Computer-Assisted Lip Reading Recognition for Hearing Impaired

Yun-Long Lay[1], Hui-Jen Yang[2], and Chern-Sheng Lin[3]

[1] Department of Electronic Engineering, National Chin-Yi University of Technology,
[2] Department of Information Management, National Chin-Yi University of Technology
No. 35, Lane 215, Sec. 1, Chung-San Rd., Taiping City. Taichung Hsien, Taiwan, 411
{yllay,yanghj}@ncut.edu.tw
[3] Department of Automatic Control Engineering, Feng Chia University
No.100, Wenhwa Rd, Seatwen, Taichung, Taiwan, 40724
lincs@fcu.edu.tw

Abstract. Within the communication process of human beings, the speaker's facial expression and lip-shape movement contains extremely rich language information. The hearing impaired, aside from using residual listening to communicate with other people, can also use lip reading as a communication tool. As the hearing impaired learn the lip reading using a computer-assisted lip reading system, they can freely learn lip reading without the constraints of time, place or situation. Therefore, we propose a computer-assisted lip reading system (CALRS) for phonetic pronunciation recognition of the correct lip-shape with an image processing method, object-oriented language and neuro-network. This system can accurately compare the lip-image of Mandarin phonetic pronunciation using Self-Organizing Map Neuro-Network (SOMNN) and extension theory to help hearing impaired correct their pronunciation.

Keywords: hearing impaired, Self-Organizing Map Neuro-Network, extension, lip reading reorganization.

1 Introduction

In research, the McGurk effect displays that the combination of sound and image would affect people's cognition of correct word utterance [2]. For instance, the Mandarin pronunciation is "ba"(巴) and the corresponding image is "ga"(嘎), then people will perceive the combination of sound and image as "ta"(他). Obviously, sound and image can simultaneously affect people's cognition of language. Moreover, many pronunciations have their unique sound and image characteristics. For instance, the Mandarin pronunciations of "ba" (爸) and "da" (搭) are easy to distinguish by lip image, but is actually not easy to distinguish by the sound. On the contrary, the pronunciations of "ba"(爸) and "bai" (掰) are easily distinguished by the sound, but not easily distinguished by the lip image. Thus, in the lip reading teaching system, the pronunciation and lip image may complement one another [2].

Although the lip-reading recognition is not absolutely accurate and reliable, normal-hearing people and hearing impaired people actually have to use it every day. If

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2009, LNCS 5616 pp. 559–568, 2009.

hearing impaired people want to enhance their communication with other people aside from wearing hearing aid instrument, lip reading recognition is one essential learning method. According to a study of the University of Manchester [8], hearing impaired people can normally use residual hearing to recognize 21% of spoken words. If the hearing impaired use hearing aid instruments or lip recognition to properly support their communication, the recognition rate of these words would be 65%. Moreover, if hearing impaired people use hearing aids and lip reading simultaneously, the recognition rate of these words would reach 90%. Obviously, lip reading recognition is important for hearing impaired people.

In order for hearing impaired people to conveniently understand pronunciation, the purpose of this research is to propose a computer-assisted lip reading system for lip reading recognition to help hearing-impaired people quickly understand pronunciation by lip reading. A dynamic lip reading image of a phoneme was grabbed, then through neuro-network and extension techniques each phoneme in the database can quickly be searched for. After the pronunciation of the hearing impaired, the system will quickly pick up the change of lip-shape by computer to help hearing impaired people modify the lip-shape while pronouncing a Mandarin phoneme. A Self-Organizing Map Neuro-Network (SOMNN) technique is used to classify and find related position of the image signal. This system also combines extension method to complete the lip recognition by the computer system for higher recognition rate. The hearing impaired can self-evaluate their pronunciation and lip reading accurately to enhance their language learning and communication capability.

2 System Structure and Research Method

This research is to implement a computer-assisted lip reading system to help hearing impaired people to study the phonetic pronunciation with a lip-reading recognition technique. Three components (the pre-process of lip-shape image separation, Neuro-Network characteristic extraction, and extension technique) were included in this system and shown in Fig. 1.

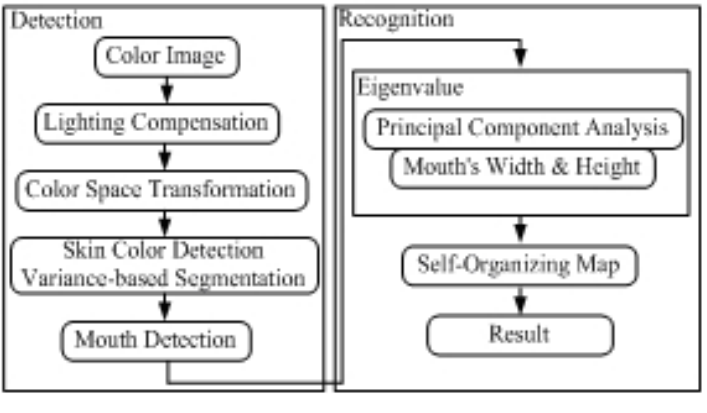

Fig. 1. The framework of lip reading recognition system

2.1 Self-Organizing Map Neuro-Network (SOMNN)

Self-Organizing Map Neuro-Network (SOMNN), proposed by Kohonen, belongs to the forward and non-supervised network. It is a competitive neuro-network, too [3], [5]. The self-organizing network algorithm uses the characteristic mapping method with free dimension input vectors to map into the characteristic-mapping graph of lesser dimension. According to the current input vectors mutually competing among the neurons, the winning neuron may obtain the chances of the adjustment to link the weight vector. The final output layer neuron would display in the output space with significant topological structures based on the input vector features.

2.2 Extension Theory

Extension was originally called a "matter-element analysis" and was established in 1983 by Professor Wen Tsai to solve the contradictory problems in the real world. Its basic concept is to use the structure and transformation of matter-elements to respond with the mutual transformation relations of "quality" and "quantity".

Extension is a formalized tool. From the perspectives of "quality" and "quantity", this theory solves the regulation and methods of ambiguity problems. The foundation of extension is a matter-element theory. In classical mathematics, extension uses observation of the data's quality and shape to express an objective formula. However, if you want to solve the ambiguity problems, these conditions are still not enough. Therefore, in order to solve problems of ambiguity, it needs to have an object, the characteristic of the object, and corresponding amount-values taken into consideration [9], [10].

The Basic Theory of a Matter-Element. In extension, a matter-element includes three primary elements. Suppose the name of object "*R*" is called "*N*", its characteristics "*c*", and its amount-value of characteristic *c* "*v*". Then the relationship of the basic-element or the mater-element is

$$R = (N, c, v) \tag{1}$$

If we assume an object $R= (N,c,v)$ is a matter-element of multiple dimensions, we can use array $C = [c_1,c_2,\ldots c_n]$ to represent when the object has n characteristics. Its corresponding amount-value is represented by value matrix $V = [v_1,v_2,\ldots v_n]$. The formula of a matter-element is

$$R = (N, C, V) = \begin{bmatrix} R_1 \\ R_2 \\ \cdots \\ R_n \end{bmatrix} = \begin{bmatrix} N, & c_1, & v_1 \\ & c_2, & v_2 \\ & \cdots & \cdots \\ & c_n, & v_n \end{bmatrix} \tag{2}$$

The formula of a matter-element can be notated as equation (5).

$$R= (N,C ,V) \tag{3}$$

3 Experiment Procedure

The system implementation uses Borland C++ Builder 6 and MATLAB 7 to develop. The hardware parts include a personal computer, CCD camera and Picolo image card to process the lip recognition. The experiment procedure includes pre-processing, the module of extension self-organizing neuro-network recognition and experiment results.

3.1 Pre-processing

After the face image is extracted from the color CCD, shown in Fig.2, the most difficult step is how to separate the lip-shape image from the complex background image. During the separation process, the lips are red and their color is possibly affected by the environment noise which would cause the lips' color change and be inconsistent. Therefore, we rely on the main color R(red) of the lips pattern to process (assume R: 120; G (green) <60 or G>120; B (blue) <40 or B>120) and make the original red color more obvious, and the green and the blue colors clear [11]. Thus, the separation of lip-shape image will not affect its accuracy because of these external factors. The quality of whole recognition system performance will be increased. From chromatics theory, color is composed of RGB (three original colors). So, if we want to constitute red, its R-value is obviously higher than its B and G values.

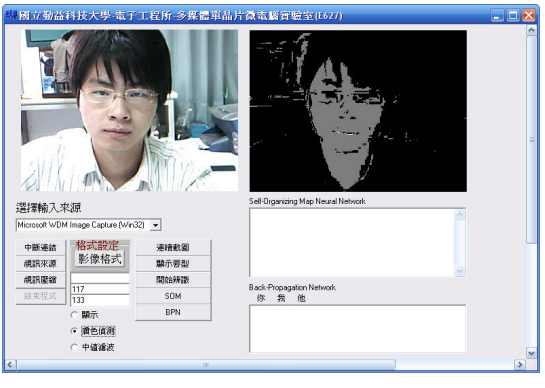

Fig. 2. The original image after CCD extraction

Taking the red area to get the image, the lip-shape image uses template matching to search for the ROI (region of interest). Fig. 3a is the graph of the ROI treating process [6]. Thus, the separated lip-shape image's corresponds to the original image coordinate position to process the transformation to grey-scale. After the transformation, the grey-scale image undergoes a shrinking process, shown in Fig. 3b. The previous obtained lip-shape image would become the lip recognition process's input data for Self-Organizing Map Neuro-Network.

3.2 The Module of Extension Self-Organizing Map Neuro-Network Recognition

After the pre-process transformation of all phonemes lip-reading image, the obtained input lip-shape image would be sent into the module of extension-SOMNN to

(a)

Fig. 3a. The graph of the ROI lip-shape extraction

(b)

Fig. 3b. Lip-shape image after the shrinking process

process. In this module, the front end uses SOMNN to process the characteristic extraction of lip reading image data. The back end uses the extension technique to deal with the distinction between the phoneme database and the inputted lip reading image. The detailed algorithm is as follows:

Step 1: setup the network parameters

Step 2: read weight value matrix *W*. This *W* is the weight value after training all phonetic notation images. It sets also up the topology coordinate of the output layer.

$$XNode\ [j]\ [k] = j$$

$$YNode\ [j]\ [k] = k \tag{4}$$

Step 3: input lip reading image of all phonemes as input vector *X*.

Step 4: find the superior output processing unit *Node* [*j**][*k**]

$$net[j][k] = \sum_{i} (X[i] - W[i][j][k])^2 \tag{5}$$

Where *W* [*i*] [*j*] [*k*] represents as the linking weight value of *i*th input unit, the network topology coordinate *k*th and the net value's smallest hidden layer process unit *Node* [*j**][*k**]. That is:

$$Node[j^*][k^*] = \min_{j \quad k} net[j][k] \tag{6}$$

Step 5: calculate output vector *Y* in the output layer and access the corresponding index value of topology.

$$\begin{cases} Y[j][k] = 1, & j = j^*; k = k^* \\ Y[j][k] = 0, & other \end{cases} \tag{7}$$

After the front-end process, the index value of the continuing lip-shape image can be obtained. Relying on these values and carrying them into the back-end Extension method processes, the distinction among the phoneme database to map the corresponding phonetic pronunciation. The recognition process is as follows:

Step 1: set up the element module of each phonetic pronunciation type.

$$R_i = \left(PN_i, C_i, V_{ij}\right) = \begin{bmatrix} PN_i, & c_1, & v_{i1} \\ & c_2, & v_{i2} \\ & \cdots & \cdots \\ & c_n, & v_{in} \end{bmatrix}, i = 1, 2, 3....10 \tag{8}$$

PN_i is the type of phonetic pronunciation, C_i is the characteristic number of the input lip image, $V_{ij} = (a_{ij}, b_{ij})$ is the classical domain of each lip image characteristic set, and $V_{pj}' = (a_{pj}, b_{pj})$ is the node domain of each lip-shape image characteristic set.

Step 2: representing the signal of waiting lip-shape image uses the following formula:

$$R_t = \left(PN, C_i, V_{ti}\right) = \begin{bmatrix} PN, & c_1, & v_{t1} \\ & c_2, & v_{t2} \\ & \cdots & \cdots \\ & c_n, & v_{tm} \end{bmatrix}, i = 1, 2, ..., 10 \tag{9}$$

Step 3: from the phonetic pronunciation related function equation (11-13) to calculate the relative degree of the lip-shape image. The equation is as follows:

$$K_{ij}\left(v_{tj}\right) = \begin{cases} \dfrac{-\rho\left(v_{tj}, V_{ij}\right)}{\left|V_{ij}\right|} & , v_{tj} \in V_{ij} \\ \dfrac{\rho\left(v_{tj}, V_{ij}\right)}{\rho\left(v_{tj}, V_{pj}\right) - \rho\left(v_{tj}, V_{ij}\right)} & , v_{tj} \notin V_{ij} \end{cases} \tag{10}$$

where

$$\left|V_{ij}\right| = \left|b_{ij} - a_{ij}\right| \qquad i = 1, 2, ..., 5\ ;\ j = 1, 2, ...10 \tag{11}$$

This research proposes the extension correlating function. When $0 \le K\left(v\right) \le 1$, it means v belongs to V. When $K(v) < 0$, it means v does not belong to V.

Step 4: based on the importance of each correlating function for lip-shape image recognition, the weight $\alpha_{i1}, \alpha_{i2}, \ldots \alpha_{i10}$ is set up. In this research, the weight value is 1/10.

Step 5: to calculate each lip-shape image's correlated level.

$$\phi_i = \sum_{j=1}^{10} \alpha_{ij} K_{ij} \ , i = 1, 2, ..., 5 \tag{12}$$

Step 6: searching for the maximum value of ϕ_i for the final lip-shape image recognition results. The equation is as follows:

$$\phi_{\max} = \max_{1 \le i \le 5} \{\phi_i\} \tag{13}$$

The primary advantage of the method proposed in this research is that the non-supervised learning method of SOMNN may automatically classify different characteristic values for all lip-language image data. Additionally, the operation process of the neuro-network does not need to increase the hidden layer in order to reach the convergent effect between the input signal and weight value. However, extension, based on the foundation of element theory, can aim at the recognition problems in the fuzzy area to make suitable distinctions and processing. For instance, the extension correlating function range of $0 < K(x) \le -1$ is not complicated in math, so the operating speed does not comparatively waste time. Therefore, relying on this characteristic information, extension can directly find out the lip-shape image related levels of all Mandarin phonetic pronunciation images and each related level of pronunciation lip language images.

4 Experiment Results

After the aforesaid experiment procedure, extension-SOMNN was applied to process the lip reading recognition of phonemes. In the examples, ten sets (ㄆ,ㄈ,ㄍ,ㄐ,ㄓ,ㄚ,ㄟ,ㄦ,ㄧ,ㄨ) of lip-reading images of phonemes, which the lip-shape is obviously changing, were chosen as training phonemes. After the system implementation, we test the recognition rate of each phoneme and total recognition rate of the system. In the experiment process, through initially CCD extracting the dynamic lip-reading image, the image size is 320 x 240 pixels. Ten important dynamic lip-reading image pictures of each phoneme were selected to process the recognition. Fig. 4 displays the dynamic lip-reading images of phoneme "ㄈ" ("fo") grabbed by CCD. After the extraction, the lip reading image would be in order from the pre-processing method to carry on the deduction sampling to the 22 x 19 pixel size.

After the completion of all lip-reading images in pre-process, SOMNN was applied to process the characteristic extraction of the lip reading image. Thus, each lip reading would have ten characteristic values of lip-shape images. For instance, Fig. 4 is the topological distribution graph with the lip reading image of phoneme "ㄈ" after SOMNN training.

The trained characteristic values were used to set up the lip-shape image database. The lip reading characteristic values of each phonetic pronunciation through the matter-element model of extension shown as R_{ij} (classic domain and node domain). In the lip-reading matter-element model, it uses Mandarin phoneme notation as the "Name" of the matter-element, the number of dynamic lip-images extracted as the "Characteristics" of the matter-element and through the training of SOMNN, get the index coordinate values as the "Value" of the matter-element. The setting of measured value range adopts the index coordinate value $\pm 10\%$. For example, the value of the 1st set

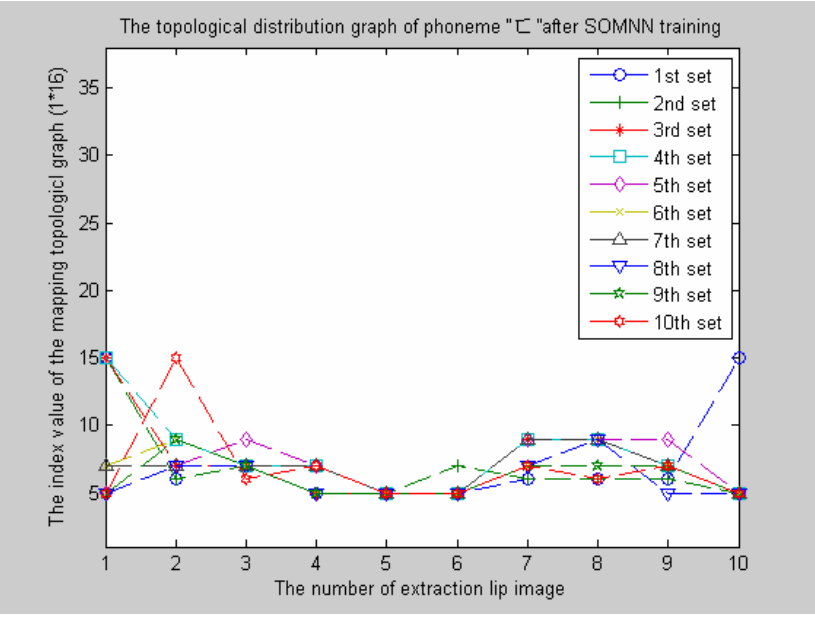

Fig. 4. The lip-reading dynamic images of phoneme "ㄈ" after the reduction sampling

of *R01* is 5 calculated by SOMNN. After the modification by the measured value range, the value is changed to 4.9~5.1. The measured value point is decided by the index coordinate value in the most number of each phoneme within the 100 group-training database.

As for testing the 100-group lip reading images, Table 1 has shown the lip reading image testing data after SOMNN calculation. Each phoneme has ten sets of data, and each piece of data is the topological distribution of the testing phoneme images, which is the input data of extension. The most highly correlated data has the maximum value and is highlighted with a shaded background. Finally, the total recognition rate of the system is 91.6%.

Table 1. The recognition results of lip reading images with 10-phoneme notation

Phoneme recognition						Total recognition rate
Phoneme	ㄆ	ㄈ	ㄍ	ㄐ	ㄓ	91.6%
Recognition rate	98%	98%	99%	70%	81%	
Phoneme	ㄚ	ㄟ	ㄦ	ㄧ	ㄨ	
Recognition rate	97%	96%	98%	97%	82%	

5 Conclusion

This research provides a lip-reading recognition system with the combination of SOMNN and extension theory. The main advantage of this research is that the non-supervising learning of SOMNN can automatically classify lip-reading image data into different characteristic value without increasing the hidden layers during the operation to get the convergent effect between the input signal and weighting values. Extension is based on the foundation of matter-element theory to recognize the fuzzy area for appropriate distinguishing and operating. It does not invoke complicated mathematics, so it will save some time during operating time. Based on the characteristic information, the relationships of 10 phonemes for Mandarin phoneme lip-shape images and individual phoneme lip-shape images were found by extension methodology. The total recognition rate is 91.6% in this research as shown in Table 1.

From the experiment results, the performance of the recognition system is only good enough for 10 phonemes, not for all 37 Mandarin phonemes. If the recognition process can be supplemented with sound recognition and high-speed photography instruments, all 37 phonemes can possibly have good recognition rates for advanced study.

Acknowledgement. Authors thank the National Science Council support this research. The research number is NSC 95-2221-E-167-001 and NSC96-2221-E-167-026-MY3.

References

1. Black, A., Taylor, P.: Festival Speech Synthesis System: System documentation (1.1.1), Human Communication Research Center Technical Report HCRC/TR-83, Edinburgh (1997)
2. Green, K.P., Gerdeman, A.: Cross-Modal Discrepancies in Coarticulation and the Integration of Speech Information: the McGurk Effect with Mismatched Vowels. Journal of Experiment Psychology: Human Perception and Performance 21(6), 1409–1426 (1995)
3. Hatzilygeroudis, I., Prentzas, J.: Integrating (Rules, Neural Networks) and Cases for Knowledge Representation and Reasoning in Expert Systems. Expert Systems with Application 27(1), 63–75 (2004)
4. Ju, Y., Yu, Y., Ju, G., Cai, W.: Extension Set and Restricting Qualifications of Matter-elements' Extension. In: IEEE CNF, vol. 1, pp. 395–398 (2005)
5. Kohonen, T.: The Self-Organizing Map. Proc. IEEE 78(9), 1464–1480 (1990)
6. Kou, X., Wang, Z., Chen, M., Ye, S.: Fully Automatic Algorithm for Region of Interest Location in Camera Calibration. Optical Engineering 41, 1220–1226 (2001)
7. Lu, Q., Yu, Y.: The Research of Data Mining Based on Extension Sets. In: IEEE CNF, vol. 2, pp. 234–247 (2005)
8. Bauman, N.: Speechreading (2003), http://www.hearinglosshelp.com/articles/speechreading.htm
9. Wang, M.-H., Ho, C.-Y.: Application of Extension Theory to PD Pattern Recognition in High-Voltage Current Transformers. In: IEEE JNL, vol. 20, pp. 1939–1946 (2005)

10. Wang, M.-H.: A Novel Extension Method for Transformer Fault Diagnosis. In: IEEE JNL, vol. 18, pp. 164–169 (2003)
11. Wang, S.L., Lau, W.H.: A Real-Time Automatic Lipreading System. IEEE CNF 2, 101–104 (2004)
12. Yang, H.-J., Lay, Y.-L.: The Implementation and Evaluation of Computer-Aided Mandarin Phonemes Training System for Hearing-Impaired Students. British Journal of Educational Technology 36(3), 537–551 (2005)