Feedback in Multimodal Self-organizing Networks Enhances Perception of Corrupted Stimuli

Andrew P. Papliński[1] and Lennart Gustafsson[2]

[1] Clayton School of Information Technology,
Monash University, Victoria 3800, Australia
app@csse.monash.edu.au

[2] Computer Science and Electrical Engineering,
Luleå University of Technology, S-971 87 Luleå, Sweden
Lennart.Gustafsson@ltu.se

Abstract. It is known from psychology and neuroscience that multimodal integration of sensory information enhances the perception of stimuli that are corrupted in one or more modalities. A prominent example of this is that auditory perception of speech is enhanced when speech is bimodal, i.e. when it also has a visual modality. The function of the cortical network processing speech in auditory and visual cortices and in multimodal association areas, is modeled with a Multimodal Self-Organizing Network (MuSON), consisting of several Kohonen Self-Organizing Maps (SOM) with both feedforward and feedback connections. Simulations with heavily corrupted phonemes and uncorrupted letters as inputs to the MuSON demonstrate a strongly enhanced auditory perception. This is explained by feedback from the bimodal area into the auditory stream, as in cortical processing.

1 Introduction

Bimodal integration of sensory information is advantageous when phenomena have qualities in two modalities. Audiovisual speech, i.e. speech both heard and seen by lip reading is a case where such sensory integration occurs. An important advantage that this integration yields is that audiovisual speech is more robust against noise, see e.g. [1]. This advantage, robustness of identification against noise in stimuli, that are sensed by more than one sensory modality is a general property of bimodal integration. Bimodal and multimodal integration has been studied extensively, for reviews, see [2].

There are areas in cortex that have long been recognized as multimodal association areas, such as the superior temporal polysensory area (STP), see e.g. [3,4], but more recently it has been established that multimodal convergence also occurs earlier in cortical sensory processing, in unimodal (which thus are not exclusively unimodal) sensory cortices [5,6].

Convergence of signals conveying auditory and visual information onto a neuron or a neural structure can be mediated in different ways. Feedforward or

A. Sattar and B.H. Kang (Eds.): AI 2006, LNAI 4304, pp. 19–28, 2006.

bottom up connections from lower levels to higher levels in the neural hierarchy and feedback or top down connections going in the opposite direction both serve to integrate information from different sensory modalities, see e.g. [7,8].

The functionality of feedforward and feedback connections has been extensively studied in vision. When a visual stimulus is presented there follows a rapid forward sweep of activity in visual cortex, with a delay of only about 10 msec for each hierarchical level [9]. The initial activity is thus determined mainly by feedforward connections. Feedback will then dynamically change the tuning of neurons even in the lowest levels.

Feedback in bimodal sensory processing has been found to be important in the bimodal processing of audiovisual speech. Speech is processed in several cortical regions, see [10] for a review. Sensory specific cortices provide one of the stages in the human language system [11,5,12]. Auditory processing for phoneme perception takes place in the left posterior Superior Temporal Sulcus (STSp) see e.g. [13,14]. Bimodal integration of audiovisual speech takes place in the multimodal association area in the Superior Temporal Sulcus (STS) and the Superior Temporal Gyrus (STG) [12], located between the sensory-specific auditory and visual areas.

In audiovisual speech there is a perceptual gain in the sensory-specific auditory cortex as compared to purely auditory speech. This corresponds to increased activity in unimodal cortices when bimodal stimuli are presented and is believed to be the result of top-down processing from the STS, modulating the activity in unimodal cortices [15]. It has been found, see [4], that while sensory convergence in higher-order bi- and multisensory regions of the superior temporal sulcus (STS) is mediated by feedforward connections, the visual input into auditory cortex is mediated by feedback connections. This enhances the perception in auditory cortex. As a comparatively recent development in human evolution language also exists in written form, i.e. language has yet another set of visual properties. Simultaneous presentation of written text and auditory speech is not as “natural” an occurrence as lip reading and auditory speech, yet, if the written text and the auditory speech are congruent, speech perception is improved, see [16,17]. Neural resources for processing of letters exist in or close to the left fusiform gyrus, see [18,19,20]. Bimodal integration of phonemes and letters takes place in the STS [21,22].

In this paper, which is a direct continuation of work presented in [23,24], we model the processing of phonemes and letters both in sensory-specific areas and in sensory integration. The network we use for this study is a multimodal self-organizing network (MuSON) which consists of unimodal maps with phonetic and graphic inputs respectively, and an integrating bimodal map. These maps correspond to the cortical architecture for phonetic processing in the STSp, processing of letters in the fusiform area and the bimodal integration in the STS. In [22] it is argued that there is feedback from the bimodal integrating area down into the sensory-specific auditory area. This feedback is also part of our model, in that we introduce a second auditory map which accepts feedforward inputs from the first auditory map as well as feedback inputs from the bimodal map.

We have earlier shown [24] in a study without feedback how templates for phonemes and letters result from self-organization of the unimodal maps and integrate into templates for the bimodal percepts in the bimodal map. We have also shown that the bimodal percepts are robust against additive noise in the letters and phonemes. The purpose of this study is to show how this robustness of the bimodal percepts can be "transferred" down in the processing stream by feedback. In essence we want to show that we hear a noisy phoneme better when we see the corresponding uncorrupted letter.

2 The Multimodal Self-Organizing Networks

Self-organizing neural networks have been inspired by the possibility of achieving information processing in ways that resemble those of biological neural systems.

In particular, pattern associators based on Hebbian learning [25] and self-organizing maps [26] show similarities with biological neural systems. Pattern associators have been employed to simulate the multimodal sensory processing in cortex [27]. Kohonen Self-Organizing Maps (SOMs) are well-recognized and intensively researched tools for mapping multidimensional stimuli onto a low dimensionality (typically 2) neuronal lattice, for an introduction and a review,

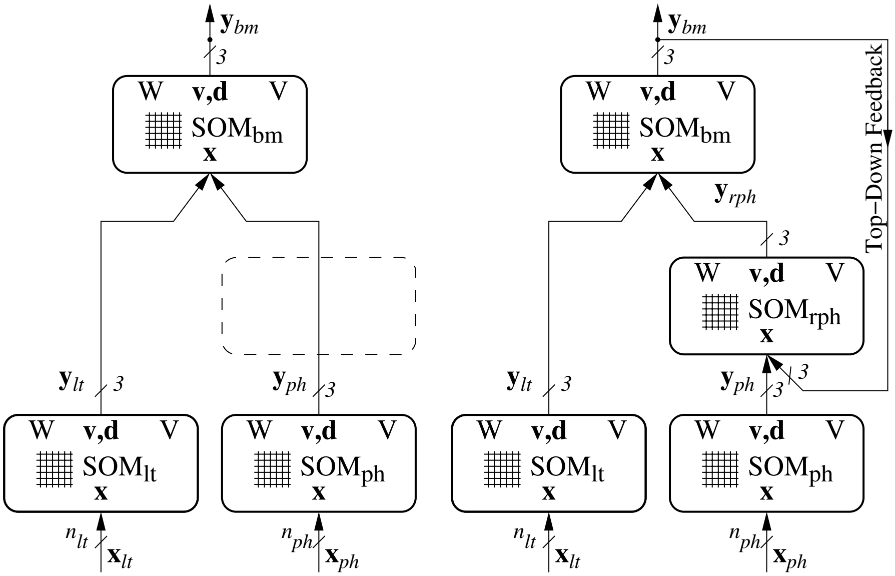

Fig. 1. Left: A two-level feedfoward-only Multimodal Self-Organizing Network (MuSON) processing auditory and visual stimuli. The auditory stimuli are processed in SOM_{ph}, and the visual stimuli in SOM_{lt}. Bimodal integration then takes place in SOM_{bm}. Right: A three-level Multimodal Self-Organizing Network (MuSON) with a feedback connection from the bimodal level to the auditory stream.

see [26]. In this paper we will employ a network of interconnected SOMs, referred to as Multimodal Self-Organizing Networks (MuSONs), see [23,24].

We consider first a feedforward Multimodal Self-Organizing Network (MuSON) as presented in the left part of Figure 1. The pre-processed sensory stimuli, $\mathbf{x}_{lt}$ and $\mathbf{x}_{ph}$ form the inputs to their respective unisensory maps, SOM$_{lt}$ and SOM$_{ph}$. Three-dimensional outputs from these maps, $\mathbf{y}_{lt}$ and $\mathbf{y}_{ph}$, are combined together to form a six-dimensional stimulus for the higher-level bimodal map, SOM$_{bm}$. The learning process takes place concurrently for each SOM, according to the well-known Kohonen learning law, see [23,24] for details. After self-organizations each map performs the mapping of the form: $\mathbf{y}(k) = g(\mathbf{x}(k); W, V)$, where $\mathbf{x}(k)$ represents the k^{th} stimulus for a given map, W is the weight map, and V describes the structure of the neuronal grid. The 3-D output signal $\mathbf{y}(k)$ combines the 2-D position of the winner with its 1-D activity.

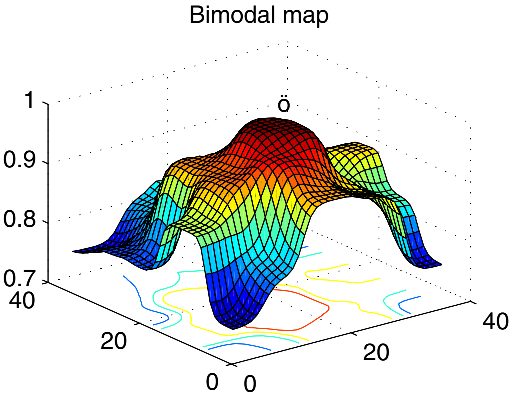

Fig. 2. The activity in the trained bimodal map when the letter and the phoneme **ö** is the input to the sensory-specific maps. The patch of neurons representing **ö** is clearly distinguished from other patches.

The position of the winner can be determined from the network (map) post-synaptic activity, $d(k) = W \cdot \mathbf{x}(k)$. As an example, in Figure 2, we show the post-synaptic activity of a trained bimodal map when the visual letter stimulus $\mathbf{x}_{lt}(k)$ and the auditory phoneme stimulus $\mathbf{x}_{ph}(k)$ representing letter/phoneme **ö** is presented. The activity in the map for one phoneme/letter combination shows one winning patch with activity descending away from this patch as illustrated in Figure 2. Such winning patches form maps as in Figure 3.

3 The Unimodal Visual Map for Letters and Auditory Map for Phonemes

The stimulus $\mathbf{x}_{lt}$ to the visual letter map is a 22-element vector obtained by a principal component analysis of scanned (21×25)-element binary character

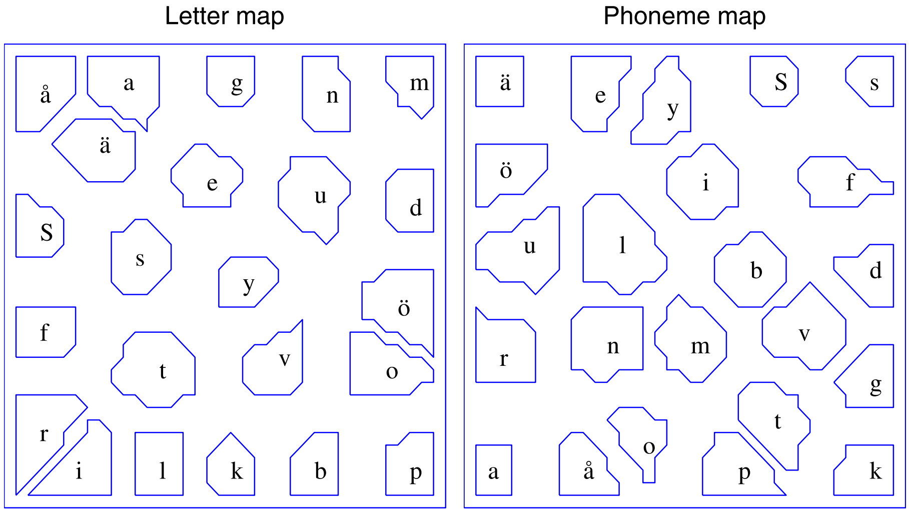

Fig. 3. Patches of highest activity for labeled letters and phonemes after self-organization on a map of 36×36 neurons

images. After self-organization is complete, the properties of the visual letter map SOM_{lt} may be best summarized as in the left part of Figure 3. The map shows the expected similarity properties — symbols that look alike are placed close to each other in the map. Note, for example, the cluster of characters **f**, **t**, **r**, **i**, **l**, **k**, **b** and **p** based predominantly on a vertical stroke. The patches in Figure 3 cover populations of neurons which show the highest activity for their respective stimuli. The neuronal populations within the patches of the letter map constitute the detectors of the respective letters.

In 1988 Kohonen presented the "phonetic typewriter", a phonotopic SOM that learned to identify Finnish phonemes [28,26]. In our study the auditory material consists of twenty-three phonemes as spoken by ten native Swedish speakers. Thirty-six melcepstral coefficients [29] represent each phoneme spoken by each speaker. These feature vectors were averaged over the speakers, yielding one thirty-six element feature vector for each phoneme. This averaged set of vectors constitutes the inputs $\mathbf{x}_{ph}$ to the auditory map SOM_{ph}. After the learning process we obtain a phoneme map as presented in the right part of Figure 3. As for the letter map the patches of neuronal populations constitute the detectors of the respective phonemes.

Note that the plosives **g**, **k**, **t** and **p**, the fricatives **s**, **S** (this is our symbol for the sh-sound as in English she) and **f** and the nasal consonants **m** and **n** form three close groups on the map. Vowels with similar spectral properties are placed close to each other. The back vowels **a**, **å** and **o** are in one group, the front vowels **u ö**, **ä**, **e**, **y** and **i** in another group with the tremulant **r** in-between. The exact placing of the groups vary from one self-organization to another, but the existence of these groups is certain.

4 The Bimodal Map Integrating Phonemes and Letters

The outputs from the auditory phoneme map and the visual letter map are combined as 6-dimensional inputs $[\mathbf{y}_{lt}\ \mathbf{y}_{ph}]$ to the bimodal map SOM$_{bm}$. Self-organization results in the map shown in Figure 4. The similarity characteristics of this map are derived from the placement of the patches in the unimodal maps and thus only indirectly reflect the features of the phonemes and letters. The fricative consonants **s**, **S** and **f** form a group in the combined map as do the nasal consonants **m** and **n**. Most, but not all, vowels form a group and those who are isolated have obviously been placed under influence from the visual letter map.

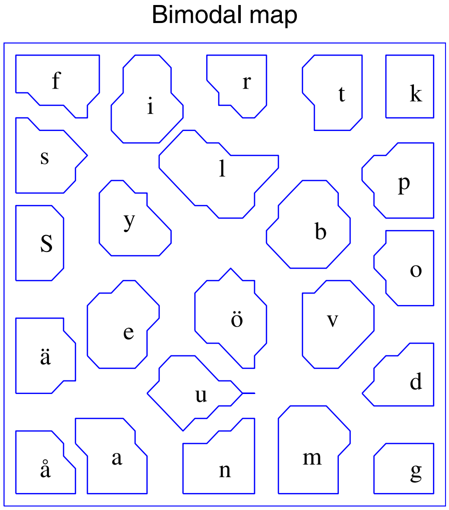

Fig. 4. Bimodal SOM$_{bm}$ map. Patches of highest activity for labeled letter/phoneme combinations after self-organization on a map of 36×36 neurons.

5 Robustness of the Bimodal Percepts Against Unimodal Disturbances

An important advantage of integration of stimuli from sensory-specific cortices into multimodal percepts in multimodal association cortices is that even large disturbances in the stimuli may be eliminated in the multimodal percepts. Our model has the same advantage, as can easily be demonstrated.

We choose to study the processing of the three letters **i**, **å** and **m** which are all uncorrupted. The corresponding phonemes **i**, **å** and **m** are heavily corrupted however, and these corrupted phonemes cause the activity on the phoneme map to move as shown in the left part of Figure 5. In the bimodal map, the right part of Figure 5, the activities have moved very little. The recognition of these bimodal percepts is much less influenced by the auditory corruptions than the recognition in the phoneme map. This holds for all other letter/phoneme combinations as well.

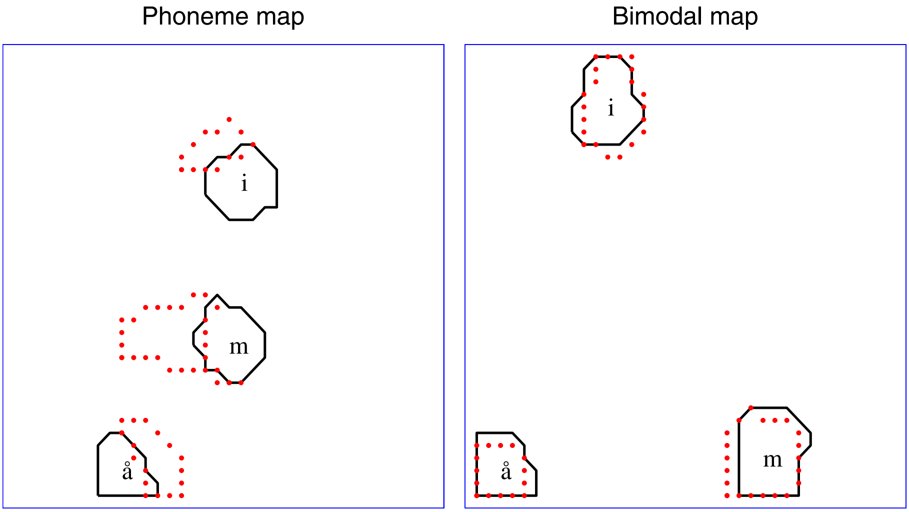

Fig. 5. The maximum activities, shown by solid lines in the maps when the inputs consist of different uncorrupted phonemes and by dotted lines when the inputs consist of heavily corrupted phonemes. Notice the difference in changes of activities in the phoneme map and in the bimodal map.

6 Introduction of Feedback and Its Significance for Auditory Perception

The robustness of the bimodal percepts, demonstrated in Figure 5, can be employed to benefit through feedback to enhance auditory perception, as is the case in cortex [15,22].

We introduce feedback in our MuSON through the re-coded phoneme map SOM$_{rph}$, see the right part of Figure 1. The 6-dimensional input stimuli to the re-coded phoneme map $[\mathbf{y}_{ph}\ \mathbf{y}_{bm}]$ is formed from the feedforward connection from the sensory phoneme map SOM$_{ph}$ and the top-down feedback connection from the bimodal map SOM$_{bm}$.

In the second phase of self-organization the re-coded phoneme map SOM$_{rph}$ is initialized to have the 23 winners in the same positions as in the phoneme map SOM$_{ph}$. The weight vectors are then trained by the Kohonen rule. Relaxation is included between two maps in the feedback loop. After self-organization the re-coded phoneme map SOM$_{rph}$ is seen to be similar to the phoneme map SOM$_{ph}$ as illustrated in Figure 6. However, the re-coded phoneme map through its feedback connections has a dramatically different property when phonemes are corrupted, as seen in Figure 7. The map shows post-synaptic activities for the three letters and phonemes, **i**, **å** and **m**, when the auditory and visual inputs to the sensory-specific maps are perfect (solid lines), and when the auditory inputs to the unisensory phoneme map are heavily corrupted (dotted lines). Note that when phonemes are corrupted the activities in the re-coded phoneme map change

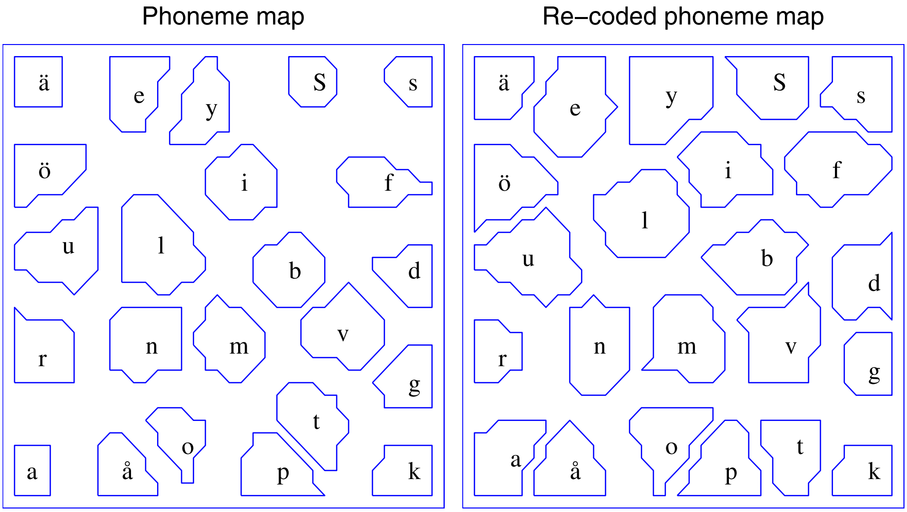

Fig. 6. The phoneme SOM_{ph} map and the corresponding re-coded phoneme SOM_{rph} map

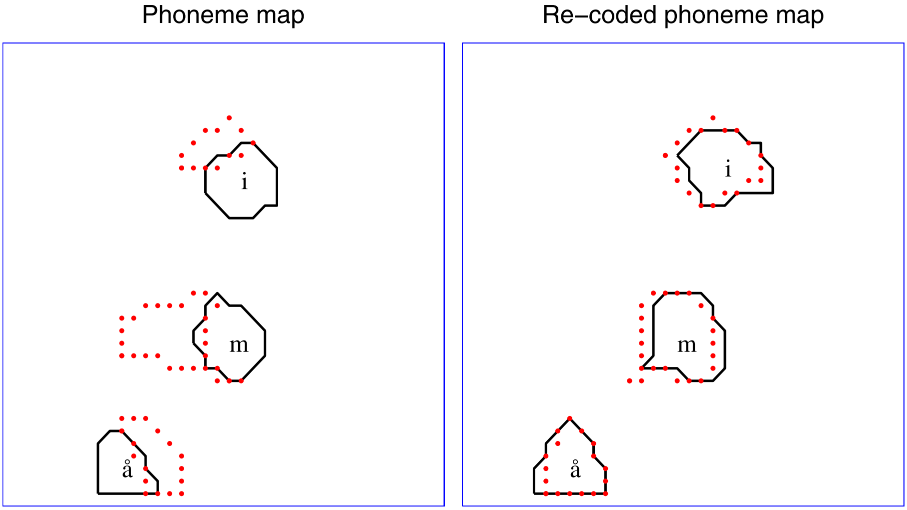

Fig. 7. The maximum activities for the three letters and phonemes **i**, **å** and **m**, shown by solid lines when the auditory and visual inputs to the sensory-specific maps are perfect and by dotted lines when the auditory inputs to the unisensory phoneme map are heavily corrupted

insignificantly compared to the activities in the phoneme map. This holds for all other letter/phoneme combinations as well.

7 Conclusion

With a Multimodal Self-Organizing Network we have simulated bimodal integration of audiovisual speech elements, phonemes and letters. We have demonstrated that the bimodal percepts are robust against corrupted phonemes and that when these robust bimodal percepts are fed back to the auditory stream the auditory perception is greatly enhanced. These results agree with known results from psychology and neuroscience.

Acknowledgment

The authors acknowledge financial support from the 2006 Monash University Small Grant Scheme.

References

1. Sumby, W., Pollack, I.: Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. **26** (1954) 212–215
2. Calvert, E.G., Spence, C., Stein, B.E.: The handbook of multisensory processes. 1st edn. MIT Press, Cambridge, MA (2004)
3. Schroeder, C., Foxe, J.: The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. Cognitive Brain Research **14** (2002) 187–198
4. Schroeder, C., Smiley, J., Fu, K., McGinnis, T., O'Connell, M., Hackett, T.: Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. Int. J. Psychophysiology **50** (2003) 5–17
5. Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P., Woodruff, P.W., Iversen, S.D., David, A.S.: Activation of auditory cortex during silent lipreading. Science **276** (1997) 593–596
6. Driver, J., Spence, C.: Crossmodal attention. Curr. Opin. Neurobiol. **8** (1998) 245–253
7. Calvert, G.A., Thesen, T.: Multisensory integration: methodological approaches and emerging principles in the human brain. J. Physiology Paris **98** (2004) 191–205
8. Foxe, J.J., Schroeder, C.E.: The case for feedforward multisensory convergence during early cortical processing. Neuroreport **16**(5) (2005) 419–423
9. Lamme, V., Roelfsema, P.: The distinct modes of vision offered by feedforward and recurrent processing. Trends Neuroscience **23** (2000) 571–579
10. Price, C.J.: The anatomy of language: contributions from functional neuroimaging. J. Anat. **197** (2000) 335–359
11. Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Springer, J.A., Kaufman, J.N., Possing, E.T.: Human temporal lobe activation by speech and nonspeech sounds. Cerebral Cortex **10** (2000) 512–528
12. Calvert, G.A., Campbell, R.: Reading speech from still and moving faces: The neural substrates of visual speech. J. Cognitive Neuroscience **15**(1) (2003) 57–70
13. Dehaene-Lambetrz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., Dehaene, S.: Neural correlates of switching from auditory to speech perception. NeuroImage **24** (2005) 21–33

14. Möttönen, R., Calvert, G.A., Jääskeläinen, I., Matthews, P.M., Thesen, T., Tuominen, J., Sams, M.: Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. NeuroImage **19** (2005)
15. Calvert, G., Campbell, R., Brammer, M.: Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology **10** (2000) 649–657
16. Frost, R., Repp, B., Katz, L.: Can speech perception be influenced by simultaneous presentation of print? J. Mem. Lang. **27** (1988) 741–755
17. Dijkstra, T., Frauenfelder, U.H., Schreuder, R.: Bidirectional grapheme-phoneme activation in a bimodal detection task. J. Physiology Paris **98**(3) (2004) 191–205
18. Gauthier, I., Tarr, M.J., Moylan, J., Skudlarski, P., Gore, J.C., Anderson, A.W.: The fusiform "face area" is part of a network that processes faces at the individual level. J. Cognitive Neuroscience **12**(3) (2000) 495–504
19. Polk, T.A., Farah, M.J.: The neural development and organization of letter recognition: Evidence from functional neuroimaging, computational modeling, and behavioral studies. PNAS **98** (1998) 847–852
20. Polk, T.A., Stallcup, M., Aguire, G.K., Alsop, D.C., D'Esposito, M., Detre, J.A., Farah, M.J.: Neural specialization for letter recognition. J. Cognitive Neuroscience **14**(2) (2002) 145–159
21. Raij, T., Uutela, K., Hari, R.: Audiovisual integration of letters in the human brain. Neuron **28** (2000) 617–625
22. van Atteveldt, N., Formisano, E., Goebel, R., Blomert, L.: Integration of letters and speech sounds in the human brain. Neuron **43** (2004) 271–282
23. Papliński, A.P., Gustafsson, L.: Multimodal feedforward self-organizing maps. In: Lecture Notes in Computer Science. Volume 3801., Springer (2005) 81–88
24. Gustafsson, L., Papliński, A.P.: Bimodal integration of phonemes and letters: an application of multimodal self-organizing networks. In: Proc. Int. Joint Conf. Neural Networks, Vancouver, Canada (2006) 704–710
25. Hopfield, J.: Neural networks and physical systems with emergent collective computational properties. PNAS USA **79** (1982) 2554–2588
26. Kohonen, T.: Self-Organising Maps. 3rd edn. Springer-Verlag, Berlin (2001)
27. Rolls, E.T.: Multisensory neuronal convergence of taste, somatosentory, visual, and auditory inputs. In Calvert, G., Spencer, C., Stein, B.E., eds.: The Handbook of multisensory processes. MIT Press (2004) 311–331
28. Kohonen, T.: The "neural" phonetic typewriter. Computer (1988) 11–22
29. Gold, B., Morgan, N.: Speech and audio signal processing. John Wiley & Sons, Inc., New York (2000)