3-14-13, Higashigotanda, Shinagawa-ku, Tokyo 141-0022, Japan*

Language is essentially multi-modal in its sensory origin, the daily conversation depending heavily on the audio-visual (AV) information. Although the perception of spoken language is primarily dominated by audition, the perception of facial expression, particularly that of the mouth, helps us comprehend speech. The McGurk effect is a striking phenomenon where the perceived phoneme is affected by the simultaneous observation of lip movement, and probably reflects the underlying AV integration process. The elucidation of the principles involved in this unique perceptual anomaly poses an interesting problem. Here we study the nature of the McGurk effect by means of neural networks (self-organizing maps, SOM) designed to extract patterns inherent in audio and visual stimuli. It is shown that a McGurk effect-like classification of incoming information occurs without any additional constraint or procedure added to the network, suggesting that the anomaly is a consequence of the AV integration process. Within this framework, an explanation is given for the asymmetric effect of AV pairs in causing the McGurk effect (fusion or combination) based on the 'distance' relationship between audio or visual information within the SOM. Our result reveals some generic features of the cognitive process of phoneme perception, and AV sensory integration in general.

Keywords: self-organizing maps; McGurk effect; audio-visual integration; phoneme perception

1. Introduction

The daily conversation, although heavily dependent on speech and auditory perception, is actually dependent on information from multiple sensory modalities, vision in particular. Humans use not only vocal information but also visually mediated information, such as gestures, eye contact and facial expressions to facilitate communication. The visual information is useful in conveying information regarding feelings and emotions. In daily conversation, seeing the mouth movements helps us recognize what is being spoken, especially during a noisy condition (e.g. Sumby & Pollack 1954).

* Author for correspondence (kenmogi@csl.sony.co.jp).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspa.2007.1910 or via http://journals.royalsociety.org.

Received 28 November 2006
Accepted 29 October 2007

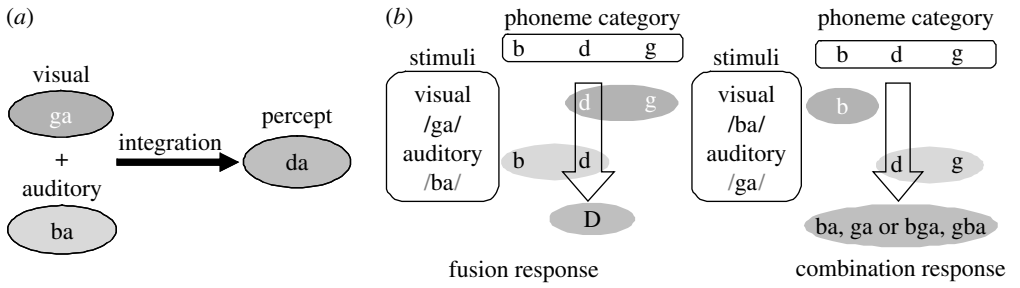

Figure 1. (*a*) Typical example of the McGurk effect. When an auditory /ba/ is dubbed with a visual /ga/, subjects usually perceive the phoneme as /da/ syllable. (*b*) The fusion and combination responses.

Recently, language processing in the context of multi-modal (in particular audio-visual, AV) sensory perception has been addressed in terms of the neural mechanisms involved (Calvert *et al.* 2000; Jones & Callan 2003; Sekiyama *et al.* 2003; Callan *et al.* 2004). These studies revealed the activation of the superior temporal sulcus and gyrus (STS/G) areas of the cortex when the AV stimuli of syllables were presented, suggesting the involvement of not only the auditory cortex but also the AV association cortex in phoneme perception. It is therefore important to investigate the nature of the AV integration process in the perception of phonemes.

The McGurk effect is a striking phenomenon under incoherent AV information, where lip-read information in speech interferes with auditory perception of phonemes (McGurk & MacDonald 1976; Sekiyama & Tohkura 1991; Munhall *et al.* 1996; Sekiyama 1997), suggesting a strong influence of vision (figure 1*a*). There are two typical responses in the McGurk effect. The AV combination of 'bilabial' (involving both lips) sounds and 'palatal' (the body of the tongue raised against the hard palate) mouth movement typically results in a 'fusion response', in which a new phoneme different from the originals is perceived. For instance, when an auditory /ba/ is dubbed to the motion picture of mouth pronouncing /ga/, the majority of subjects perceive the phoneme /da/ in adults (McGurk & MacDonald 1976). On the other hand, when an auditory /ga/ is paired with the visual articulation of /ba/ syllable, subjects basically perceive either of the phonemes /ba/ or /ga/, and sometimes report the perception of the phoneme /bga/ or /gba/ ('combination response', MacDonald & McGurk 1978; figure 1*b*). Therefore, there is a marked asymmetry as regards the audio or visual representation of a given set of phonemes in the neural mechanism leading to the McGurk effect. The asymmetric effect of AV pairs causing the fused or combination response is one of the most intriguing properties of this perceptual anomaly, and needs to be accounted for by any theories pertaining to explain this phenomenon.

The appearance of the McGurk effect may be related to the language development in general. The original study by McGurk & MacDonald (1976) showed that the preschool and school children (i.e. 3–5 and 7–8-year-old groups) reported a weaker McGurk effect than did the adults. Hockley & Polka (1994) reported the visual influence on the developmental improvement across ages in AV speech perception. As regards the differences across languages, several

studies (Sekiyama & Tohkura 1991, 1993; Sekiyama 1997) suggested differential occurrences of the McGurk effect across various languages (e.g. Japanese, American and Chinese). Sekiyama & Tohkura (1993) showed that, as a whole, the McGurk effect was weaker in Japanese subjects than in Americans.

Evidence reported in studies on the development of speech in terms of AV perception suggests that the learning process of phoneme perception in AV information might be one of the factors inducing the McGurk effect. In this respect, it is an interesting question whether the AV information plays an important role in language acquisition. Developmental studies showed deviations in the language development and/or delayed language acquisition for blind children (Mills 1983, 1988; Andersen *et al.* 1993; Warren 1994). Delayed acquisition has been reported for the production of speech sounds, the comprehension of semantic concepts including sighted terms (e.g. referential vocabulary and the lack of overgeneralizations). Studies on language developments in blind children have reported less-developed auditory perception in some cases (Mills 1988). These studies suggest that visual information accompanying speech confers a benefit for the language acquisition.

Assuming that the visual information is important in language acquisition, when does the AV integration start in the course of language development? AV matching ability in speech has been reported in 2 (Patterson & Werker 2003), 4 month old (Kuhl & Meltzoff 1982) and 2.5 to 5 month old infants (Burnham & Dodd 1998). Evidence reported in these studies suggests that infants can match features of auditory and visual speech events before they start to acquire native-language perception at the age of 6–12 months (Kuhl 2000, 2007). Some studies suggest that infants are able to integrate AV information in speech (Desjardins & Werker 1996; Rosenblum *et al.* 1997; Burnham 1998; Burnham & Dodd 2004). Among these studies, Burnham & Dodd (2004) reported on the emergent percept in the McGurk effect in a habituation-test paradigm.

In summary, studies on language development and acquisition have shown that phoneme perception is an actively learned cognitive ability dependent primarily on audition, but significantly employing AV information. The visual influence might result in the accompanying anomalies in phoneme perception as is exemplified in the McGurk effect, the exhibition of which does not require any special training. Thus, it is possible that the McGurk effect is a by-product of the language acquisition, fuelled by the exposure to coherent AV stimuli. If this hypothesis is true, there would be no special neural circuit responsible for the McGurk effect, which would arise from the inherent properties of phoneme perception system in general. It is therefore interesting to investigate whether a network trained for general phoneme categorization would exhibit a McGurk-like effect.

There have been model studies using both auditory and visual speech information in the fields of psychophysics (Robert-Ribes *et al.* 1996; Schwartz *et al.* 1998) and speech recognition (Hennecke & Stork 1996). Some models have attempted to explain many results of psychophysical data (e.g. fuzzy logic model for perception, 'FLMP' model of Massaro (1998) and Chen & Massaro (2004), and the 'TRACE' model of Campbell (1988)). Models in speech recognition have been developed from the viewpoint of engineering applications (e.g. Brooke 1998; Iwano *et al.* 2001).

When building an artificial system of speech recognition, some important design questions must be addressed, e.g. the manner in which information from AV channels are integrated. There are two important points to be clarified in constructing a model architecture. Firstly, there is the question of 'early or late' integration. Some systems integrate 'before the recognition' (early). Others integrate by evaluating probabilities separately and combining them in some manner (late). In psychological models, early integration has been considered to be more consistent with the results of behavioural experiments than late integration (Schwartz *et al.* 1998). Secondly, there is a question as to whether two auditory and visual features evaluated are transformed into a common or amodal representation before the recognition. Schwartz *et al.* (1998) argued for the necessity of a common format based on psychophysical evidence. On the other hand, Bernstein *et al.* (2004) and Bernstein (2005) disputed the existence of a common format from the viewpoint of brain science, and suggested that auditory and visual speech processing result in separate modality-specific representations, which are then linked or associated to configure the general perception.

There have been some studies of the McGurk effect based on neural network models. For instance, Renart *et al.* (1999) studied the McGurk effect in an analogue neuron model with two input modules and a bimodal module. They suggested that the activities of the bimodal module could account for the combination response of the effect. However, the model did not reproduce the fusion response, nor the fact that specific pairs of phonemes lead to either the combination or fusion responses. To the best of our knowledge, no models of phoneme perception have accounted for the basic feature of the McGurk effect in the context of neural networks.

Here we reproduce the rudimentary feature of the McGurk effect by self-organizing maps (SOM, Kohonen 1995). We started from the assumption of early integration with a common metric. We study by means of a neural network model the role of phoneme learning based on the AV information, eventually leading to the McGurk effect. In terms of the brain's network, our model is aimed to reproduce some rudimentary properties of the AV information processing in the peripheral auditory cortex including the STS/G (Calvert *et al.* 2000; Jones & Callan 2003; Sekiyama *et al.* 2003; Callan *et al.* 2004).

We extracted audio and visual properties from the actual video capturing of a human speaker, obtaining a vector representation of the features. We then constructed a neural network simulating phoneme perception using the coherent AV information (figure 2). The results of our analysis reproduce the basic features of the McGurk effect, notably the asymmetric effect of AV pairs causing the fused or combination response, and suggest some underlying properties of the sensory integration process that leads to this striking anomaly, revealing the principles of phoneme perception involved.

2. Method

(*a*) *Self-organizing maps*

We used the SOM to simulate the human phoneme perception process. SOM has been employed as an effective tool to model the organizing process of the functional maps in the brain (Kohonen 1988, 1995) and is an instance of

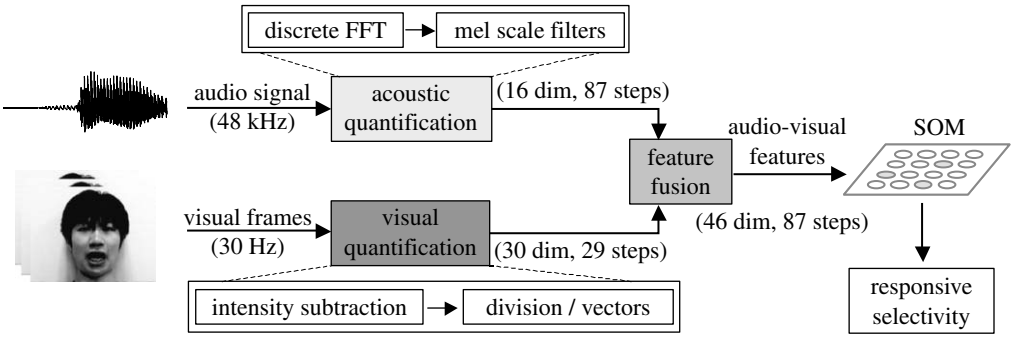

Figure 2. Production of data vectors from the video followed by their application to the self-organizing maps. dim, dimension; FFT, fast Fourier transform.

unsupervised learning systems. Anatomical, physiological and theoretical studies support the hypothesis that the cerebral cortex employs unsupervised learning (Doya 1999). The SOM model thus attempts to emulate the AV information processing of phonemes in the cerebral cortex (i.e. the auditory cortex, STS/G).

The SOM is known to be robust as regards the defects in data and is able to calculate the best-match unit for vector data including some defects. When applied to the categorization of phonemes, the SOM reproduces classification through unsupervised learning. After a training phase in which the self-organization takes place, SOM is able to categorize the data into some groups according to similarities.

In employing the SOM, it has been known that it is crucial to tune the learning parameters for the map construction and visualization. We initialized the reference vectors linearly, where the vectors on the map are initialized along the greatest eigenvectors calculated from the input data. We also employed a batch learning algorithm without a learning-rate parameter. In this treatment, the SOM has no convergence problems and yields asymptotic values for the reference vectors (Kohonen 1995). We verified that the use of linear initialization and batch algorithm did not change the basic properties of the map construction. We had essentially the same results when compared with the basic SOM algorithm although the map construction is influenced by the treatment of learning parameters and the initial state (data not shown). We used the mask function in the SOM Toolbox for AV data vectors.

(*b*) *Response selectivity*

When the SOM was trained, the vectors were scattered onto the SOM to detect the 'winner' unit. Each unit thus responded to the short-time segment of the data. Units on the map did not represent the responses for the whole sequence of a particular phoneme. For instance, a unit on the map would respond to the /b/ part of /ba/ syllable and another unit would respond to the /a/ part of the same. In general, each unit could be responsive for not only a particular part of a syllable but also similar parts of other syllables.

In order to estimate how the phonemes are categorized on the SOM, it is necessary to take an overview of the map. We estimated the responsive properties of the SOM by calculating the 'response selectivity', the ratio of the magnitude of response for the set of syllables fed into the network.

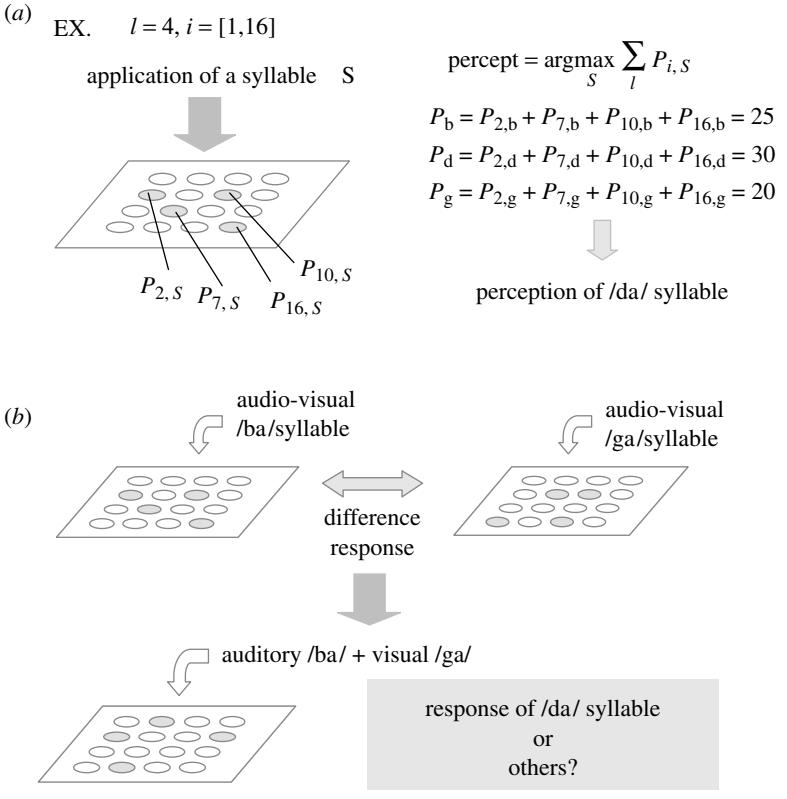

Figure 3. The response selectivity. (*a*) Application of a sample syllable S to the SOM in training. (*b*) The representation of phonemes on the self-organizing maps.

Hubel & Wiesel (1962) reported that neurons in the primary visual cortex have selectivity for direction and orientation of line segments. In the following works in neurophysiology, response selectivity has become an important concept for analysis. Let the index P denote the response selectivity. Based on P, the phoneme classification is characterized in the following way.

(i) All sample vectors of the input data are scattered to the best-matching units on the map trained by the input data.
(ii) The numbers of the responses for each phoneme in units are counted and gathered for each unit.
(iii) The ratio of numbers for each phoneme in each unit is taken to define the P value of response selectivity.
(iv) New input data are applied to the maps.
(v) The P values of the matching units for each vector of the input data are summarized.
(vi) The perceived phoneme is determined by the maximal value derived from summations of the P values for each syllable.

Let

$$x_j(1 \leq j \leq N * L),$$

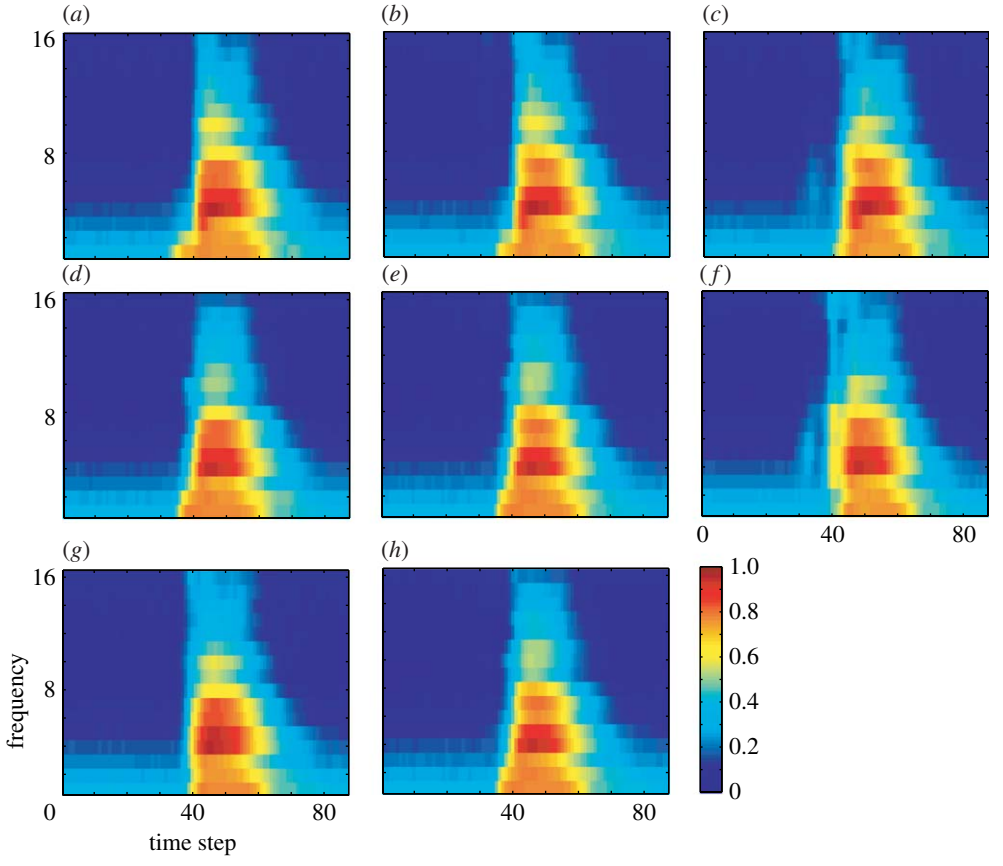

Figure 4. Average auditory vector representations of syllables extracted from video sounds. Each map consists of 16-dimensional vectors with 87 time steps (syllables (*a*) /ba/, (*b*) /da/, (*c*) /ga/, (*d*) /ma/, (*e*) /na/ (*f*) /ka/, (*g*) /pa/ and (*h*) /ta/). The value of these vectors in each dimension has been normalized to be between 0 and 1.

denote an input data vector of the phoneme S, where N and L are the number of sample data and the number of vectors for each phoneme, respectively. The 'winner' unit c_j is chosen for each vector as

$$c_j = \underset{i}{\operatorname{argmin}}\{\|x_j - m_i\|\},$$

where

$$m_j (1 \leq j \leq M),$$

is the reference vector of each unit on the map. M represents the map size. The number of c_j for each unit i is then counted and divided by NL. The procedure is repeated NL times.

$$P_{i,S} = \frac{p_{i,S}}{\sum_S p_{i,S}} \quad (S = /b, d, g/\text{or}/m, n, k/\text{or}/p, t, k/),$$

then represents the response selectivity of each unit i for phoneme. The P values are here normalized to be between 0 and 1. The contribution of each unit for the final perception of a phoneme is assumed to be equal.

When the new set of sample data $y_l (1 \leq l \leq L)$ of a syllable is applied to the map, the best-match unit c is chosen by the formula below,

$$c_l = \underset{i}{\operatorname{argmin}}\{\|y_l - m_i\|\}.$$

$P_{c_l,S}$ for unit c_l is then calculated. The procedure is repeated L times, after which all the $P_{c_l},_S$ are summed.

Finally, the percept represented by the network is obtained as

$$\text{percept} = \underset{S}{\operatorname{argmax}} \sum_l P_{i,S}.$$

Thus, the phoneme represented by the network is determined by the relative values of $P_{i,S}$ for each phoneme (figure 3*a*).

(*c*) *Extraction of feature vectors from the video*

AV data were recorded by a digital video camera (Sony DCR-TRV30) and a condenser microphone (Sony ECM-MS907). These data were edited by a software (Premiere v. 6.0) on a PC (IBM Think pad X31). Matlab was used to process all the data, with the SOM Toolbox used for the core of basic SOM algorithm (Copyright (C) 2000–2005 by Esa Alhoniemi, Johan Himberg, Juha Parhankangas and Juha Vesanto, under the terms of the GNU General Public License.)

All auditory and visual stimuli were recorded from a male native Japanese speaker. By that time, 30 blocks of eight syllables, /ba/, /da/, /ga/, /ma/, /na/, /pa/, /ta/ and /ka/ were uttered and recorded. The visual and auditory tracks were then aligned automatically. We first cut a segment of 1 s (30 frames). Two criteria were used to balance the timing in view of the different VOT (voice onset time) and speech lip movements in the syllables.

(i) The burst point of utterance was detected as a mark point after the onset of VOT in the auditory track. As the segment to be used for analysis, 13 frames before and 17 frames after the mark point were then taken.
(ii) The segments were further adjusted so that the 14th frame in the image track includes a face image with the mouth open or in the process of opening.

We used the Fourier spectrograms of sound signals as the information of auditory stimuli and referred to the TDNN model (Waibel *et al.* 1989) for the pretreatment in processing the vectors.

The sampling rate was 48 000 Hz. The quantization was 16 bits and the stereo voice sounds were transformed into monaural sounds on the PC. Discrete Fourier transform was then applied to create the spectrograms. We used 16 mel scale spectral coefficients for the input vectors (Waibel & Yegnanarayana 1981; Waibel *et al.* 1989). The mel scale coefficients were computed from the power spectrum by computing the log energies in each mel scale energy band (Rabiner & Juang 1993), the mel scale filter in the band being a window of triangle shape. The mel scale energy bands ranged from 70 Hz to 6 kHz, i.e. effectively equivalent to a low-pass filter. As previously mentioned, the duration of auditory signal was 1 s. Typically, vocalization started at approximately 400 ms after the onset and continued for approximately 500 ms. The sound signal was Hanning windowed and the discrete fast Fourier transform was computed every 11.1 ms, the window length being

42.7 ms. Therefore, the data of one syllable have 87 time steps on account of the relationship between the window size and the length of sound signals (figure 4). Although the signals processed were varying in terms of the length, pitch and loudness, modification was not applied to these features when the signal was transformed into vectors. The vectors were normalized between 0 and 1 in the preparation for application to the SOM.

We quantified the image data into vectors by subtracting the grey colour intensities between frames. The visual signal was extracted from the same video recording as was used for the auditory signal. The original images in the video had a colour gradation of 8 bits, with 320×240 pixels resolution. Each frame was converted into monochrome by Premiere v. 6.0. The absolute values of the difference in the intensities of pixels were then calculated. The images were divided into 30 areas horizontally. The magnitudes of difference in each area were summed to create 30-dimensional vectors (figure 5*a*). Here, the horizontal division was chosen rather than the vertical division as the facial movement is larger in the vertical direction. The resolution of the horizontal division was determined after preliminary investigations to sufficiently represent the differential facial motion in the utterance of syllables (figure 5*b*).

As already noted, we have assumed here an early integration of AV information, presumably to occur within the peripheral auditory cortex including the STS/G in the brain. These regions have neurons that respond to both auditory and visual stimuli (e.g. Beauchamp *et al.* 2004; Ghazanfar & Schroeder 2006). Ghazanfar *et al.* (2005) reported in a monkey study that enhanced responses of neurons in the STS occurred for shorter voice-onset time, and that suppressed responses occurred when voice-onset time was longer, suggesting that the time-synchronous information received from two modalities are processed. Consequently, we concatenated the time-synchronous audio and visual features in the training of the SOM for phoneme categorization.

In the perception of phonemes, in that of consonants in particular, human subjects are sensitive to transients of sound frequencies of short duration. Boemio *et al.* (2005) showed in an fMRI study that the left superior temporal sulcus in the brain is sensitive to the short-time transient stimuli of sounds.

We combined the auditory and visual data along the time sequence. Auditory data had 16-dimensional vectors with the time step of 11 ms. Visual data had 30-dimensional vectors with the time step of 33 ms. We aligned the auditory and visual data along the time sequence and combined auditory data with visual data to produce 46-dimensional vectors. Since the sampling rate of visual data was one-third that of auditory data, the visual data was aligned in increments of three steps (figure 6).

(*d*) *Procedure of training and presentation of data to the SOM*

We constructed three groups of phonemes (/b,d,g/, /m,n,k/ and /p,t,k/) from eight syllables. The syllables in these groups have been known as typical inducer pairs and the resulting perceived phonemes in the McGurk effect (MacDonald & McGurk 1978), where presentation of bilabial sounds (/b/, /m/, /p/) dubbed with the movies of facial motion vocalizing palatal sounds (/g/, /k/), result in the perception of alveolar phonemes (/d/, /n/, /t/).

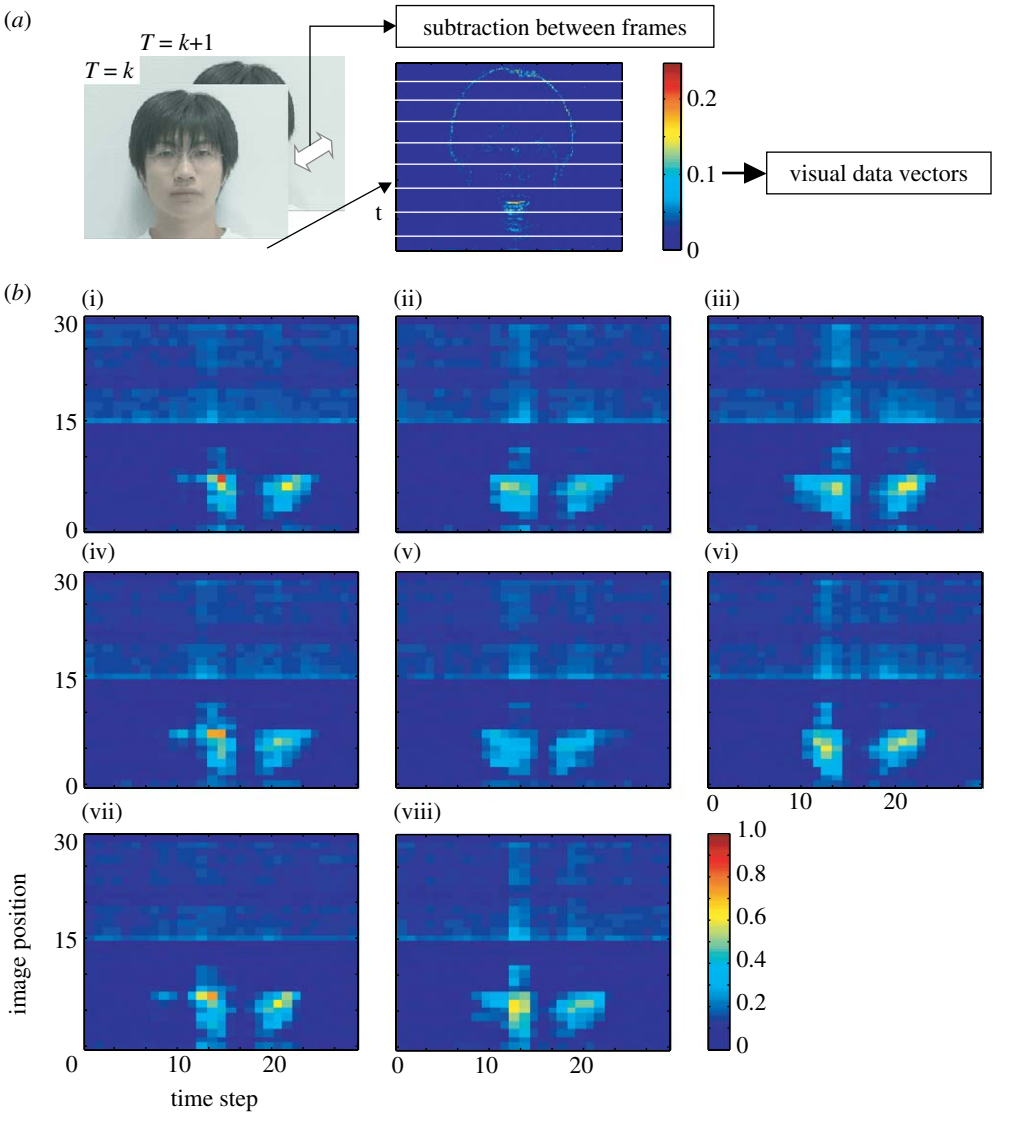

Figure 5. (*a*) Subtraction between the frames. Visual data were created by a horizontal division of the images. (*b*) Average visual data vector representations of the syllables, which consisted of parameters of 30 dimensions extending over 29 time steps (visual (i) /ba/, (ii) /da/, (iii) /ga/, (iv) /ma/, (v) /na/, (vi) /ka/, (vii) /pa/ and (viii) /ta/). The vertical axis represents the image position, where the lower part corresponds to the speaker's lower face position including the jaw and the mouth, and the upper part represents the forehead and eyes.

We applied the data to SOM, which had 30×28 units (figure 7). There were three conditions in stimuli presentation for the SOM. In the AV condition, the visual as well as audio information was fed into the system. In the visual masking (AV (visual masking)) condition, visual vector data were masked out when the best-match units were calculated on the SOM. In the auditory masking (AV (auditory masking)) condition, the auditory vector was masked out. In addition to these conditions, the

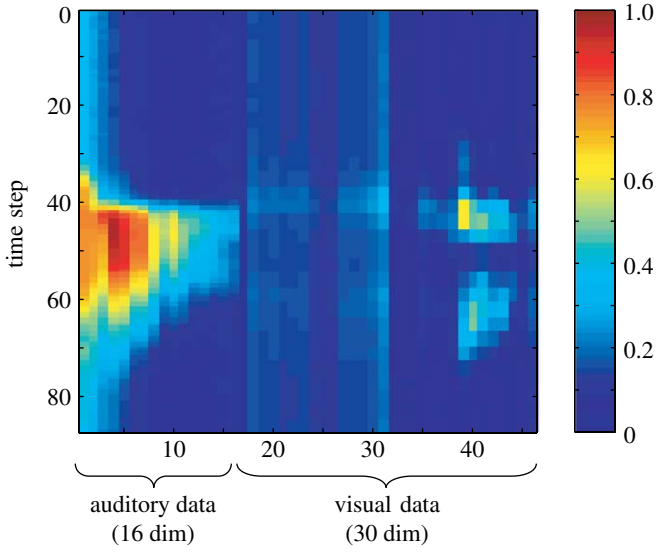

Figure 6. AV vector representation of a syllable (that for /ba/ shown here). dim, dimension.

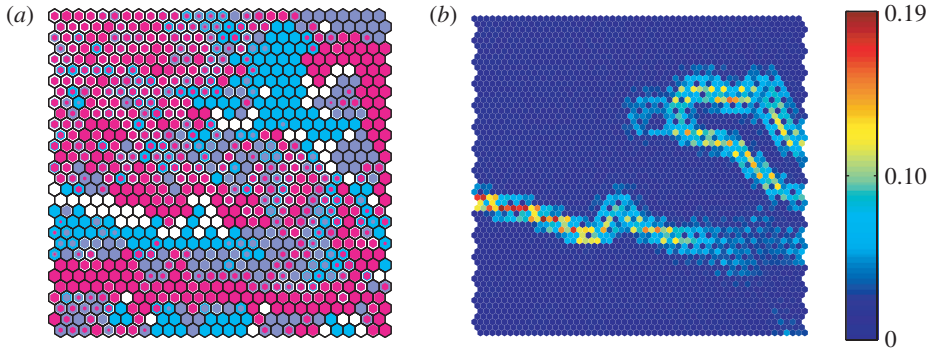

Figure 7. (*a*) Responses of the units for three phonemes (/b,d,g/). The hexagonal lattice denotes units in self-organizing maps. The colour of the hexagons on the map represents the responses to the syllables, the diameters of an area of a particular colour indicating the ratios of the magnitude of the responses to the syllable in question. Light blue, the response for /ba/; purple, the response for /da/; red, that for /ga/. (*b*) The *U*-matrix maps on the SOM created by the SOM toolbox.

SOMs were separately trained in auditory-only (A-only) condition and visual-only (V-only) condition. The performances of the SOM trained in all conditions were then examined by means of fivefold cross validation with 90 sample data (30 samples each for three phonemes).

In addition, in order to understand how the McGurk effect is induced, we trained and tested the SOM in three conditions. In simulating a situation closer to what actually happens in natural language acquisition and perception, we fed the incoherent AV data expected to cause the McGurk effect to the SOM trained in the AV (condition 1). Even in a trained network, it is an open question whether the simultaneous presentation of AV stimuli is essential in the induction of the McGurk effect or not. To check this point, we presented the auditory and visual data separately to the SOM trained by AV data

(condition 2). We then calculated the P values by using the masked input. When the P values were calculated, we fed the one-modality masked data to the SOM and obtained the P values for each modality data. The P values were then normalized to equally incorporate the contribution of each modality into phoneme perception. The final P value was then defined to be the sum of P values for each modality. Finally, in condition 3, two separate SOMs were trained by auditory and visual data separately in order to test the validity of the late-integration hypothesis. The whole P values of responsive selectivity were then calculated by summing the two P values normalized to equalize the contribution of the two SOMs.

In addition, the whole P values in the conditions 2 and 3 were also calculated by means of multiplying the P values from auditory and visual data presentation in order to compare them with results from previous models (e.g. FLMP, Massaro 1998). This calculation is followed by the equation,

$$P_{\mathrm{AV}}(C_i) = p_{\mathrm{A}}(C_i)p_{\mathrm{V}}(C_i)/\sum_j p_{\mathrm{A}}(C_j)p_{\mathrm{V}}(C_j).$$

In the above equation, C_i and C_j are phonetic categories in this model; P_{AV} are the model probability of phonemes responses for the each category C_i; and p_{A} and p_{V} indicate the P values on the SOM for auditory and visual stimuli.

3. Results

In all three typical pairs inducing the McGurk effect, evaluation of the performance of the trained SOM by means of fivefold cross validation in the AV, the AV (visual masking) and the AV (auditory masking) conditions (table 1*a*) showed that the performance in the AV condition was significantly higher than that in other conditions. In other words, the AV presentation made the phoneme perception more robust. This result is consistent with the study by Sumby & Pollack (1954). On one hand, the correlation between the total performance of the SOM in the AV (visual masking) and in the AV (auditory making) condition is not consistent among the phoneme groups studied, suggesting that neither the auditory nor visual information is dominant in the determination of the phoneme categorization in the SOM trained by AV information.

The performance in the AV condition was higher than that of the A- or the V-only condition. Thus, the AV information is more robust in estimating the phonemes than the unisensory information. Among the single modality data, the performance of SOM in the A-only condition was higher than that in the V-only condition in all three phoneme groups.

In condition 1, we applied AV pairs expected to cause fused response in the McGurk effect to the SOM trained by all AV data. When the typical fusion pairs (auditory /ba/ paired with visual /ga/, auditory /ma/ paired with visual /ka/ and auditory /pa/ paired with visual /ka/) were presented, the ratios of units representing the fused phonemes were 46.9, 34.4 and 49.3% for /da/, /na/ and /ta/, respectively. Conversely, when the combination pairs (auditory /ga/ paired with visual /ba/, auditory /ka/ paired with visual /ma/ and auditory /ka/ paired with visual /pa/) were applied, the ratios of units representing the fused

Table 1. (*a*) The performance of the classifier for three groups of inducing pairs and the perceived phoneme (/b,d,g/, /m,n,k/, /p,t,k/) causing the typical McGurk effect. AV, both the audio and the visual data vectors are available for searching best-match units on the SOM; AV (visual masking), visual data vectors are unavailable; AV (auditory masking), auditory data vectors are unavailable; A only, the SOM was trained by the auditory data only; V only, the SOM was trained by the visual data only. (*b*) The application of the AV combinations expected to cause the McGurk effect. Condition 1: the SOM is trained with the AV data. After the training, the AV data are presented to the SOM simultaneously. Condition 2: the SOM is trained by the AV data. After the training, the audio and visual data are presented to the SOM separately. Condition 3: two SOMs are separately created for the audio and visual data. (*c*) The whole P values in condition 2 and condition 3 were calculated by multiplying the P values from two separate unimodalities.

(*a*)

the performance of classifier (fivefold cross validation) (%)	/b,d,g/	/m,n,k/	/p,t,k/
AV	96	98	94
AV (visual masking)	88	90	71
AV (auditory masking)	89	78	77
A only	84	92	83.0
V only	81	78	72

(*b*)

McGurk effect (%)	presentation pair	/b,d,g/			/m,n,k/			/p,t,k/		
		/b/	/d/	/g/	/m/	/n/	/k/	/p/	/t/	/k/
condition 1	fusion	20.9	46.9	32.2	23.4	34.4	42.1	9.6	49.3	41.1
	combination	52.2	2.3	45.4	47.0	0.6	52.4	83.0	1.2	15.8
condition 2	fusion	21.2	4.9	73.9	38.8	2.9	58.3	19.8	7.7	72.6
	combination	61.8	0.2	38.0	38.1	0.0	61.9	80.7	0.4	18.9
condition 3	fusion	0.4	0.0	99.6	5.9	0.0	94.1	0.4	0.0	99.6
	combination	90.3	0.0	9.7	89.0	0.0	11.0	95.6	0.0	4.3

(*c*)

		/b/	/d/	/g/	/m/	/n/	/k/	/p/	/t/	/k/
condition 2 (multiply)	fusion	23.6	7.4	69.0	43.2	5.1	51.7	19.1	11.3	69.6
	combination	62.0	0.7	36.9	40.6	0.2	59.2	81.1	1.0	17.9
condition 3 (multiply)	fusion	6.0	0.0	94.0	18.6	0.0	81.4	0.6	0.2	99.2
	combination	86.8	0.0	13.2	87.0	0.0	13.0	95.6	0.0	4.3

phonemes were 2.3, 0.6 and 1.2% for /da/, /na/ and /ta/, respectively (upper section of table 1*b*). These results reproduce the asymmetric effects in the induction of the McGurk effect.

When the audio and visual data were presented independently to the SOM trained by AV data (condition 2), ratios of units representing the fused phonemes were 4.9, 2.9 and 7.7% for /da/, /na/ and /ta/, respectively, in the presence of the fused pairs. In the presence of combination pairs, the ratios were 0.2, 0 and 0.4%.

We applied auditory and visual data to the SOMs trained by auditory and visual data separately (condition 3). We then combined the P values from each

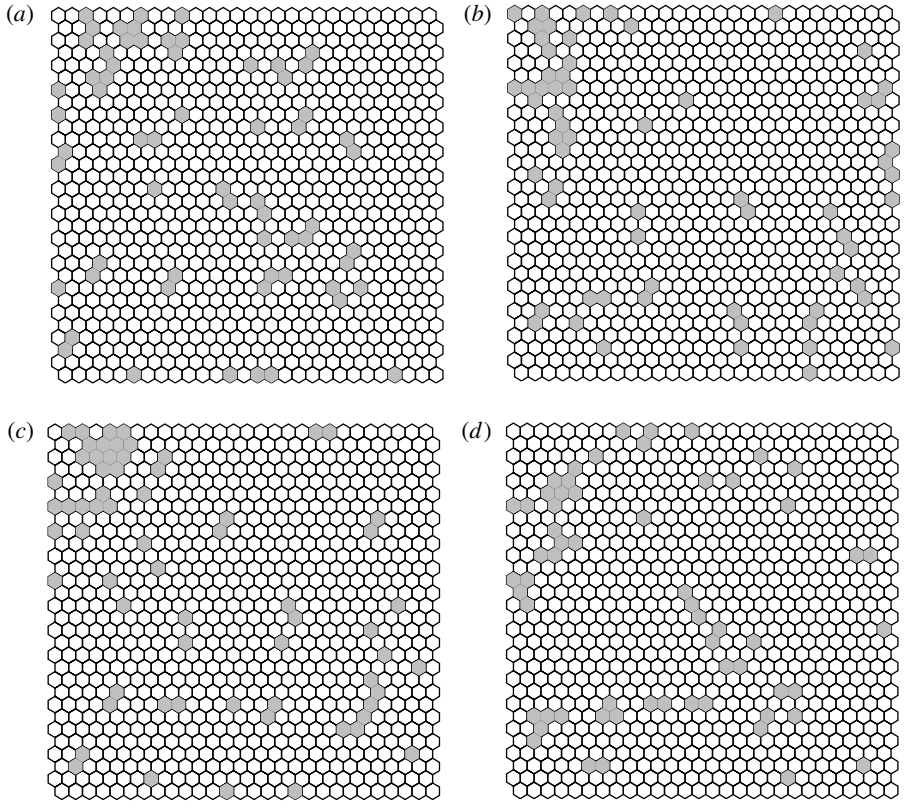

Figure 8. SOM responses for various combinations of AV /ba/ and /ga/. The maximal values of the summarized P values predict the resulting perception of the syllable in humans. (*a*) Response to AV /ba/ (P_b, P_d, P_g, 44.6, 21.9, 20.6, respectively). (*b*) Response to AV /ga/ (P_b, P_d, P_g, 16.3, 21.2, 49.5, respectively) (*c*) Response to the combination of auditory /ba/ and visual /ga/ (fusion pair; P_b, P_d, P_g, 27.2, 35.6, 24.2, respectively). (*d*) Response to the combination of auditory /ga/ and visual /ba/ (combination pair; P_b, P_d, P_g, 30.6, 27.9, 28.4, respectively).

modality to create the classification profile in the presence of fusion and combination pairs. The ratio of phoneme perception was approximately 0% for all fused alveolar syllables, resulting in a null effect (lower section of table 1*b*).

The alternative methods of P values calculation either by taking the sum or by multiplication revealed that the results were not significantly different (table 1*b*,*c*).

Figure 8 shows the responses of the SOM to various combinations of auditory and visual signals. Response to AV /ba/ (figure 8*a*), AV /ga/ (figure 8*b*), the combination of auditory /ba/ and visual /ga/ (fusion pair; figure 8*c*) and the combination of auditory /ga/ and visual /ba/ (combination pair; figure 8*d*). Note that the units corresponding to incoherent AV pairs are different from those corresponding to coherent AV pairs, suggesting that new patterns of activity were induced in the SOM by the incoherent AV pairs.

In the results reported above, the SOM exhibited activities consistent with the induction of a McGurk-like effect. In particular, the asymmetric effect of combination and fused pairs has been reproduced. It is interesting to ask how this

Table 2. Similarity indexes for three groups (/b,d,g/, /m,n,k/, /p,t,k/).

similarity index	/b,d,g/			/m,n,k/			/p,t,k/		
$f(S_i, S_j)$	(/b/,/d/)	(/d/,/g/)	(/g/,/b/)	(/m/,/n/)	(/n/,/k/)	(/k/,/m/)	(/p/,/t/)	(/t/,/k/)	(/k/,/p/)
f	47.9	78.0	88.3	48.8	67.5	85.1	61.7	65.9	69.9
f_A	13.3	48.1	47.2	14.5	42.4	42.4	25.2	45.3	31.2
f_V	34.6	29.8	41.1	34.3	25.2	42.7	36.5	20.6	38.7

asymmetry arose. Here we analyse the origin of the asymmetric response of the SOM to fusion pairs and combination pairs in the modelled McGurk effect.

To facilitate analysis, we define an index of 'similarity' for the syllable data. The similarity index f is defined as

$$f(S_i, S_j) = \sum_{n}^{N}\sum_{m}^{M}|x_{S_i}(n, m) - x_{S_j}(n, m)|$$

$$f(S_i, S_j) = f(S_j, S_i),$$

where S_i is a phoneme; x is a matrix composed of $N=16$ dimensions; and $M=87$ time steps.

If the value of similarity index is small, the two syllables are similar to each other. For example, the value of $f(/b/, /d/)$ was 47.9, whereas the value of $f(/g/, /b/)$ was 88.3 (table 2). Therefore, the averaged /ba/ signal was more similar to that of averaged /da/ than the averaged /ga/.

The total f value is a composite of two parts

$$f(S_i, S_j) = f_A(S_i, S_j) + f_V(S_i, S_j),$$

where f_A is the auditory part and f_V is the visual part. When calculated from the data, we found the following relationships between them (table 2):

$$f_A(/b/, /d/) < f_A(/d/, /g/), \quad f_V(/b/, /d/) > f_V(/d/, /g/),$$
$$f_A(/m/, /n/) < f_A(/n/, /k/), \quad f_V(/m/, /n/) > f_V(/n/, /k/),$$
$$f_A(/p/, /t/) < f_A(/t/, /k/), \quad f_V(/p/, /t/) > f_V(/t/, /k/).$$

As regards the auditory part, f_A $(/b/, /d/)$ is smaller than f_A $(/d/, /g/)$. On the other hand, f_V $(/b/, /d/)$ is larger than f_V $(/d/, /g/)$, i.e. the averaged /da/ is more similar to /ga/ than /ba/ in the visual part.

This asymmetry of the distance relationship between the auditory and visual signals of the three phonemes is consistent with the occurrence of the fusion and combination responses in the McGurk effect.

When a fusion pair (auditory /ba/ and visual /ga/) is presented, the perception of /da/ is reasonable from the point of view of information similarity, as the distance between auditory /ba/ and auditory /da/, as well as visual /ga/ and visual /da/ is small. On the other hand, when a combination pair (auditory /ga/ and visual /ba/) is presented, the perception of the fused phoneme /da/ is less likely to occur, as the distance between auditory /ga/ and auditory /da/,

and that between visual /da/ and visual /ba/ are larger than that in the case of fusion pairs.

The same can be said for the fusion and combination pairs of /m,n,k/ and /p,t,k/ phoneme groups.

4. Discussion

In this paper, we addressed the nature and origin of phoneme perception as it emerges in the McGurk effect, and revealed some features of this perceptual anomaly and categorization in multi-modal (in this case AV) sensory integration in general. The performance of SOM studied here has suggested that the McGurk effect, although a striking anomaly in itself, is actually a consequence of the fact that the phoneme classification is acquired in the context of AV speech information.

One important issue that emerges is the role of AV integration in phoneme perception. The performances of the A-only condition were more than 80% across three groups of phonemes (table 1*a*), suggesting that auditory information alone may be sufficient in learning the phoneme perception. However, the performances of the AV condition were higher than that of the A-only condition (table 1*a*). Thus, although it is possible to establish phoneme categorization without visual information, there might be an advantage of the AV integration in language acquisition.

Blind children, compared with sighted children, are at a disadvantage in language acquisition (Mills 1983, 1988; Andersen *et al.* 1993; Warren 1994). Infants have been shown to use auditory–visual modalities in language development (e.g. Burnham 1998; Kuhl 2000). These sets of evidence suggest that visual information is helpful in the acquisition of language, supported by the association of the audio and visual information in speech. The McGurk effect, in view of the importance of AV integration in the cognition of language, can be regarded as a consequence of the set-up of the cortical network to address the problem of robustness in speech perception, which needs to be carried out in the daily environment full of noise.

Norrix *et al.* (2006) reported that the McGurk effect occurred less for adults with language-learning disabilities (LLD), who showed a poor integration for auditory–visual speech. Hamilton *et al.* (2006) reported that a patient who had an acquired deficit of auditory–visual speech integration showed no McGurk effect. The differential modes of AV integration in various languages might result in varied occurrence of the McGurk effect. Indeed, Sekiyama & Tohkura (1991, 1993) reported that the occurrence of the McGurk effect was less for Japanese subjects than for English-speaking ones. This finding has been suggested to be explained by the social fact that the Japanese are less likely to look at the face of the speaker in daily conversation.

The importance of AV integration is related to the process of speech articulation, possibly through the significance of the observation of speech production activating the mirror system (Rizzolatti & Craighero 2004). There has been a discussion on the age-related increase of the McGurk effect (McGurk *et al.* 1976; Hockley & Polka 1994). It has been suggested that the acquisition of the ability of AV information integration in speech intensifies the employment of visual information in phoneme perception. Burnham & Dodd (2004) pointed out

that the developmental increase of the McGurk effect is related to the articulation experience, which in turn correlates with the use of visual information in speech perception. Liberman & Mattingly (1985) proposed the *motor theory* of speech perception, pointing out the importance of act of speech in the perception of phonemes. Recently, Sams *et al.* (2005) reported that a small degree of McGurk effect was observed when the subjects just silently articulated the syllables without any visual feedback. These studies suggest that there is a close relationship between the perception and production of speech, including the McGurk effect.

It is interesting to ask whether the McGurk effect is acquired in the process of learning, or predestined to some degree. The results of the SOM model are consistent with the idea that the learning of AV speech perception plays an important role in the process leading to the McGurk effect. On the other hand, there are evidences which suggest that cognitive processes causing the McGurk effect start quite early in development. Infants exhibit behaviours consistent with integrated perception of phonemes as is observed in the McGurk effect (Rosenblum *et al.* 1997; Burnham 1998; Burnham & Dodd 2004), 4 month olds apparently exhibiting McGurk effect-like perception (Desjardins & Werker 1996). Two month old infants already have AV matching ability in speech (Patterson & Werker 2003). Thus, the probable scenario at present is that the learning of AV speech perception does not generate the McGurk effect de novo, but nurtures it.

It is an open question what relative advantages an early or late-integration scheme holds. Schwartz *et al.* (1998) suggested from the psychophysical evidences (e.g. speech rate) that AV fusion must be realized at an early stage of processing. Hennecke & Stork (1996), on the other hand, stated that experiments using HMMs generally show that a late integration results in better performance. In our simulation, we assumed a mechanism of 'early integration', where the AV integration starts before the final recognition step. The comparison between conditions 1 and 3 (table 1*b*) reveals that the early integration leads to a more robust induction of the McGurk effect. The existence of the McGurk effect seems to be consistent with an early integration, at least in within constraints employed in our present model.

Some psychophysical models of speech perception have been proposed, e.g. the FLMP (Massaro & Cohen 1983; Massaro 1998; Chen & Massaro 2004) and the TRACE (McClelland & Elman 1986; Campbell 1988) models. The present SOM model resembles Massaro's FLMP at the conceptual level. Both models posit a common format in the AV integration (cf. Bernstein *et al.* 2004), and the prototypes of phonemes are important in both models in determining the phoneme perception. On the other hand, there are some architectural differences. The FLMP assumes that the prototypes are built in the model as memory. Conversely, the SOM model organizes the phoneme prototypes by learning the AV speech data constructed by video images. In addition, FLMP assumes a late integration whereas the SOM model is based on an early integration.

The TRACE model can account for the fusion response in the McGurk effect (Campbell 1988). This is an interactive model, where information processing is executed by means of mutually excitatory and inhibition connections on a number of simple processing units. In this model, auditory and visual information are translated into the feature spaces of seven and one dimensions, respectively.

The dimensions represent generic features (e.g. power, vocalic, diffuse, mouth open). Each dimension is treated in terms of a continuous variable that ranges from one to eight, in a manner similar to vectors in the SOM models. The explanation for the fusion and combination responses in TRACE invokes the similarities between the features of phonemes and the asymmetries between modalities, sharing some properties with the present analysis of the SOM model. The differences between the two models can be found in the attributes of vectors and the architecture. Vectors of TRACE are simplified, i.e. the visual information being represented by just one dimension (the mouth open or closed). In addition, TRACE is not a neural network model and has no common format for the AV information (cf. Schwartz *et al.* 1998).

The results of our simulation showed McGurk-like responses on the SOM trained by AV information. In order to put our results in the empirical context, it is helpful to compare the degrees of the occurrence of the McGurk-like responses in our model with that from the psychophysical data. The reported occurrences of the McGurk effect in the psychophysical data range from approximately 20 to 98% (e.g. McGurk *et al.* 1976; Sekiyama & Tohkura 1991; Burnham 1998), depending on subjects and stimuli. The occurrence of McGurk-like responses in the SOM model (e.g. 46.9, 34.4, 49.3%, table 1*b*), while not insignificant, stops short of reproducing the most robust range of empirical results.

It is worth noting, in comparing with the empirical data, several circumstances possibly affecting the figures. The actual subject's behaviour is a product of many complex factors affecting cortical processing (e.g. top-down effects, attention), some of which are not taken into consideration in our present model. For one thing, the model does not have any top-down mechanism. The resulting response is always the same for a particular pair of AV stimuli, whereas in human subjects the response can be different from time to time. In order to account for both the variability and robustness of the responses of human subjects, more factors need to be incorporated into the model.

Time is an important element in AV integration. Green & Kuhl (1991) reported in a psychophysical study that the response time (RT) for incoherent AV stimuli (e.g. a visual /ga/ paired with a auditory /ba/, resulting in the perception of the fused phoneme /da/) was longer than that of the coherent stimuli (an AV /ba/ syllable). This result suggests that the process of perception generated by the input of incoherent AV stimuli is slightly different from that of the coherent AV stimuli. Some studies reported that the McGurk effect is robust in terms of the deviation of presentation time. The acceptable range of time deviation in causing the McGurk effect is from −60 to +180 ms (Munhall *et al.* 1996) and from −30 to +170 ms (Van Wassenhove *et al.* 2007), negative values indicating that the sound precedes the vision.

In order to investigate how our model behaves with respect to time, we conducted additional simulations involving temporal deviations of the stimuli. The results (data not shown) indicated that the model is able to reproduce the fusion responses within the temporal deviation of approximately −40 to +40 ms, consistent with the general tendencies of the physiological data. The model in its present form, however, is not able to reproduce the temporally asymmetric occurrence of fusion responses when the sound lags behind vision by a larger delay, suggesting the need for additional mechanisms to account for the brain's robust performance under time deviation, e.g. some forms of audio and

visual working memory, especially that for vision (Baddeley & Della Sala 1996; Baddeley *et al.* 1998).

Brain imaging studies have suggested phoneme perception in peripheral auditory cortex (BA 41, 42) and the STS/G (BA 21, 22, 37; Calvert *et al.* 2000; Jones & Callan 2003; Sekiyama *et al.* 2003; Callan *et al.* 2004; Omata & Mogi 2005*a*,*b*). Auditory–visual neurons have been reported in the cortex (Beauchamp *et al.* 2004; Ghazanfar *et al.* 2005; Ghazanfar & Schroeder 2006). Pekkola *et al.* (2005) reported in an fMRI study for humans that visual speech perception activated Heschl's gyri and the primary auditory cortex. The occurrence of the McGurk effect has been correlated with activities in the posterior part of the left superior temporal sulcus (Sekiyama *et al.* 2003). These results reveal the cortical circuits involved in the occurrence of McGurk effect, and AV integration in general. In our present study, we assumed that the SOM model would correspond to cortical areas such as the peripheral auditory cortex and the STS/G, in which the neurons respond to both the auditory and visual information are conveyed from each unimodality. Although the properties of the response of units found in the simulation is consistent with the activities of neurons in these areas, further specifications must be incorporated in order to facilitate more direct and detailed comparisons.

One of the issues that could possible arise when making the model more realistic is scalability. Would the model behave in a reasonable manner when more phonemes are made to be represented in the network? In order to address this problem, we conducted a preliminary experiment to investigate whether the McGurk-like responses are maintained in a larger map including all the eight phonemes used in the original simulation. The results showed that the McGurk-like responses were retained in the larger map, although the degrees of the McGurk effect occurrence were smaller (in the 20–30% range) than in the case of 'smaller' map pertaining to the specific set of phonemes.

In the current simulation, we assumed a common format of AV processing. It is an open question whether the AV speech processing is explained by a common format or by a modality-specific theory. Bernstein *et al.* (2004) claimed that the common format is inadequate to account for the constraints in the brain, and that auditory and visual speech processing must result in separate modality-specific representations, which are then associated by dynamically created assemblies. Our results suggest that a common format is adequate to account for the occurrence of the McGurk effect-like responses.

In this paper, we have demonstrated that the SOM can reproduce some generic features of phoneme classification, including anomalies such as the McGurk effect. The results of our simulation indicate that the learning of AV speech is an inducing factor of the McGurk effect. In this respect, it might be considered that the McGurk illusion is a 'natural by-product' of phoneme categorization based on AV integration, although the specific dependence on the learning process is not clear at present. It is to be noted that our simulation has suggested that the asymmetry between the 'fusion' and 'combination' pairs in the McGurk effect is due to the inherent asymmetry in the similarity of the signals embedded in the audio and visual make-up of the phonemes. The detailed features of the McGurk effect, in the view suggested by the present study, are not ad hoc artefacts of the cortical network involved, but are the explicit representations of the nature of the signals to be classified by the network.

Note. The video clips used for analysis in this paper are available online as electronic supplementary material.

We thank H. Ito, A. Onzo, Y. Sekine, T. Sudo and T. Yanagawa, and the anonymous referees for their valuable comments on the manuscript.

References

Andersen, E., Dunlea, A. & Kekeli, L. 1993 The impact of input: language acquisition in the visually impaired. *First Lang.* **13**, 23–49. (doi:10.1177/014272379301303703)

Baddeley, A. D. & Della Sala, S. 1996 Working memory and executive control. *Phil. Trans. R. Soc. A* **351**, 1397–1404. (doi:10.1098/rstb.1996.0123)

Baddeley, A. D., Gathercole, S. E. & Papagno, C. 1998 The phonological loop as a language learning device. *Psychol. Rev.* **105**, 158–173. (doi:10.1037/0033-295X.105.1.158)

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H. & Martin, A. 2004 Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat. Neurosci.* **7**, 1190–1192. (doi:10.1038/nn1333)

Bernstein, L. E. 2005 Some principles of the speech perceiving brain. In *Handbook of speech perception* (eds D. B. Pisoni & R. E. Remez), pp. 79–98. London, UK: Blackwell.

Bernstein, L. E., Auer Jr, E. T. & Moore, J. K. 2004 Audiovisual speech binding: convergence or association? In *Handbook of multisensory processing* (eds G. Calvert, C. Spence & B. E. Stein), pp. 203–223. Cambridge, MA: MIT Press.

Boemio, A., Fromm, S., Braun, A. & Poeppel, D. 2005 Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat. Neurosci.* **8**, 389–395. (doi:10.1038/nn1409)

Brooke, M. N. 1998 Computational aspects of visual speech: machines that can speechread and simulate talking faces. In *Hearing by eye II* (eds R. Campbell, B. Dodd & D. Burnham), ch. 5, pp. 109–122. East Sussex, UK: Psychology Press.

Burnham, D. 1998 Language specificity in the development of auditory–visual speech perception. In *Hearing by eye II* (eds R. Campbell, B. Dodd & D. Burnham), ch. 2, pp. 27–60. East Sussex, UK: Psychology Press.

Burnham, D. & Dodd, B. 1998 Familiarity and novelty in infant cross-language studies: factors, problems, and a possible solution. *Adv. Infancy Res.* **12**, 170–187.

Burnham, D. & Dodd, B. 2004 Auditory–visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Dev. Psychobiol.* **45**, 204–220. (doi:10.1002/dev.20032)

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M. & Vatikiotis-Bateson, E. 2004 Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *J. Cognit. Neurosci.* **16**, 805–816. (doi:10.1162/089892904970771)

Calvert, G. A., Campbell, R. & Brammer, M. J. 2000 Evidence from function magnetic resonance imaging of crossmodal bindings in the human heteromodal cortex. *Curr. Biol.* **10**, 649–657. (doi:10.1016/S0960-9822(00)00513-3)

Campbell, R. 1988 Tracing lip movements: making speech visible. *Visible Lang.* **22**, 32–57.

Chen, T. H. & Massaro, D. W. 2004 Mandarin speech perception by ear and eye follows a universal principle. *Percept. Psychophys.* **66**, 820–836.

Desjardins, R. N. & Werker, J. F. 1996 4-month-old female infants are influenced by visible speech. *Poster presented at the Int. Conf. of Infant Studies, Providence RI.*

Doya, K. 1999 What are the computations of the cerebllum, the basal ganglia and the cerebral cortex? *Neural Netw.* **12**, 961–974. (doi:10.1016/S0893-6080(99)00046-5)

Ghazanfar, A. A. & Schroeder, C. E. 2006 Is neocortex essentially multisensory? *Trends Cognit. Sci.* **10**, 278–285. (doi:10.1016/j.tics.2006.04.008)

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L. & Logothetis, N. K. 2005 Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J. Neurosci.* **25**, 5004–5012. (doi:10.1523/JNEUROSCI.0799-05.2005)

Green, K. P. & Kuhl, P. K. 1991 Integral processing of visual place and auditory voicing information during phonetic perception. *J. Exp. Psychol.: Hum. Percept. Perform* **17**, 278–288. (doi:10.1037/0096-1523.17.1.278)

Hamilton, R. H., Shenton, J. T. & Coslett, H. B. 2006 An acquired deficit of audiovisual speech processing. *Brain Lang.* **98**, 66–73. (doi:10.1016/j.bandl.2006.02.001)

Hennecke, M. E. & Stork, D. G. 1996 Visionary speech: looking ahead to practical speechreading systems. In *Speechreading by man and machine: models, systems and applications* (ed. D. G. Stork), NATO Advanced Study Institute Series, vol. 150, pp. 331–349. Berlin, Germany: Springer.

Hockley, N. S. & Polka, L. 1994 A developmental study of audiovisual speech perception using the McGurk paradigm. *Poster presented at the 12th Meeting of the Acoustical Society of America, Austin, TX*. 96, 3309.

Hubel, D. H. & Wiesel, T. N. 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154.

Iwano, K., Tamura, S. & Furui, S. 2001 Bimodal speech recognition using lip movement measured by optical-flow analysis. In *Proc. HSC2001, Kyoto, Japan*, pp. 187–190.

Jones, A. J. & Callan, E. D. 2003 Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Cognit. Neurosci. Neuropsychol.* **14**, 1129–1133.

Kohonen, T. 1988 Neural phonetic typewriter. *IEEE Comput.* **21**, 11–22.

Kohonen, T. 1995 *Self-organizing maps*. Heidelberg, Germany: Springer.

Kuhl, P. K. 2000 A new view of language acquisition. *Proc. Natl Acad. Sci. USA* **97**, 11 850–11 857. (doi:10.1073/pnas.97.22.11850)

Kuhl, P. K. 2007 Is speech learning 'gated' by the social brain? *Dev. Sci.* **10**, 110–120. (doi:10.1111/j.1467-7687.2007.00572.x)

Kuhl, P. K. & Meltzoff, A. N. 1982 The bimodal perception of speech in infancy. *Science* **218**, 1138–1141. (doi:10.1126/science.7146899)

Liberman, A. M. & Mattingly, I. G. 1985 The motor theory of speech perception revised. *Cognition* **21**, 1–36. (doi:10.1016/0010-0277(85)90021-6)

MacDonald, J. & McGurk, H. 1978 Visual influences on speech perception processes. *Percept. Psychophys.* **24**, 253–257.

Massaro, D. W. 1998 *Perceiving talking faces: from speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W. & Cohen, M. M. 1983 Evaluation and integration of visual and auditory information in speech perception. *J. Exp. Psychol.* **9**, 753–771.

McClelland, J. L. & Elman, J. L. 1986 The TRACE model of speech perception. *Cognit. Psychol.* **18**, 1–86. (doi:10.1016/0010-0285(86)90015-0)

McGurk, H. & MacDonald, J. 1976 Hearing lips and seeing voices. *Nature* **264**, 746–748. (doi:10.1038/264746a0)

Mills, A. E.(ed.) 1983 *Language acquisition in the blind child*. London, UK; Canberra, Australia: Croom Helm.

Mills, A. E. 1988 Visual handicap. In *Language development in exceptional circumstances* (eds D. Bishop & K. Mogford), pp. 150–163. Edinburgh, UK: Churchill Livingstone.

Munhall, K. G., Gribble, P., Sacco, L. & Ward, M. 1996 Temporal constraints on the McGurk effect. *Percept. Psychophys.* **58**, 351–362.

Norrix, W. L., Plante, E. & Vance, R. 2006 Auditory–visual speech integration by adults with and without language-learning disabilities. *J. Commun. Disord.* **39**, 22–36. (doi:10.1016/j.jcomdis.2005.05.003)

Omata, K. & Mogi, K. 2005*a* Audiovisual integration in Dichotic listening. In *Proc. of 9th European Conf. on Speech Communication and Technology (Interspeech), Lisboa, Portugal*, pp. 1717–1720.

Omata, K. & Mogi, K. 2005*b* Robust phoneme perception in complex stimulus conditions. In *Proc. 12th Int. Conf. on Neural Information Processing, Taipei, Taiwan*, pp. 521–526.

Patterson, M. L. & Werker, J. F. 2003 Two-month-old infants match phonetic information in lips and voice. *Dev. Sci.* **6**, 191–196. (doi:10.1111/1467-7687.00271)

Pekkola, J., Ojanen, V., Autti, T., Jaaskelaine, I. P., Mottonen, R., Tarkiainen, A. & Sams, M. 2005 Primary auditory cortex activation by visual speech: an fMRI study at 3T. *Neuroreport* **16**, 125–128. (doi:10.1097/00001756-200502080-00010)

Rabiner, L. & Juang, B. H. 1993 *Fundamentals of speech recognition*. Upper Saddle River, NJ: PTR Prentice-Hall, Inc.

Renart, A., Parga, N. & Rolls, T. E. 1999 Associative memory properties of multiple cortical modules. *Network: Comput. Neural Syst.* **10**, 237–255. (doi:10.1088/0954-898X/10/3/303)

Rizzolatti, G. & Craighero, L. 2004 The mirror-neuron system. *Annu. Rev. Neurosci.* **27**, 169–192. (doi:10.1146/annurev.neuro.27.070203.144230)

Robert-Ribes, J., Piquemal, M., Schwartz, J.-L. & Escudier, P. 1996 Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition. In *Speechreading by man and machine: models, systems and applications* (ed. D. G. Stork), NATO Advanced Study Institute Series, pp. 193–210. Berlin, Germany: Springer.

Rosenblum, L. D., Schmuckler, M. A. & Johnson, J. A. 1997 The McGurk effect in infants. *Percept. Psychophys.* **59**, 347–357.

Sams, M., Mottonen, R. & Sihvonen, T. 2005 Seeing and hearing others and oneself talk. *Cognit. Brain Res.* **23**, 429–435. (doi:10.1016/j.cogbrainres.2004.11.006)

Schwartz, J.-L., Robert-Ribes, J. & Escudier, P. 1998 Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. In *Hearing by eye II* (eds R. Campbell, B. Dodd & D. Burnham), ch. 4, pp. 85–108. East Sussex, UK: Psychology Press.

Sekiyama, K. 1997 Cultural and linguistic factors in audiovisual speech processing: the McGurk effect in Chinese subjects. *Percept. Psychophys.* **59**, 73–80.

Sekiyama, K. & Tohkura, Y. 1991 McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. Acoust. Soc. Am.* **90(4 Pt 1)**, 1797–1805. (doi:10.1121/1.401660)

Sekiyama, K. & Tohkura, Y. 1993 Inter-language differences in the influence of visual cues in speech perception. *J. Phonet.* **21**, 427–444.

Sekiyama, K., Kannoc, I., Miura, S. & Sugita, Y. 2003 Auditory–visual speech perception examined by fMRI and PET. *Neurosci. Res.* **47**, 277–287. (doi:10.1016/S0168-0102(03)00214-1)

Sumby, W. H. & Pollack, I. 1954 Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26**, 212–215. (doi:10.1121/1.1907309)

Van Wassenhove, V., Grant, K. W. & Poeppel, D. 2007 Temporal window of integration in audio-visual speech perception. *Neuropsychologia* **45**, 598–607. (doi:10.1016/j.neuropsychologia.2006.01.001)

Waibel, A. & Yegnanarayana, B. 1981 Comparative study of nonlinear time warping techniques in isolated word speech recognition systems. Technical report, Carnegie-Mellon University.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. & Lang, K. J. 1989 Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust. Speech Signal Process.* **37**, 328–339. (doi:10.1109/29.21701)

Warren, D. H. 1994 *Blindness and children: an individual differences approach*. New York, NY: Cambridge University Press.