2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A Self-Organized Artificial Neural Network Architecture that Generates the McGurk Effect

Lennart Gustafsson
Computer Science, Electrical and Space Eng.
Luleå University of Technology, Sweden
Email: lgus@ltu.se

Tamas Jantvik
Decuria
11129 Stockholm, Sweden
Email: tamas@decuria.se

Andrew P. Papliński
Clayton School of Information Technology
Monash University, VIC 3800, Australia
Email: Andrew.Paplinski@monash.edu

***Abstract*—A neural network architecture, subjected to incongruent stimuli in the form of lip reading of spoken syllables and listening to different spoken syllables, is shown to generate the well-known McGurk effect, e.g. visual /ga/ and auditory /ba/ is perceived as /da/ by the network. The neural network is based on an architecture which has previously been successfully applied to sensory integration of congruent stimuli and is here extended to take into account that lip reading groups consonants into equivalence classes, bilabial, dento-labial and nonlabial consonants, rather than distinguishing between individual consonants.**

***Index Terms*—McGurk effect, Neural networks, lip reading**

I. Introduction

A. General properties of sensory integration

It is well established that experiencing an event or an object through more than one sensory modality brings several advantages: identification is more rapid [1], identification is more robust against corrupted stimuli [2] and threshold for detection is lower, for an overview see [3]. All this is true when stimuli are congruent, i.e., stimuli to the different sensory modalities emanate from the same object or event. In the opposite case, when the stimuli are incongruent, identification of each stimulus takes longer time [4] and becomes less certain.

B. The McGurk phenomenon

Exposing a test person to a filmed speaker saying a syllable while the sound of a different syllable has been dubbed onto the film is a well-studied case of incongruent stimuli. In a paper from 1976 [5] McGurk reported the discovery that the combination of auditory /ba-ba/ and visual /ga-ga/ in the large majority of adult test persons (98%) resulted in /da-da/ being perceived. The voiceless case of auditory /pa-pa/ and visual /ka-ka/ resulted in the percept /ta-ta/ in the majority of test persons, in this case, however, some (6%) perceived the auditory percept and others (7%) the visual percept.

In a second paper from 1978 [6] MacDonald enlarged the scope of the examinations and identified a third McGurk case: when the auditory syllable is /ma/ and the visual syllable is /da/, /ga/, /ka/, /na/, or /ta/, then /na/ is perceived by a great majority of the test persons. In the reversed cases, i.e. where the auditory syllable is /ga-ga/ and the visual syllable /ba-ba/ , the auditory syllable was perceived by a majority of the test persons, but a significant number perceived combinations such as /bda/, and a few perceived the visually presented syllable. The corresponding results were obtained for the two other reversed cases.

C. McGurk effect under different listening conditions

The results by [5] have since been replicated a number of times, see e.g. [7], [8], with similar results. [9] made an extensive experiment with the Japanese syllables /ba/, /pa/, /ma/, /wa/, /da/, /na/, /ra/, /ga/ and /ka/ (the phoneme /w/ is a bilabial consonant in Japanese). They found that "auditory wins" in general when listening conditions are good. By adding auditory noise many McGurk cases were induced though.

When using the terms congruity and incongruity one naturally implies that the stimuli coincide in time. There is, however, an appreciable tolerance for temporal asynchrony which varies between individuals [10], [11].

D. The McGurk effect as an experimental tool

The McGurk effect has also been used as a tool to study bimodal integration as such. One such topic is the development of sensory integration. Already [5] studied how the fused percepts /da/ and /ta/ differed between age groups and found that adult test persons had many more fused percepts than both pre-school and primary school children while the latter more frequently experienced the presented auditory percept. In a more recent study [12] it was found that children with learning disabilities were less likely to hear a fused percept than normal-learning children, children with learning disabilities more often reported hearing the visual stimulus.

An excellent review of research on the McGurk effect in the larger context of audiovisual integration can be found in the work of [13]. The McGurk effect has recently been employed in electrophysiological investigations of brain processes that demonstrate visual influence in auditory perception [14].

E. Explanations for the McGurk effect

A qualitative explanation for the McGurk effect is rather obvious: "Thus, in a ba-voice/ga-lips presentation, there is visual information for [ga] and [da] and auditory information with features common to [da] and [ba]. By responding to the common information in both modalities, a subject would arrive at the unifying percept [da]." This was stated already

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

by McGurk and MacDonald. They were, however mistaken when they claimed "...in the absence of auditory input, lip movements for [ga] are frequently misread as [da]" (see e.g the extensive confusion matrices for visually perceived consonants presented in the work of [15] where very few such misreadings are reported).

In their second paper [6] retracted their first explanation. Referring to the work of [16] they noted that there are visual similarities between syllables containing one of the nonlabial consonants /da/, /ga/, /ta/, and /ka/ and visual similarities between syllables containing one of the bilabial consonants /ba/, /pa/, and /ma/ such that syllables within each group could not be distinguished. Syllables were easily distinguishable between groups though. This is a special case of "equivalence classes" (see below). MacDonald and McGurk formulated a "manner-place" hypothesis to explain the McGurk perceptions of auditory bilabial and visual nonlabial stimuli. The auditory stimulus yields information on the "manner of articulation" ("voiced or voiceless", "oral or nasal", "stopped or continuant", etc.) whereas information about the "place of articulation" comes from the visual stimulus. The hypothesis argues that, "at an as yet unknown level of processing, information from the two sources is combined and synthesized, resulting in the 'auditory' perception of a best fit solution". The authors note that this hypothesis does not account for the perceptions of the combination of an auditory nonlabial and a visual bilabial stimulus.

[17] presented an argument based on hypothetical numerical values of visual and auditory similarity between the syllables /ba/, /ga/ and /da/ and implemented it in fuzzy logic. The argument results in the suggestion that /da/ is a good compromise between visual /ga/ and auditory /ba/. The verbal explanation employed feedback from bimodal processing to unimodal processing, also suggested by a number of other authors (see e.g. the work of [18]). Feedback from bimodal processing to unimodal processing will also be a part of our neural network architecture. A number of experiments to chart activities in cortex when a subject is exposed to McGurk inducing stimuli have been conducted. These experiments have resulted in an ever more complicated picture.

In a paper [19] a verbal explanation based on a neural architecture is proposed:

"These results suggest an architecture in which the STS contains small patches of neurons that respond to specific syllables. Activity across multiple syllable patches would be compared using a winner-take-all algorithm, with the most active patch determining perception. Each patch might receive input from neurons in visual and auditory association areas coding for specific visemes and phonemes. During presentation of congruent auditory–visual speech, input from auditory and visual neurons would be integrated, improving sensitivity. During presentation of incongruent McGurk stimuli, this process could result in unexpected percepts. For instance, if an STS patch representing 'da' received input from both auditory 'ba' and visual 'ga' neurons, the patch would have a large response during presentation of the 'ba' plus 'ga' McGurk stimulus, producing a 'da' percept" [19] (p.2417).

This suggestion agrees well with the simulations presented in this paper. However, feedback from the STS patch to the auditory patch is necessary and the concept of "equivalence class" must be included in the architecture to produce the desired results.

Two recent presentations [20], [21] also use neural networks to model the McGurk effect. They are very different from the treatment presented here, in several respects. Firstly, error correcting weight updating rules are employed in [20], [21], whereas our treatment employs self-organization. Secondly, their networks are single networks whereas ours is a connection of networks with two association levels (unimodal and bimodal respectively) and feedback from the bimodal level to the unimodal. In both these respect we believe our model is more biologically motivated. Thirdly, our stimuli have been video recorded and the features of sound and mouth movements directly extracted from these (see below). In [20], [21] the voice features include "the voice, the manner and the place of articulation". The visemes, i.e. the visual expressions, are in [20], [21] represented by "randomly generated vectors".

F. Equivalence classes

An experimental study of lip reading or speechreading carried out by [22] has established that many consonants cannot be distinguished from each other by vision alone (cited by [23]). [24] found that "only 4 visually-contrastive units are available consistently to the lipreader: bilabial, rounded-labial, labial-dental, and nonlabial." The concepts "viseme" and "equivalence class" have been introduced [25], [23], [26] to summarize the results of such studies in a lucid manner. Since the term "viseme" sometimes is interpreted as "visible consonant phonemes" [27], sometimes as "... group phonemes into categories called visemes" [27], we will use the unambiguous term "equivalence class" here. The consonants in an equivalence class can be identified with a high degree of certainty (from almost 100 % to approx. 70 %) as belonging to that class and are thus rarely mistaken as belonging to another class. A study has shown that under optimal conditions eight equivalence classes of English consonants can be identified [15].

Since the speaker and the test persons for this study are all native Swedish speakers we will refer to a few Swedish studies [28], [29]; a good summary can be found in the work of [30]. It was found that in a casual manner of speaking and ordinary lighting situation it is judicious to employ only three equivalence classes for consonants: bilabial consonants (b, p, m), dento-labial consonants (f, v), and non-labial consonants (n, s, sh, k, d, t, r, j, h, g, l and a few others, not used in our study). Under optimal testing conditions, with a hyper-articulating speaker, seven equivalence classes could be identified. Our video recording was arguably of the casual kind and we therefore employ the three equivalence classes in this paper. This point will be further argued under Materials.

II. Materials

One Swedish female speaker with particularly clear diction read the syllables /ba/, /da/, /fa/, /ga/, /ha/, /ja/, /ka/, /la/, /ma/, /na/, /pa/, /ra/, /sa/, /sha/, /ta/ and /va/. Ten Swedish test persons with normal hearing and eye sight were exposed to this series twice, in randomized order, under three conditions: auditory only, visual only and audiovisual. All involved had given their written, informed consent to participate in the experiment. One test person had misunderstood the test procedure and her result was not included in the total.

The auditory and audiovisual syllables were almost 100 % correctly identified. As expected, many visual syllables were incorrectly identified. The visual results are summarized in the Confusion matrix [23] shown in Table I.

TABLE I
CONFUSION MATRIX [23] SHOWING HOW SUBJECTS IDENTIFIED SYLLABLES PRESENTED ONLY VISUALLY

		VISUALLY PRESENTED STIMULUS															
		fa	va	ba	pa	ma	da	ga	ha	ja	ka	na	la	ra	sa	sja	ta
IDENTIFIED STIMULUS	fa	13	8	1	1		1										
	va	4	6	1			2	1	1				1	1			
	ba	1		4	4	4											1
	pa		1	9	7	7											
	ma			4	6	7											
	da							1	1		1	1	1	2			2
	ga								2	2	1	4		1	2	1	
	ha							4	3		5		1			1	2
	ja		1				1	2	1	4	2			4		1	2
	ka										2	2	1	2		1	1
	na						1		2	1		4				1	
	la						1	2	1		2	2	6				
	ra						1	2	2		1		3	1	2		1
	sa		1				5	1	1	4			2	3	6	4	1
	sja		1				3	1		6	1		1		6	6	5
	ta						1	1		1	1	2		1		2	2
	blank						2	3	4		2	3	2	3	1		1

We see that three equivalence classes can be identified in the matrix;

1) Out of 36 exposures to the dento-labial class (/fa/ and /va/) 31 were perceived as belonging to that class.
2) Out of 54 exposures to the bilabial class (/ba/, /pa/ and /ma/) 52 were perceived as belonging to that class.
3) Out of 198 exposures to the nonlabial class (/da/, /ga/, /ha/, /ja/, /ka/, /la/, /na/, /ra/, /sa/, /sja/ and /ta/) 167 were perceived as belonging to that class.

We also note that the sibilant fricatives /sa/ and /sja/ were perceived as belonging to that group in 22 out of 36 exposures. We refrain from denoting this as a separate group.

The identification of individual syllables is poor and mostly far below 50 % correct, the only exception being /fa/.

It will become clear from our simulations that the fact that we visually perceive consonants in equivalence classes rather than as individual consonants is essential in explaining the McGurk effect.

A. A summary of the image preprocessing of our stimuli

It has been shown that lip formation, teeth exposure and tongue movements are the most important features for identifying phonemes [25]. In our case we are satisfied with identifying equivalence classes of consonants. For this task a small set of features is found to be sufficient. The lips, shown in Figure 1, are approximated by an ellipse and the maximum

Fig. 1. Example frame showing our visual stimuli, with markers for outer lip boundary and teeth.

eccentricity and the minimum minor axis are extracted. The exposure of the upper teeth, when visible alone, is also a feature. The visual feature vector consists of four elements. When these vectors are fed to a self-organized feature map [31] with very few nodes available, one node will be assigned to the bilabial consonants, a second node to the dento-labial consonants and a third node to the non-labial consonants, as indeed would be expected. Figure 2 shows the result of the organization of a 4×4 self-organized feature map when trained with these visual features. The units on the borders are not active.

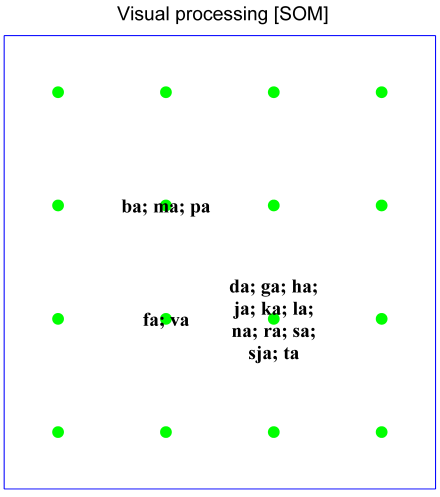
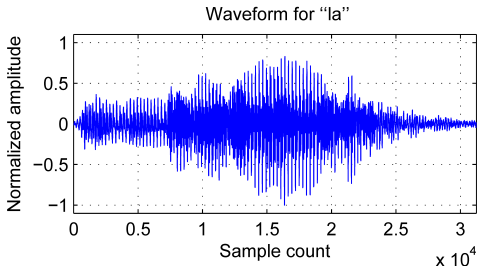


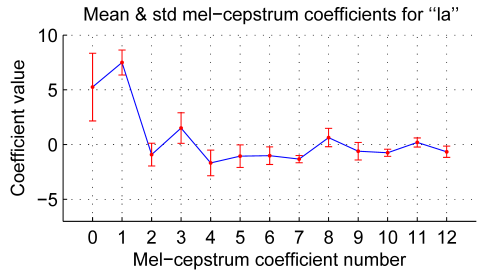
Fig. 2. Organization of a 4×4 self-organized feature map when trained with our visual features (the units on the borders are inactive). The units' labels show which unit respond the strongest for the listed stimuli.

B. A summary of the auditory preprocessing of our stimuli

A set of seventy-three melcepstral sequences, each thirteen elements long and concatenated to form a feature vector of nine hundred and forty-nine elements, was determined from the auditory time function for each syllable. For analysis reasons all vectors should be of equal length and therefore shorter syllables were zero padded to equal the longest syllable (sja). A discussion of the mel-cepstrum and its use in representation of speech is given by [32]. For computational reasons these vectors were subjected to principal component analysis (PCA) and the vector inputs to the auditory SOM are formed using the fifteen principal component scores that have the largest variance. The waveform of the syllable "la", and the element-wise mean and standard deviation of the resulting mel-cepstrum sequence is shown in Figure 3(a) and Figure 3(b), respectively.



(a) Waveform for "la"



(b) Mean mel-cepstrum coefficients

Fig. 3. (top) Waveform of an example auditory stimulus; the syllable "la". (bottom) Element-wise mean and standard deviation of the mel-cepstrum coefficients extracted from the auditory stimulus "la" shown at the top.

III. A SELF-ORGANIZED ARTIFICIAL NEURAL NETWORK ARCHITECTURE ADAPTED FOR STUDYING THE MCGURK EFFECT

The architecture depicted in Figure 4 originate from our earlier work on sensory integration in two or more modalities [33], [34], [35], [36] and all non-trivial technical details can be found in [36].

There are three modules on two levels in Figure 4. The unimodal level contains two modules, the left one for visual processing and the right one for auditory processing. The acronym SOM stands for (Kohonen) Self-Organizing Map and the acronym SumSOM stands for Summing SOM. The output from the SOM is determined by the Winner Take All (WTA) operation. Both the position and the activity level of the winning node are determined and conveyed to the bimodal level. The activity level is a measure of the similarity between the input and the ideal stimuli the map has self-organized to detect in the visual and auditory maps respectively. The SumSOM for auditory processing sums the activities caused by the auditory input and the bimodal output. The bimodal SumSOM has been self-organized under the influence of congruent visual and auditory stimuli and therefore when the stimuli are congruent the bimodal processing yields an output that reinforces the auditory input through the feedback. In the case of incongruent visual and auditory stimuli the feedback may cause the auditory processing to choose another winner than the actual auditory stimulus.

Simulations with this architecture, using congruent letters and phonemes, have demonstrated that time to identification is shorter and that identification in noise is more robust than made possible by unimodal stimuli. We will now test the usefulness of this architecture when stimuli are incongruent.

The information conveyed from the unimodal processing level to the bimodal level is restricted to the positions and activity levels of the winners (the winner has the highest activity level in the map, where activity level is a measure of the similarity between the input and the ideal stimuli that the map has self-organized to detect) in the visual and auditory maps respectively.

In the McGurk case of visual /ga/ and auditory /ba/ it is possible that /da/ becomes the winner in the bimodal map as there will be contributions to the activity level of /da/ from both visual /ga/ and auditory /ba/ while visual /ga/ contributes very little to auditory /ba/ in the bimodal map, and vice versa. We have, however, found that simulations do not yield the consistent McGurk cases as experimentally found [6], [9]. We have also seen that test persons do not identify the visual consonants correctly, but only in equivalence classes and we therefore adapt the original architecture in Figure 4 to include a stage where the visual syllables are grouped in three equivalence classes.

We have also added feedback from the bimodal processing to the visual processing. This is not necessary to generate the McGurk effect but it enables the architecture to allow "that the interaction between hearing and lipreading is genuinely bidirectional" as has recently been reported [37].

Simulations with this slightly extended architecture have often but not sufficiently often exhibit the experimentally found fused syllable in the bimodal processing which then through feedback to auditory processing causes the established McGurk effect.

We, however, notice that auditory /ba/ causes the second highest activity in the /da/-patch in the auditory processing SOM. Likewise the /ta/-patch has the second highest activity caused by auditory /pa/ and the /na/-patch has the second highest activity level caused by auditory /ma/, see Figure 5(a-c). This is obviously a contributing explanatory factor of the McGurk effect.

Our architecture has been designed to exhibit the experimen-

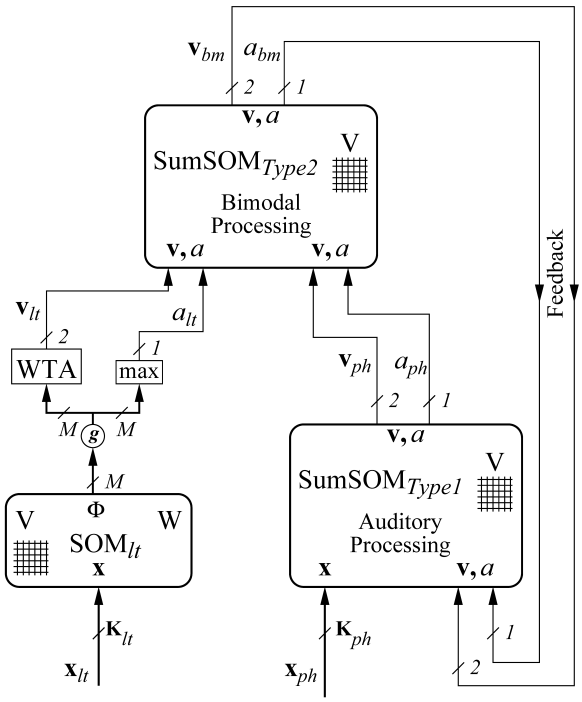
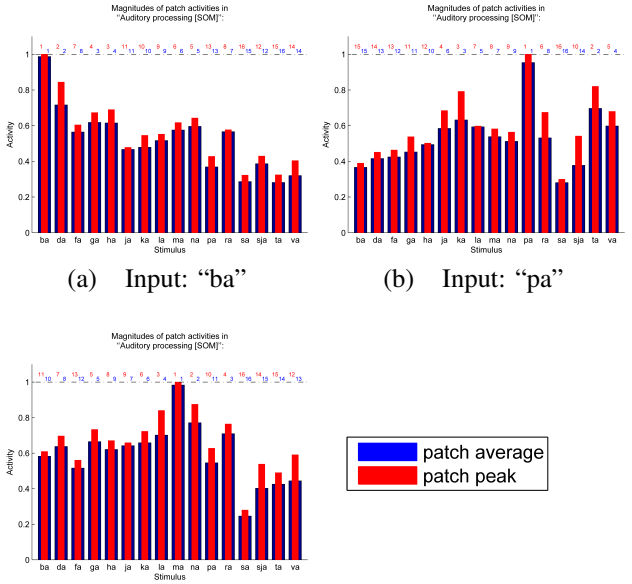


Fig. 4. A two-level MMSON [36] with feedback processing auditory and visual stimuli (adopted from the work of [36]).



(a) Input: "ba" (b) Input: "pa"

(c) Input: "ma"

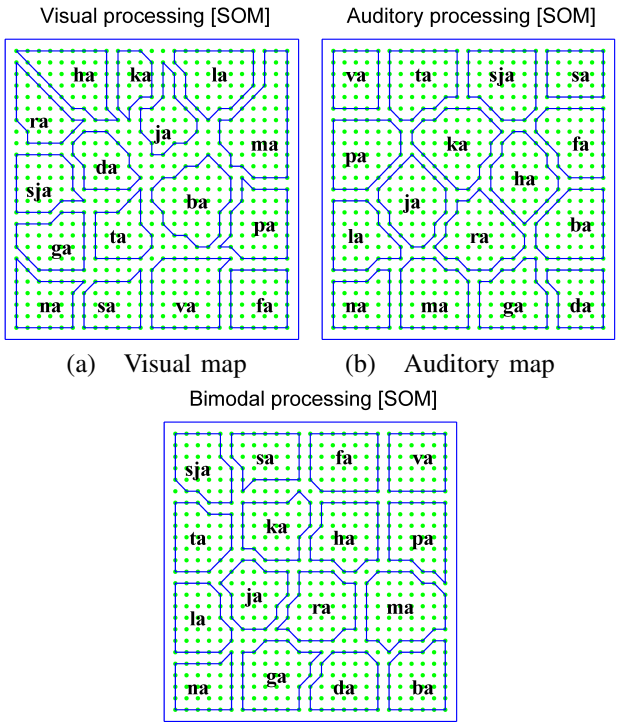
Fig. 5. The bar plots (a), (b) and (c) show the peak (red) and average (blue) activities in the self-organized feature map organized on auditory stimuli when presented with the feature vectors representing "ba", "pa", and "ma", respectively.

tally established characteristics of sensory integration while being as computationally efficient as possible. To correctly exhibit the McGurk effect the requisite demand for computational efficiency must be eased slightly. By forwarding also the position and activity of the patch with the second highest activity from auditory to bimodal processing the McGurk effect is consistently generated.

The dynamics of the perceptual process is shown in Figure 7(a-f) where initial (Figure 7(a-c)) and final winners (Figure 7(d-f)) are shown. As expected the /da/-patch wins in all the final maps.

IV. Simulation Results

The visual and the auditory self-organizations resulting from learning congruent stimuli are shown in Figures 6(a) and 6(b), and that at the bimodal level in Figure 6(c). The similarity



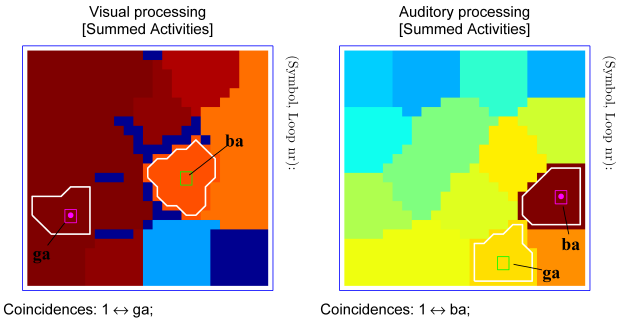
(a) Visual map (b) Auditory map

(c) Bimodal map

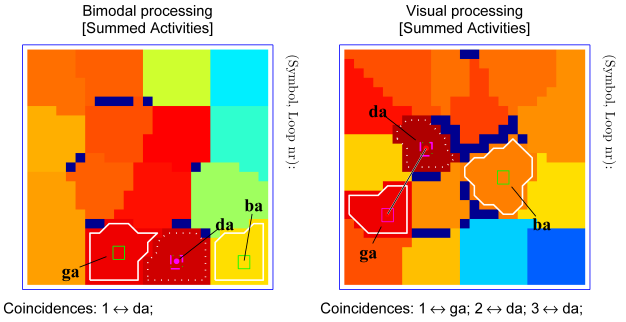
Fig. 6. Areas of classification for labelled letters, syllables and letter/syllable combinations after self-organization. The labels share their positions with the ideal neurons. In all three modules the response field consist of the output signals of 25×25 neurons.

properties at the unimodal levels are evident; bilabial, dento-labial and non-labial consonants are grouped together in three groups and syllables that sound alike are placed close to each other. The three visual groups are reduced to three equivalence classes before data are conveyed to the bimodal level.

The classic McGurk case – visual /ga/ and auditory /ba/ – causes the initial activities at the unimodal and bimodal levels as shown in Figures 7(a–c). Naturally /ba/ has the highest initial activity in the auditory map, but since the bimodal map shows the highest activity at /da/, this is fed back to the auditory processing map and the final result is that /da/ reaches the highest activity also in the auditory processing map; this is shown in Figures 7(d–f). This result was consistently obtained

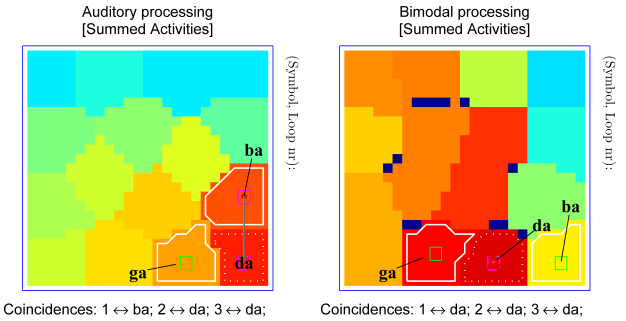


(a) Visual module's and (b) Auditory module's initial responses



(c) Bimodal module's initial responses

(d) Visual module's stabilized responses



(e) Auditory module's and (f) Bimodal module's stabilized responses

Fig. 7. The effect of presenting "ga" to the visual processing module and "ba" to the auditory processing module. (a), (b), and (c) show the activity levels in the visual, auditory and bimodal processing modules after the first feed-forward sweep, respectively. (d), (e), and (f) show the activity levels in the visual, auditory and bimodal processing modules when the dynamics have stabilized. The activity levels are temperature coded (dark blue represents the lowest activity while dark red represents the highest activity), and the winner neurons are indicated by magenta filled circles in a magenta square. Patches are laid out as in Figure 6.

throughout the simulations (a total of 20). The two other McGurk cases, i.e. auditory /pa/ and visual /ka/ perceived as /ta/ and auditory /ma/ and visual /ga/ perceived as /na/, were also correspondingly and consistently modeled.

We also report that the "reversed" McGurk cases, i.e. the cases where the auditory syllable was /ga/, /ka/ or /ga/ respectively with the visual syllables /ba/, /pa/ or /ma/ respectively, resulted in the visual stimulus being heard. This is mainly in agreement with the results reported by [9], and by [5] (for adult test persons) but not with the results of [6]. For these cases there is a spread in the experimental results which we don't see in our simulation results. There is, however, one simulation charastistic of these cases which stands out. The activity in both unimodal and bimodal processing units are considerably lower than they are for the McGurk cases (which in turn show lower activity levels than the congruent cases). This means that the inputs are not as clearly identified by the artificial neural network. It is tempting to believe that the same phenomenon reveals itself to test persons, resulting in more fragmented results than obtained in the McGurk cases or the congruent cases.

V. Discussion

In our neural network architecture we have implemented knowledge from psychological experiments as the resolution of visual syllables was reduced to three equivalence classes. In future work the architecture may be developed so that it self-organizes into the suitable resolution. This, however, demands tests with a number of casually pronouncing speakers. With only one speaker the visual SOM will self-organize to yield full resolution, i.e. all visual syllables will be individually identifiable. With a number of speakers we would expect overlaps between syllables from different speakers and individual syllable recognition would be lost.

We obtained a 100% McGurk effect in our simulations. This is somewhat higher than the percentages obtained in psychological testing. Our results were obtained from different self-organizations and simulations with the same network architecture. While it is judicious to use three equivalence classes for consonants it is possible that some test persons are better than the majority at visually recognizing consonants. Tests with different numbers of equivalence classes might yield slightly different results from ours; we leave this for future research.

VI. Conclusion

A neural network architecture, developed for the study of sensory integration of congruent stimuli has been applied to incongruent stimuli. With minor modifications of the architecture it has been shown to generate the well-known McGurk effect.

Acknowledgements

The authors acknowledge the financial support of the Ph.D. Polis collaboration program between Luleå University of Technology and Monash University in Melbourne. The authors also acknowledge both the feedback and the encouraging support from Jerker Delsing at Luleå University of Technology.

References

[1] M. Hershenson, "Reaction time as a measure of intersensory facilitation," *Journal of Experimental Psychology*, vol. 63, no. 3, pp. 289 – 293, March 1962. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022101507645753

[2] D. E. Callan, A. M. Callan, C. Kroos, and E. Vatikiotis-Bateson, "Multimodal contribution to speech perception revealed by independent component analysis: a single-sweep EEG case study," *Cognitive Brain Research*, vol. 10, no. 3, pp. 349 – 353, January 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926641000000549

[3] N. Bolognini, F. Frassinetti, A. Serino, and E. Làdavas, ""Acoustical vision" of below threshold stimuli: interaction among spatially converging audiovisual inputs," *Experimental Brain Research*, vol. 160, no. 3, pp. 273–282, 2005. [Online]. Available: http://dx.doi.org/10.1007/s00221-004-2005-z

[4] A. Aleksandrov, E. Dmitrieva, and L. Stankevich, "Experimental formation of visual-auditory associations on phoneme recognition," *Neuroscience and Behavioral Physiology*, vol. 40, no. 9, pp. 998–1002, 2010. [Online]. Available: http://dx.doi.org/10.1007/s11055-010-9359-4

[5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, Dec. 1976. [Online]. Available: http://dx.doi.org/10.1038/264746a0

[6] J. MacDonald and H. McGurk, "Visual influences on speech perception processes," *Attention, Perception, & Psychophysics*, vol. 24, no. 3, pp. 253–257, 1978. [Online]. Available: http://dx.doi.org/10.3758/BF03206096

[7] M. Sams, R. Aulanko, M. Hämäläinen, R. Hari, O. V. Lounasmaa, S.-T. Lu, and J. Simola, "Seeing speech: visual information from lip movements modifies activity in the human auditory cortex," *Neuroscience Letters*, vol. 127, no. 1, pp. 141 – 145, June 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/030439409190914F

[8] T. Chen and D. Massaro, "Mandarin speech perception by ear and eye follows a universal principle," *Attention, Perception, & Psychophysics*, vol. 66, no. 5, pp. 820–836, 2004. [Online]. Available: http://dx.doi.org/10.3758/BF03194976

[9] K. Sekiyama and Y. Tohkura, "McGurk effect in non-english listeners: Few visual effects for japanese subjects hearing japanese syllables of high auditory intelligibility," *Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 1797–1805, 1991. [Online]. Available: http://dx.doi.org/10.1121/1.401660

[10] K. Munhall, P. Gribble, L. Sacco, and M. Ward, "Temporal constraints on the mcGurk effect," *Attention, Perception, & Psychophysics*, vol. 58, no. 3, pp. 351–362, 1996. [Online]. Available: http://dx.doi.org/10.3758/BF03206811

[11] L. M. Miller and M. D'Esposito, "Perceptual fusion and stimulus coincidence in the cross-modal integration of speech," *The Journal of Neuroscience*, vol. 25, no. 25, pp. 5884–5893, 2005. [Online]. Available: http://www.jneurosci.org/content/25/25/5884.abstract

[12] E. A. Hayes, K. Tiippana, T. G. Nicol, M. Sams, and N. Kraus, "Integration of heard and seen speech: a factor in learning disabilities in children," *Neuroscience Letters*, vol. 351, no. 1, pp. 46–50, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304394003009716

[13] R. Campbell, "The processing of audio-visual speech: empirical and neural bases," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 1001–1010, 2008. [Online]. Available: http://rstb.royalsocietypublishing.org/content/363/1493/1001.abstract

[14] E. Smith, S. Duede, S. Hanrahan, T. Davis, P. House, and B. Greger, "Seeing is believing: Neural representations of visual stimuli in human auditory cortex correlate with illusory auditory perceptions," *PLOS ONE*, vol. 8, no. 9, September 2013.

[15] J. J. Williams, J. C. Rutledge, A. K. Katsaggelos, and D. C. Garstecki, "Frame rate and viseme analysis for multimedia applications to assist speechreading," *The Journal of VLSI Signal Processing*, vol. 20, no. 1–2, pp. 7–23, 1998. [Online]. Available: http://dx.doi.org/10.1023/A:1008062122135

[16] C. A. Binnie, A. A. Montgomery, and P. L. Jackson, "Auditory and visual contributions to the perception of consonants," *J Speech Hear Res*, vol. 17, no. 4, pp. 619–630, 1974. [Online]. Available: http://jslhr.asha.org/cgi/content/abstract/17/4/619

[17] R. Bovo, A. Ciorba, S. Prosser, and A. Martini, "The mcGurk phenomenon in Italian listeners," *Acta Otorhinolaryngologica Italica*, vol. 29, no. 4, pp. 203–208, August 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/20161878/

[18] M. M. Benoit, T. Raij, F.-H. Lin, I. P. Jääskeläinen, and S. Stufflebeam, "Primary and multisensory cortical activity is correlated with audiovisual percepts," *Human Brain Mapping*, vol. 31, no. 4, pp. 526–538, September 2010. [Online]. Available: http://dx.doi.org/10.1002/hbm.20884

[19] M. S. Beauchamp, A. R. Nath, and S. Pasalar, "fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the mcGurk effect," *The Journal of Neuroscience*, vol. 30, no. 7, pp. 2414–2417, 2010. [Online]. Available: http://www.jneurosci.org/content/30/7/2414.abstract

[20] I. Sporea and A. Gruning, "Modelling the McGurk effect," in *Proc. 18th ESANN*, 2010, pp. 1–6.

[21] ——, "A distributed model of memory for the McGurk effect," in *Proc. IJCNN*. IEEE, 2010, pp. 1–4.

[22] F. DeLand, *The story of lip-reading: its genesis and development*. The Volta Bureau, 1931.

[23] C. G. Fisher, "Confusions among visually perceived consonants," *J Speech Hear Res*, vol. 11, no. 4, pp. 796–804, 1968. [Online]. Available: http://jslhr.asha.org/cgi/content/abstract/11/4/796

[24] M. F. Woodward and C. G. Barber, "Phoneme perception in lipreading," *Journal of Speech, Language, and Hearing Research*, vol. 3, no. 3, pp. 212–222, September 1960. [Online]. Available: http://jslhr.highwire.org/cgi/content/citation/3/3/212

[25] K. W. Berger, *Speechreading: principles and methods*, 1st ed. Baltimore, MD: National Educational Press, 1972, iSBN: 0879710055.

[26] E. T. Auer, Jr and L. E. Bernstein, "Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness," *The Journal of the Acoustical Society of America*, vol. 102, no. 6, pp. 3704–3710, 1997. [Online]. Available: http://link.aip.org/link/?JAS/102/3704/1

[27] C. S. Campbell and D. W. Massaro, "Perception of visible speech: influence of spatial quantization," *Perception*, vol. 26, no. 5, pp. 627–644, April 1997. [Online]. Available: http://www.perceptionweb.com/abstract.cgi?id=p260627

[28] S. Amcoff, "Visuell perception av talljud och avläsestöd för hörselskadade (in Swedish)," LSH Uppsala; Pedagogiska Institutionen, Tech. Rep., 1970.

[29] J. Mártony, "On speechreading of Swedish consonants and vowels," KTH (STL-QPSR), Tech. Rep., March 1974.

[30] T. Öhman, "An audio-visual speech database and automatic measurement of visual speech," KTH (TMH-QPSR), Tech. Rep., January 1998.

[31] T. Kohonen, *Self-Organizing Maps*, 3rd ed., ser. Springer Series in Information Sciences. New York: Springer, 2001, vol. 30, iSBN 3-540-67921-9, ISSN 0720-678X.

[32] B. Gold and N. Morgan, *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, Inc., 2000.

[33] A. P. Papliński and L. Gustafsson, "Feedback in multimodal self-organizing networks enhances perception of corrupted stimuli," in *Lect. Notes in Artif. Intell.*, vol. 4304. Springer, 2006, pp. 19–28.

[34] L. Gustafsson and A. P. Papliński, "Bimodal integration of phonemes and letters: an application of multimodal self-organizing networks," in *Proc. Int. Joint Conf. Neural Networks*, Vancouver, Canada, July 2006, pp. 704–710.

[35] L. Gustafsson, T. Jantvik, and A. P. Papliński, "A multimodal self-organizing network for sensory integration of letters and phonemes," in *Proc. IASTED Int. Conf. Artif. Intell. Soft Comp.*, Palma De Mallorca, Spain, Aug. 2007, pp. 25–31.

[36] T. Jantvik, L. Gustafsson, and A. P. Papliński, "A self-organized artificial neural network architecture for sensory integration with applications to letter–phoneme integration," *Neural Computation*, vol. 23, no. 8, pp. 2101–2139, August 2011.

[37] M. Baart and J. Vroomen, "Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading," *Neuroscience letters*, vol. 471, no. 2, pp. 100–103, 2010.