Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007

Speaker-dependent Bimodal Integration of Chinese Phonemes and Letters Using Multimodal Self-organizing Networks

Sharon M. Chou, Andrew P. Papliński and Lennart Gustafsson

Abstract— **We present a model of integration of auditory and visual information as in the human cortex. More specifically, we demonstrate a possible way in which the phonetic symbols and associated Mandarin Chinese phonemes pronounced by different speakers are mapped onto the model of cortical areas. Our model has been strongly influenced by recent fMRI studies on integration of letters and speech sounds in the human brain. The model is based on multimodal self-organizing networks (MuSoNs) which were introduced in our previous works and proved to be a convenient tool to describe and study mapping and integration of sensory information as in the cortex. The model also shows how phonemes pronounced by different speakers are mapped onto overlapping cortical areas.**

I. Introduction

In this paper, which is a direct continuation of [1], we present a model of integration of auditory and visual information as in the human cortex. More specifically, we demonstrate a possible way in which the phonetic symbols and associated Mandarin Chinese phonemes pronounced by different speakers are mapped onto two-dimensional patches modelling the cortical areas. Our model has been strongly influenced by recent fMRI studies on integration of letters and speech sounds in the human brain as presented, for example, in [2], [3]. The model is based on multimodal self-organizing networks (MuSoNs) which were introduced in [4], [5], [1] and proved to be a convenient tool to describe and study mapping and integration of sensory information as in the human cortex. Such a mapping transforms multidimensional sensory stimuli into a two-dimensional "cortical" representation. Such a representation is unified for all modalities and performs abstraction (compression) of information. Our model demonstrates robustness of the bi-modal percepts and shows that humans hear a noisy phoneme better when they see the corresponding uncorrupted letter.

In our surrounding environment, many objects and events have different features that stimulate a variety of sensory information. This sensory information is integrated into humans brain to form unified multisensory percepts [6]. These percepts assist us with the recognition of objects and events, and enhance our ability to detect task-relevant stimuli. Review of the extensive studies of multimodal and bi-modal integration in the human brain can be found in [7].

In particular, it is well-known that the human brain has areas for processing auditory speech located in the sensory-specific auditory cortex and corresponding areas for processing visual speech located in the visual cortex [8], [2], [3]. Such brain areas are modeled by our self-organizing networks.

The processing of speech sounds for phoneme perception and associated phonetic symbols is the key element in the cognitive process of language. Our model demonstrates robustness of the bi-modal percepts and shows that humans understand a noisy phoneme better when they see the corresponding uncorrupted letter. We also show how phonemes pronounced by different speakers are mapped onto overlapping cortical areas.

Sharon M. Chou and Andrew P. Papliński are with the Clayton School of Information Technology, Monash University, Australia (email: {Sharon.Chou, Andrew.Paplinski}@infotech.monash.edu.au). Lennart Gustafsson is with the Dept. Computer Science and Electrical Engineering, Luleå University of Technology, Sweden (email: Lennart.Gustafsson@ltu.se)

II. Chinese Phonemes and Phonetic Symbols (Letters)

Languages differ in the degree to which visual symbols match specific phonological representations [9]. There is a good correspondence between letters or graphemes and phonemes in many written alphabetical languages, such as Italian, Russian or Swedish that was considered in [4]. Hence, once one understands how individual letters and their combinations are pronounced, one can correctly pronounce written words, even without understanding their meaning [9].

However, Mandarin Chinese is much different than other languages. Chinese characters are the symbols, each symbol, or character representing a morpheme, rather than phoneme. In some cases, this phonetic component is the same as the pronunciation of the character's meaning, but it many cases the character has a different pronunciation [9].

In addition, tones in Chinese are defined in terms of the rhythmic rise and fall of pitch, or the pitch contour of the voiced part of the characters, such that if the initial (onset) is voiced, the tone begins with the initial and spreads over the whole syllable, and if the initial is voiceless, the tone is spread over the final rhyme only. Tones are essential to pronouncing characters and deriving their meaning in reading. For example, the syllable */ma/* with tone 1 means *'mother'*, with tone 2 means *'linen'*, with tone 3 means *'horse'*, and with tone 4 means *'blame'* [10].

Furthermore, due to the above tonal fact, some characters must share pronunciations [11], for example, the following characters all pronounced as */li3/* (see Table I), 理(reason), 裡(inside), 娌(sister in law). Thus, pronouncing Chinese characters necessarily involves making reference to stored representations of each particular character, rather than assembling phonological sub-components into words [9]. In such cases, phonological processing, along with visual discrimination, is needed for reading Chinese [12]. In this

study, however, we restrict our considerations to phonemes and related phonetic symbols.

TABLE I
PINYIN AND BOPOMOFO EQUIVALENCE [13]

Consonants						Vowels					
b	ㄅ	玻	d	ㄉ	得	a	ㄚ	啊	ai	ㄞ	哀
p	ㄆ	坡	t	ㄊ	特	o	ㄛ	喔	ei	ㄟ	欸
m	ㄇ	摸	n	ㄋ	訥	e	ㄜ	鵝	au	ㄠ	熬
f	ㄈ	佛	l	ㄌ	勒	eh	ㄝ	耶	ou	ㄡ	憂
g	ㄍ	哥	j	ㄐ	基	an	ㄢ	安	er	ㄦ	兒
k	ㄎ	科	q	ㄑ	欺	en	ㄣ	恩	i	ㄧ	衣
h	ㄏ	喝	x	ㄒ	希	ang	ㄤ	昂	u	ㄨ	烏
						eng	ㄥ	亨	yu	ㄩ	迂
zh	ㄓ	知	z	ㄗ	資						
ch	ㄔ	蚩	c	ㄘ	雌						
sh	ㄕ	詩	s	ㄙ	思						
r	ㄖ	日									

There are two Chinese equivalent phonetic representation systems, Zhu-Yin-Fu-Hao and Pinyin as presented in Table I. Zhu-Yin-Fu-Hao constitutes a set of characters (phonetic symbols) used to annotate the Chinese sounds, and is primarily used in Taiwan [14]. It is also known as Zhuyin or Bopomofo (ㄅㄆㄇㄈ) based on the sounds the first four syllables of phonetic symbols [15]. Pinyin is a phonetic system that uses the English alphabet to represent phonetic symbols, and is mainly used in Mainland China. In Pinyin the English letter 'V' is not used, and three consonants are represented by combinations of two letters — ㄓ(zh), ㄔ(ch), and ㄕ(sh) [16], [15].

Both systems were designed on the same Mandarin dialect, and were introduced as an aid for children and as the phonetic symbols in dictionaries. There are 37 phonemes and corresponding phonetic symbols in Zhuyin or Pinyin. Hence those two systems have one to one relationship as presented in Table I.

In this paper, we use the Zhu-Yin-Fu-Hao symbols as visual representation of phonemes. However, for simplicity and the benefit of non-Chinese speakers, the PinYin symbols are used to annotate auditory representation of phonemes in our self-organizing maps.

III. AUDIO-VISUAL SELF-ORGANIZING NETWORKS

A multimodal self-organizing network (MuSoN) is built from interconnected modules, typically, Kohonen self-organizing maps (SOMs) [17] as the one presented in Figure 1. The network presented in Figure 1 is specialized in modelling integration of auditory and visual stimuli and will be referred to as an audio-visual self-organizing network (avSoN). It consists of four modules with a top-down feedback connection. There are two sensory level modules: visual SoM$_{lt}$ processing letters, and auditory SoM$_{ph}$ processing phonemes. Sensory modules perform mapping of the Bopomofo phonetic symbols and associated Chinese phonemes represented by multidimensional stimuli $\mathbf{x}_{lt}$ and $\mathbf{x}_{ph}$, respectively, into two-dimensional maps that represent

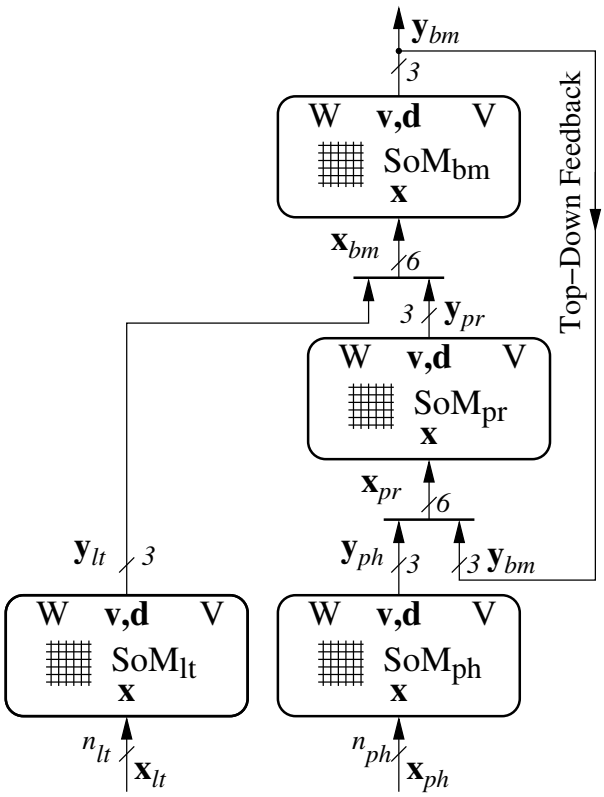

Fig. 1. An audio-visual self-organizing network (avSoN)

activated areas of cortex. The number of audio-visual stimuli is 37, as shown in Table I.

The outputs from the sensory modules, $\mathbf{y}_{lt}$ and $\mathbf{y}_{ph}$ represents the relative geometric position of the centre of the activated areas (position of the winner on the neuronal grid) and the related level of the post-synaptic activation. Efferent signals from the sensory modules form six-dimensional inputs to the bi-modal module, SoM$_{bm}$, $\mathbf{x}_{bm} = [\mathbf{y}_{lt} \quad \mathbf{y}_{pr}]$, where $\mathbf{y}_{pr}$ are different signals from the intermediate re-coded phoneme module, SoM$_{pr}$. The bi-modal module integrates stimuli from the visual and the auditory modules and produces its output, $\mathbf{y}_{bm}$, representing the position of the winning neuron and its relative activity.

The re-coded module is used to introduce the top-down feedback in the network and its inputs $\mathbf{x}_{pr} = [\mathbf{y}_{ph} \ \mathbf{y}_{bm}]$ are formed from the feedforward connection from the sensory phoneme module SoM$_{ph}$ and the top-down feedback connection from the bi-modal module SoM$_{bm}$.

It needs to be emphasized that the multi-modal network as presented in Figure 1 consists of four separate self-organizing maps/modules. It is not a hierarchical partitioning of a single SOM as in other works, e.g., [18], [19]. The response of each constituent module to a specific stimulus is, in our case, a 3-D vector representing a "cortical" position and the related strength, (see [1], [5] for details).

In our recent experiments we use the 40×40 neuronal grid in each module. That gives the realistic representation of 37 different audio-visual stimuli on the respective maps.

The learning process is based on the well-known Kohonen dot-product learning law, see [4], [5], [1] for details and is organized in two basic steps. Firstly, the maps are trained in the open loop, the phoneme module being connected directly to the bi-modal module ($\mathbf{y}_{pr} = \mathbf{y}_{ph}$).

In the second phase of self-organization the re-coded phoneme module SoM$_{pr}$ is initialized so that the re-coded map has the winners in the same positions as in the phoneme map SoM$_{ph}$. The weight vectors are then trained by the dot-product Kohonen rule. Relaxation is included between two modules connected in the feedback loop.

IV. Formation of the Sensory Maps

The data for the **visual letter map** has been obtained from the representation of 37 Zhuyin letters (ㄅㄆㄇㄈ) letters as shown in Table I. Each letter is first represented as a 33×24 binary image, and then, in order to reduce the dimensionality of the letter stimuli, a principal component analysis (PCA) is carried out, so that finally each $\mathbf{x}_{lt}$ stimulus is represented by a 37-component vector.

It is worth of noting that the initial PCA-based dimensionality reduction is used here purely for the computational efficiency. We have also tried two other approaches, namely, to excite the sensory letter module with the raw binary 33×24 vector, and to use directional filters (e.g. horizontal, vertical and diagonal) to compress initial visual information. In both cases, the resulting visual letter map have been perfectly acceptable from the point of view of their topological properties.

After self-organization is completed, which typically takes a few hundreds epochs, the resulting visual letter map SOM$_{lt}$ is similar to the one presented in Figure 2.

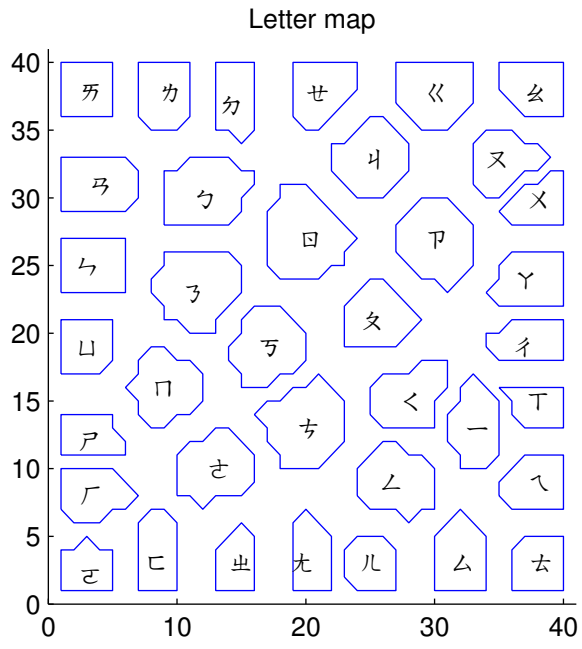

Fig. 2. Sensory letter map showing patches of high post-synaptic activities on 40×40 neuronal grids

The visual letter map shows the patches of high post-synaptic activity for each stimulus. The neuronal populations within the patches of the letter map constitute the detectors of the respective letters. The patches representing letters are topographically ordered so that letters that look alike are placed next to each other according to the shape of letter and its vertical or horizontal stroke. For instance, the three characters ㄉ(**d**), ㄌ(**l**), and ㄞ(**ai**) located at the upper-left corner of the letter map have the most similar stroke features. Therefore, they have been grouped together. The next similar characters are ㄢ(**an**) and ㄣ(**en**) are placed next to the previous group. While ㄢ(**an**) has similarity between ㄣ(**en**) and ㄞ(**ai**), therefore, its position is in between these two characters. It is easy to identify a number of similar letters placed together, for example, the group of the characters ㄘ(**c**), ㄎ(**k**) and ㄅ(**b**), and the group of the characters of ㄏ(**h**) and ㄕ(**sh**), ㄥ(**eng**) and ㄙ(**s**), ㄡ(**ou**), ㄨ(**u**) and so on.

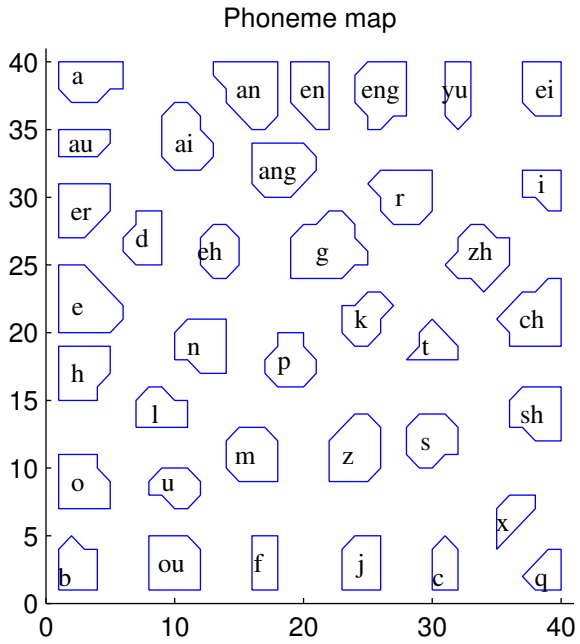

Fig. 3. Sensory phoneme map showing patches of high post-synaptic activities on 40×40 neuronal grids

The **auditory** material consists of 37 phonemes as spoken by two native Chinese speakers who read 37 syllables as shown in the third column of Table I. From each recorded syllable, the initial phoneme was extracted manually. Each phoneme consists of about a few thousand samples and is subsequently converted into thirty six melcepstral coefficients [20] using three overlapping time windows, each contributing twelve melcepstral coefficients.

During training of the auditory phoneme module/map SoMph, the vectors of the melcepstral coefficients were averaged over the speakers yielding one thirty-six element stimulus, $\mathbf{x}_{ph}$. After the learning process is completed a typical phoneme map has the form as shown in Figure 3. As in the letter map, neuronal populations in patches have

developed during self-organization as phoneme detectors.

The auditory phoneme map reflects to the articulatory features in Chinese speaking. Note, for example, the group of the similarly soundinh nasal finals ㄢ(**an**), ㄣ(**en**), ㄤ(**ang**) and ㄥ(**eng**) located next to each other in the upper part of the phoneme map. The grouping of phonemes with similar articulatory features also applies to palatal affricate consonants ㄐ(**j**), ㄑ(**q**) and ㄒ(**x**), the dental consonants ㄗ(**z**), and ㄘ(**c**) and ㄙ(**s**), as well as to the retroflex consonants ㄓ(**zh**), ㄔ(**ch**), ㄕ(**sh**) and ㄖ(**r**).

V. Bimodal integration of phonemes and phonetic symbols for different speakers

The bi-modal module S0M$_{bm}$ of the avSoN (see Figure 1) integrates information from the visual and and auditory modules and produces the map which captures similarities in both modalities. Typical bi-modal map produced by S0M$_{bm}$ is similar to the one presented in Figure 4.

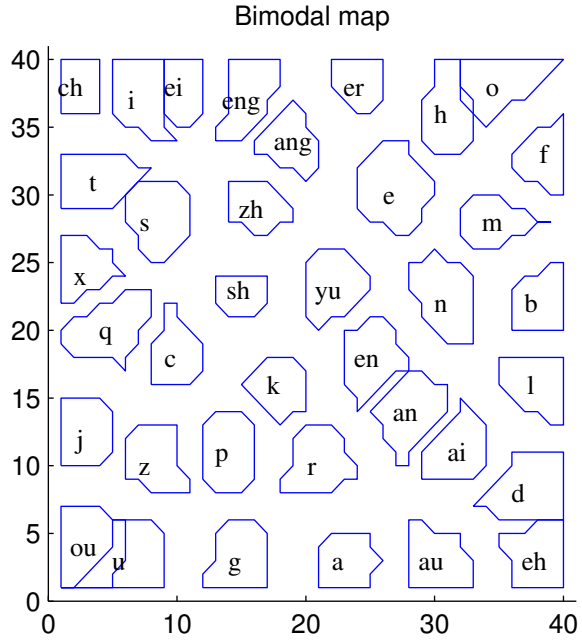

Fig. 4. The bi-modal map integrating visual (phonetic symbols) and auditory (phonemes) stimuli

Note, for example, from Figure 4, the cluster of phonemes/letters consisting of ㄢ(**an**), ㄞ(**ai**) and ㄣ(**en**) located in the lower right-hand side of the bi-modal map. The phoneme/symbol ㄢ(**an**) has visual features similar to ㄞ(**ai**) and auditory features similar to ㄣ(**en**) therefore they have been integrated into the cluster next to each other. Similarly, note the group in which the phoneme/letter ㄦ(**er**) has the visual features similar to ㄤ(**ang**) and the auditory features similar to ㄜ(**e**). Hence, ㄦ(**er**) has been categorized into the cluster between ㄤ(**ang**) and ㄜ(**e**).

This classification also applies to dental consonants group ㄗ(**z**), ㄘ(**c**) and ㄙ(**s**) located in the left-hand side of the bi-modal map. Note, ㄗ(**z**) has visual features (vertical stroke) similar to ㄐ(**j**), so ㄗ(**z**) has been categorised into the cluster next to ㄐ(**j**), same as ㄙ(**s**) and ㄊ(**t**).

Some groups such as ㄡ(**ou**) and ㄨ(**u**), and ㄅ(**b**), ㄌ(**l**) and ㄉ(**d**) are formed based predominantly on the similarity of the visual features. Conversely, some other groups like the one consisting of ㄐ(**j**), ㄑ(**q**) and ㄒ(**x**) are formed based on the predominance of the auditory similarity. Note also the similarly sounding phonemes ㄓ(**zh**), ㄗ(**z**) and ㄔ(**ch**), ㄘ(**c**) that have been shifted apart under the influence from the visual letter map.

The **re-coded phoneme map** is similar to the phoneme map, however, encodes also information from the visual modality through the feedback from the bi-modal module. It is best observed when the network trained for the auditory stimuli **averaged over speakers** is now exposed to auditory stimuli for the **individual speakers**. The resulting three maps: the phoneme map, the re-coded phoneme map and the bi-modal map are presented in Figures 5 and 6.

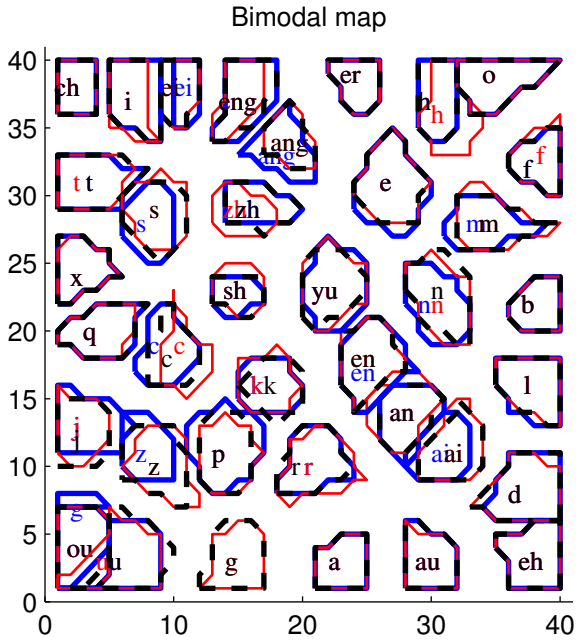

Fig. 5. Responses of the bi-modal map to auditory stimuli produced by individual speakers

In Figures 5 and 6 the black dashed lines represent responses for the average speaker's vector from the two speakers, the blue lines and red lines are used to represent each individual speaker respectively.

In order to demonstrate how the avSoN improves perception of distorted utterances produced by different speakers, we examine first in the sensory phoneme map in Figure 6 the three patches produced for the phoneme ㄐ(**j**). Note that the patches occupy significantly different, non-overlapping positions in the lower right-hand side of the phoneme map. This indicates potential difficulties in understanding such a phoneme. However, in the bi-modal map in Figure 5, where the output from the phoneme map is combined with the output from the letter map, the three patches for **j** (ㄐ)

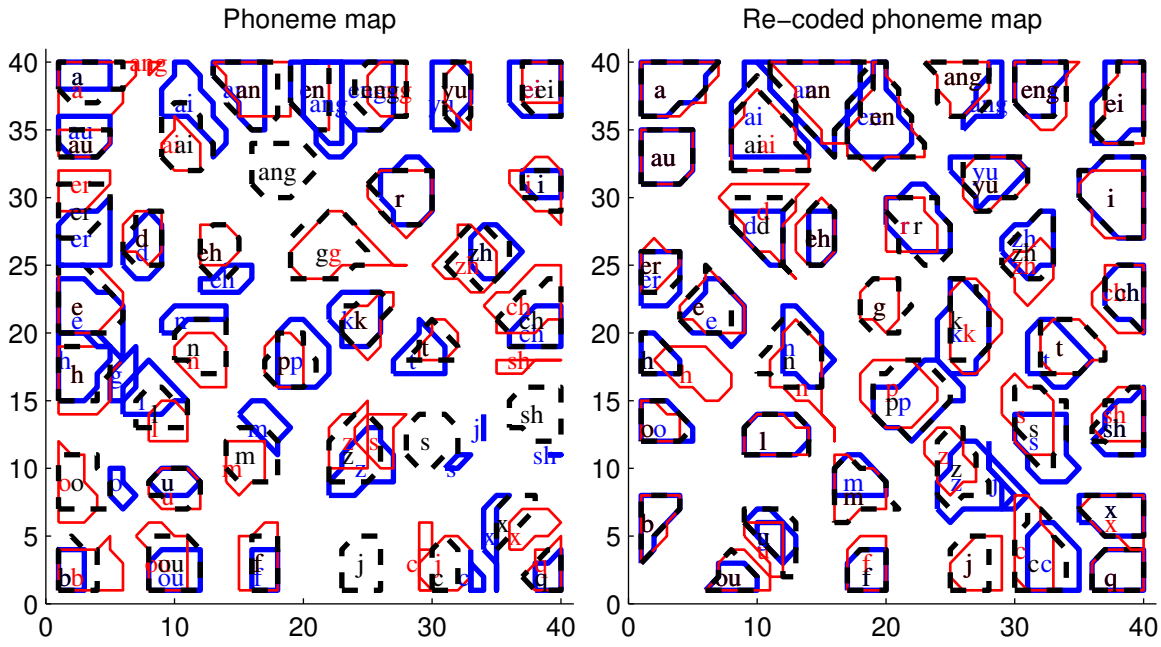

Fig. 6. Responses of the phoneme map and the re-coded phoneme map to auditory stimuli produced by individual speakers

overlap almost perfectly. Such results is now passed down through the feedback connection to the re-coded phoneme map, and the significant improvement in interpretation of two different utterances of **j** (ㄐ) can be observed in Figure 6.

Similarly, note in Figures 5 and 6 the non-overlapping triplets of patches for the phonemes ㄙ(**s**), ㄕ(**sh**) and ㄤ(**ang**). After combining in the bi-modal map the auditory material with the visual one, the relevant triplets of patches for the above phonemes in the re-coded phoneme map overlap very well which demonstrates a significant improvement in perception of those phonemes.

In addition to the four phonemes considered above not also that for the speaker 2, the pronunciation of ㄙ(**s**) is similar to ㄗ(**z**) and ㄔ(**ch**) is similar to ㄕ(**sh**) as well as ㄘ(**c**) and ㄐ(**j**). The ambiguity auditory input for the phoneme ㄙ(**s**) and ㄗ(**z**) is that both phoneme has similar articulatory features as dental fricative. The only different ㄗ(**z**) is unaspirated. The same as for ㄔ(**ch**) and ㄕ(**sh**), both phonemes are retroflex fricative, however ㄔ(**ch**) is retroflex aspirated affricate. Hence, it is difficult to distinguish between speaker and listener in the real world as well. For the speaker 1, ㄐ(**j**) and ㄘ(**c**) is slightly moved away from average speaker, and the ㄍ(**g**) is located far from the speaker 2 and the average speaker.

After the feedback process enhances the accuracy of auditory perception, for the speaker 2, the inaccurate perception of auditory input for the characters ㄙ(**s**), ㄗ(**z**), ㄔ(**ch**), ㄕ(**sh**) and ㄘ(**c**) as well as ㄐ(**j**) has improved after the feedback visual information is added. For the speaker 1, where ㄐ(**j**) has moved closer to the average and the ㄘ(**c**) is overlapping with Speaker 2 and average speaker. These results demonstrate how the audio-visual Self-organizing Network with the top down feedback improves the performance of speech recognition.

VI. Conclusion

Self-Organizing Networks consisting of Self-Organizing Modules/Maps can be conveniently used to model high level of the brain perceptual activities, in particular, integration of the visual and phonetic sensory information. During the self-organization process the sensory and multimodal modules form maps that represent their stimuli as two-dimensional cortical-like patches, each patch acting as the detector of the specific stimuli.

Our model demonstrates how phonemes pronounced by different speakers are mapped onto overlapping patches representing the cortical areas, the degree of similarity of the stimuli being expressed by the degree of overlapping of the relevant patches. The model also shows how understanding of the mispronounced phonemes improves when the visual modality supplies additional clues.

References

[1] 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.

[2] M. S. Beauchamp, "See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex," *Current Opinion in Neurobiology*, vol. 15, pp. 145–153, 2005.

[3] N. van Atteveldt, E. Formisano, R. Goebel, and L. Blomert, "Integration of letters and speech sounds in the human brain," *Neuron*, vol. 43, pp. 271–282, July 2004.

[4] A. P. Papliński and L. Gustafsson, "Multimodal feedforward self-organizing maps," in *Lect. Notes in Comp. Sci.*, vol. 3801. Springer, 2005, pp. 81–88.

[5] 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.

[6] W. Teder-Sälejärvi, F. Di Russo, J. McDonald, and S. Hillyard, "Effects of spatial congruity on audio-visual multimodal integration," *Journal of Cognitive Neuroscience*, vol. 17:9, pp. 1396–1409, 2005.

[7] E. G. Calvert, C. Spence, and B. E. Stein, *The handbook of multisensory processes*, 1st ed. Cambridge, MA: MIT Press, 2004.

[8] G. A. Calvert and R. Campbell, "Reading speech from still and moving faces: The neural substrates of visual speech," *Cognitive Neuroscience*, vol. 15, pp. 57–70, 2003.

[9] S. Bookheimer, "How the brain reads Chinese characters," *Neuroreport*, vol. 12, no. 1, Jan. 2001.

[10] C. K. Leong, P. W. Cheng, and H. T. Li, "The role of sensitivity to rhymes, phonemes and tones in reading English and Chinese pseudowords," *Reading and Writing*, vol. 18, pp. 1–26, 2005.

[11] L. Johnson, "Reading acquisition in Mandarin Chinese," 2005, http://convention.asha.org/2005/handouts/293_Johnson_Laura_071593_121405034828.pdf .

[12] W. Siok and P. Fletcher, "The role of phonological awareness and visual orthographic skills in Chinese reading acquisition," *Developmental Psychology*, vol. 37:6, pp. 886–889, 2001.

[13] K.-Y. Ho, "Mandarin," http://www.glue.umd.edu/~kwyho/phonetics_mandarin.htm .

[14] "East Asian scripts," The Unicode Consortium, 2006, http://www.unicode.org/versions/Unicode4.0.0/ch11.pdf.

[15] "Transliteration/Romanization Systems," 2003, http://www.eslisland.com/life/TransliterationRomanization-Zhuyin-Pinyin.htm .

[16] T. Xie, "A brief introduction to Romanized spelling system of Chinese — Pinyin," Dept. Asian and Asian American Studies, California State University, Long Beach, 2004, www.csulb.edu/~txie/pcr/soundsys/introtopinyin.htm.

[17] T. Kohonen, *Self-Organising Maps*, 3rd ed. Berlin: Springer-Verlag, 2001.

[18] P. Xu, C.-H. Chang, and A. Papliński, "Self-organizing topological tree for on-line vector quantization and data clustering," *IEEE Tran. System, Man and Cybernetics, Part B: Cybernetics*, vol. 35, no. 3, pp. 515–526, June 2005.

[19] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, "Dynamic self-organizing maps with controlled growth for knowledge discovery," *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 601–614, May 2000.

[20] B. Gold and N. Morgan, *Speech and audio signal processing*. New York: John Wiley & Sons, Inc., 2000.