The real and illusory in phoneme perception

Kei Omata[1)2)], Ken Mogi[2)1)]
1) Tokyo Institute of Technology Department of computational Intelligence and Systems science
2) Sony computer science laboratories, Inc.
omata@brn.dis.titech.ac.jp

Abstract

Phonemes are building blocks of linguistic conscious perception. The McGurk effect suggests that visual articulation can affect the phoneme perception. Recently, brain imaging studies have suggested that the McGurk effect is due to activities in the posterior part of the left superior temporal sulcus and gyrus, where neurons responding to both auditory and visual information are found.

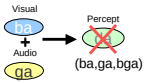
Here we formulate a model of the McGurk effect by means of the Self-organizing map (SOM). We examined whether putting the SOM in a series of learning process of phonemes would naturally induce a McGurk effect-like activities. We report the reconstruction of the McGurk effect in a manner consistent with the experimentally observed phenomenology of phoneme perception.

Our results suggest a mechanism underlying real and illusory perceptions in the context of learning.

Features of the McGurk effect

Fusion pair

Combination pair



(ba,ga,bga)

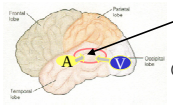
Auditory	Visual	Percept
Ba +	Ga =	Da
Ma +	Ka =	Na
Pa +	Ka =	Ta
bilabial	palatal	alveolar

(McGurk&MacDonald,1978)

⇨ The articulations of consonants are related with the McGurk effect.

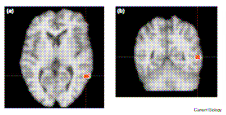
⇨ The fusion response occurs when the specific audio-visual pairs are presented.

The McGurk effect in Brain Science

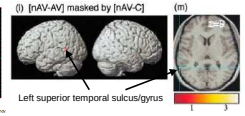


Multi-sensory region

(Scott. 2004, Sams, 2005, Mottonen,2006)

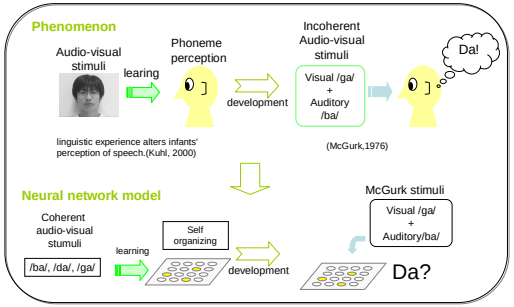


Audio-visual integration (Calvert et al, 2000)

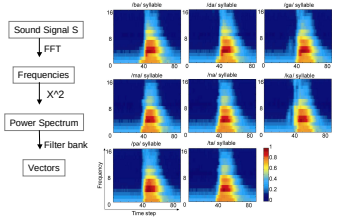


The McGurk effect (Sekiyama et al, 2003)

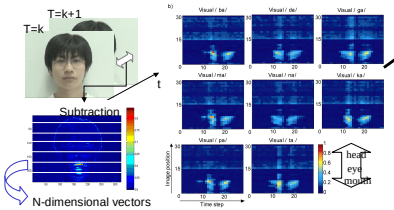
Background & Concept



The auditory vectors

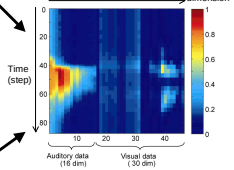


The visual vectors



Learning

Ex. An audio-visual vector

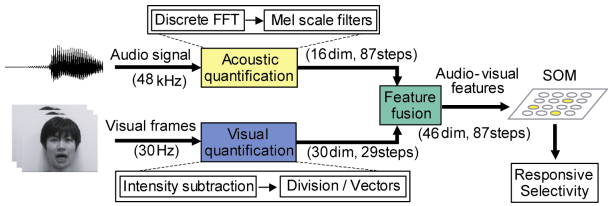


Time-synchronous audio and visual vectors are concatenated

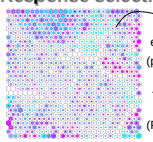
(Beauchamp et al., 2004, Ghazanfar et al., 2005,2006)

Method

1. **The construction of the phoneme perception system using the Self organizing map ('SOM', Kohonen, 1995)**
2. **Application of the McGurk stimuli to the maps**



Response selectivity P



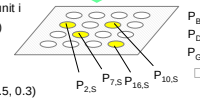
ex. responses for the unit i

$(P_{i,B}, P_{i,D}, P_{i,G}) = (2, 5, 3)$

$P_{i,S} = \frac{P_{i,S}}{\sum_S P_{i,S}}$ thus,

$(P_{i,B}, P_{i,D}, P_{i,G}) = (0.2, 0.5, 0.3)$

Ex. Application for a syllable S



$Percept = \max_S \sum_i P_{i,S}$

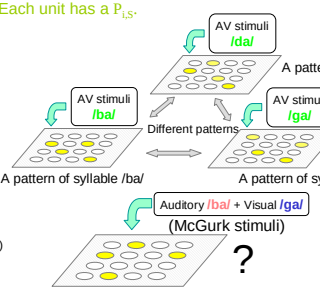
$P_B = P_{2,B}+P_{7,B}+P_{10,B}+P_{16,B} = 25$

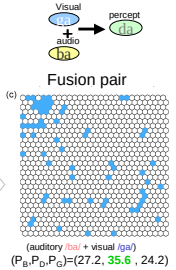
$P_D = P_{2,D}+P_{7,D}+P_{10,D}+P_{16,D} = 30$

$P_G = P_{2,G}+P_{7,G}+P_{10,G}+P_{16,G} = 20$

⇨ A Syllable /da/

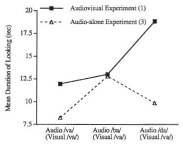
- Calculate response selectivities for all units.
- Each unit has a $P_{i,S}$.



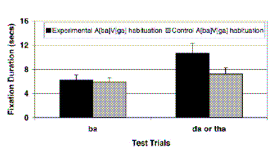


Genetic or Learning?

When does the McGurk effect appear?

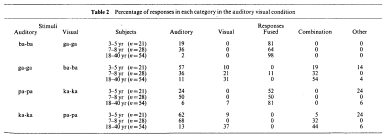


(Rosenblum, 1997)



(Burnham & Dodd, 2004)

Is language experiences important for the McGurk effect ?



(McGurk & MacDonald, 1976)

- Infants have an matching ability of auditory and visual speech information and also integrate the audio and visual information.
  (e.g. Kuhl & Meltzoff, 1982, Rosenblum et al. 1997, Burnham et al, 2004)
- Adults tend to report the McGurk effect more than infants.
  (e.g. Hockly et. al, 1994, Massaro et. al,1986, McGurk et al., 1976)

Hypothesis: Learning of the audio-visual speech perception may cause the McGurk effect.

Results

generalization capability (3 syllables ×30 utterances)

The performance of classifier (%)	/b,d,g/	/m,n,k/	/p,t,k/
5-fold Cross Validation	96	98	94

- 5-fold Cross-validation
- the ratios of discrimination in every group are more than 94%.

The McGurk effect (30 visual streams x 30 syllables)

McGurk effect(%)	/b,d,g/			/m,n,k/			/p,t,k/		
Presentation Pair	/b/	/d/	/g/	/m/	/n/	/k/	/p/	/t/	/k/
Fusion	20.9	46.8	32.2	23.4	34.4	42.1	9.6	49.1	41.1
Combination	52.2	2.3	45.4	47.0	0.6	52.4	83.0	1.2	15.8

- The fusion pairs were perceived as the third phone (alveolar).
- The combination pairs were perceived as either the presented phonemes.

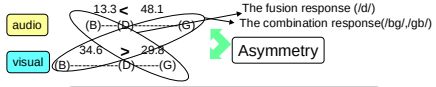
These results suggested that the McGurk effect occurred, and that the fusion pairs were similar to the alveolar consonants.

Analysis

similarity index: subtraction between the sample vectors of averaged syllables.
Distance : sum of the absolute values of subtraction for each intensity between frames.

$$f(S_i,S_j) = \sum_n^N \sum_m^M |x_{Si}(n,m) - x_{Sj}(n,m)| \quad f(S_i,S_j) = f_A(S_i,S_j) + f_V(S_i,S_j)$$

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



The McGurk effect may be based on the asymmetry of the similarities

Discussion/Conclusion

Stimuli

- There was an asymmetric relationship between the similarity of auditory features and that of visual features among the phonemes.
- The asymmetry relationships were found in all the three groups.

⇨ The existence of the asymmetric relationship between the audio and visual stimuli may induce the fusion responses of the McGurk effect when the particular audio-visual stimuli are presented.

System

- We hypothesized that time-synchronous audio-visual information is concatenated in the model.

⇨ The results suggested that the phoneme perception system requires a mechanism to combine auditory and visual information

Learning of audio-visual speech is an inducing factor of the McGurk effect.

Reference
McGurk, H. and MacDonald,J. 'Hearing lips and seeing voices',Nature,264:746-748 1976
Kohonen, T. 1995 *Self-Organizing Maps*, Springer-Verlag, Heidelberg