Expert Systems with Applications 36 (2009) 4804–4809

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

The radar-graphic speech learning system for hearing impaired

Hui-Jen Yang [a,*], Yun-Long Lay [b], Chern-Sheng Lin [c], Pei-Yuan Hong [b]

[a] Department of Information Management, National Chinyi University of Technology, No. 35, Lane 215, Sec. 1, Chung-San Road, Taiping City, Taichung Hsien 411, Taiwan
[b] Department of Electronic Engineering, National Chin-Yi University of Technology, No. 35, Lane 215, Sec. 1, Chung-San Road, Taiping City, Taichung Hsien 411, Taiwan
[c] Department of Automatic Control Engineering, Feng Chia University, No. 100, Wenhwa Road, Seatwen, Taichung 40724, Taiwan

ARTICLE INFO

Keywords:
Radar map
Hearing impaired
Neural network
Radar-graphic displaying system

ABSTRACT

Speech learning is an important foundation for literacy ability. In general, language training needs a professional instrument to analyze speech for supporting the pronunciation of hearing impaired. However, non-professional speech spectrum equipment is very expensive and its output is not easy for hearing impaired to understand and learn. The purpose of this research is to implement a radar-graphic displaying system (RDS) to support the speech learning for hearing-impaired people at low cost and better performance. The components of RDS include a computer connects to a microphone as input device to capture the speech features; a neural network is used to extract the features for speech recognition; a radar map displays the voice message on the screen to support the hearing impaired to learn speech. A system performance evaluation of RDS was performed after the system was implemented.

1. Introduction

Hearing ability is based on the onset of hearing loss and can be divided into congenital inheritance and afterward occurred. If the hearing ability occurred afterwards, then the hearing-impaired people still possess language ability, but they can not clearly hear what people say. If the hearing-impairment is a congenital inheritance, they would have hard time to learn speech. According to the critical period hypothesis (CPH), children have an idea period to learn speech and language. The best time of language learning is from age 8 to 12 (Johnson, Christie, & YawKey, 1998; Newport, 1991). After the CPH, to learn language becomes a difficult issue for children. Thus, the hindrance of hearing and language will affect individual's learning performance and recognition. If hearing-impaired people can not receive good education and be guided properly, they can not communicate with people. It is hard for hearing impaired people to get a job. They will feel frustrate. Thus, it will increase the social load and social problems. Based on the statistical data, one of one-thousand newborn babies is sick with the congenital hearing handicap, which has much relation to recessive inheritance. Hearing-impaired people are not completely deaf and still have some residual hearing. They can not use their own auditory ability or hearing-aid devices to monitor their pronunciation correctly. Language learning of hearing-impaired people depends on their feeling or guesses to determine the pronunciation, so they can not pronounce properly and correctly. Currently there is no effective method or tool to prevent and remedy hearing loss. Therefore, implementing a computer-aided learning system to improve their language learning and training is an effective and convenient method.

The radar-graphic displaying system for speech learning (RDS), implemented in this study is suitable for hearing-impaired students to learn Chinese pronunciation. The system includes a computer with graphic-display screen and microphone to increase the learning performance for hearing-impaired students. This system is designed to simplify the process of language learning through the home computer. The system calculates the speech features through voice recognition technique to compare single-word sounds with the sounds which are pronounced by hearing-impaired users. Neuro-network analysis was used to compare each single word in the database allowing a larger range of tolerance. Each feature of single words with 2-dimensional coordinates display on the screen, which is visually similar to a radar map with a fixed position. Hearing-impaired students can operate the system by themselves. They can use a microphone to extract the signal through the graphic-displaying mode to see whether their pronunciation is correct or not. Hearing-impaired students can self-monitor and self-modify their pronunciation. This system is suitable for low-grade primary school students to learn proper Chinese pronunciation.

2. System framework

Speech recognition initially should extract the speech features to establish the speech feature database for neural network training data (Rabiner & Juang, 1993). Back propagation neural network

* Corresponding author. Tel.: +886 4 23924505x7923/7912; fax: +886 4 23923725.
E-mail address: yanghj@ncut.edu.tw (H.-J. Yang).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.05.053

(BPN) and self-organizing maps (SOM) are applied in this system. BPN is used to recognize the speech and SOM is used to create the radar map. The speech features database is separately delivered to BPN and SOM for training. BPN gets the adjusted weighting values and SOM obtains the speech features. The trained weighting values are then delivered to BPN as a testing recognition reference. The 2-dimensional topology is transferred into the radar chart coordinates distribution graph. When the user operates the system, a microphone is used to input the speech signal and then the software extracts the speech features and sends them to BPN for recognition. The result is displayed in the radar map. The system framework is shown in Fig. 1.

3. Feature extraction

During the speech recognition process, speech recognition extraction is the first step to develop the speech feature database as a neural-network training database. An efficient method to represent the appropriate feature parameters is necessary to process the speech data. The obtained feature parameters will indirectly affect the speech recognition rate, through the sampling process speech signals are extracted from soundcard inside the PC. The sampling rate of the system is 8 K/s. The resolution is 8 bits, which means the speech signal waveform is 8000 dots per second. The extraction of speech features must go through some complicated processes. The sampling process initially goes through the starting-point and the ending-point to detect the start and end position of speech. The speech signal interval is every 15 ms to construct the sound-frame, following with the procedures including endpoint detection, segmentation, pre-emphasize, Hamming window, autocorrelation, LPC analysis and Cepstrum to get the speech feature coefficients (Patil, 1998). Fig. 2 shows the extraction flowchart of speech features. The speech data is represented by the suitable features parameter. Generally, the size of speech data is very large and can not be stored as the reference sample for speech recognition. The speech features parameter is replaced by Cepstrum coefficients.

3.1. Energy

In general, the speech segmentation method uses energy and zero-crossing rate. $E(n)$ can calculate the speech signal variability in a short-time to determine voice or non-voice.

$$E(n) = \sum_{m=0}^{N-1} [W(m)X(n-m)]^2 \tag{1}$$

Among which N is the length of a short-time signal, $W(m)$ is the window to choose a specific short-time pronunciation signal $X(n)$.

3.2. Zero-crossing rate

Zero-crossing rate determines the speech voice or non-voice. We can calculate the speech signal zero-crossing frequency and energy detection endpoint for segmentation.

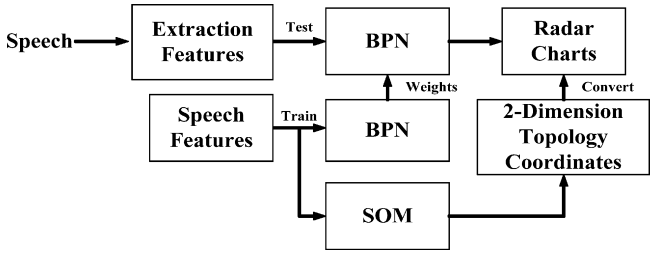

Fig. 1. Framework of the graphic speech learning system.

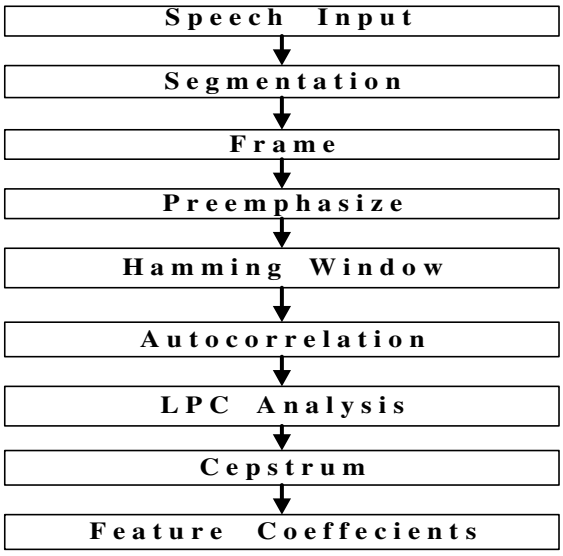

Fig. 2. Speech feature parameter extraction flowchart.

$$Z_x(m) = \frac{1}{N} \sum_{n=m-N+1}^{m} |S[x(n)] - S[x(n-1)]| \tag{2}$$

S: sign function, if neighbor signal is different from S then S is 2 else is 0.

3.3. Pre-emphasize

Speech signal spectrum frequency follows a descending 6 dB/oct. On the contrary, the auditory level has ascending 6 dB/oct characteristic. For effectively simulating the ear's automatic gain control (AGC) function, speech signal must be processed with pre-emphasize achieve +6 dB/oct.

$$H(Z) = 1 - aZ^{-1}, 0.9 \le a \le 1 \tag{3}$$

where a is the pre-emphasis parameter.

This processing is usually obtained by filtering the speech signal with a first order FIR (finite impulse response) filter whose transfer function in the z-domain.

3.4. Hamming window

When processing, speech signals must be divided into individual frames, with each frame about 20–30 ms and no overlaps. The sampling rate is 240 dots/s. For a more effective speech frame, we must add a Hamming window. Hence, the endpoint of each frame keeps the same characteristics as the middle part and gets rid of the discontinuities of the speech signal on the both endpoints. The Hamming window is calculated by Eq. (4).

$$W(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right), & 0 \le n \le N-1 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

3.5. Autocorrelation

Autocorrelation finds the speech signal waveform structure after the Hamming window process.

$$\phi_i(k) = \sum_{n=0}^{N-1} X(n+i)X(n+i+k) 0 \le k \le N-1 \tag{5}$$

Eq. (5) is a speech starting from i, $0 \le k \le N-1$, n is the signal length 20–40 ms.

3.6. Linear prediction code (LPC) analysis

LPC analysis decreases the errors of actual and predicted speech signals. The method is to measure the pitch period and resonance frequency and gets the useful speech parameters quickly.

$$S(n) = \sum_{k=1}^{p} \alpha_k S(n-k) + GU(n) \tag{6}$$

where U_n is a digital filter, G is the digital filter amplitude gain, p is the LPC prediction order and α_k is the LPC coefficients.

3.7. Cepstrum

Cepstrum is a method to convert the speech signal spectrum characteristics from the detailed variation and peaks of the waves. The peaks of a speech signal's wave appear in the low Cepstrum and the detailed variation appears in the high Cepstrum. Eq. (7) is the Cepstrum coefficient.

$$C(\tau) = F^{-1} \log |X(k)| = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{j\frac{2\pi kn}{N}}, \quad 0 \le n \le N-1 \tag{7}$$

where N is the Cepstrum coefficients and $X(k)$ is the speech signal.

Cepstrum coefficients have the characteristics of a discrete pronunciation voice box model and simulating signal to precisely calculate the vocal parameters. The obtained effective speech parameters can be applied in the speech recognition.

4. Neural network

An artificial neural network is one kind of simulation of human biological-type systems. The neural network (NN) is an interconnected group of artificial neurons. The advantages of NN are high-speed computing, large memory, high learning ability and high error-tolerance. The types of NN's learning method can be divided into supervised and unsupervised learning network. Supervised neural-networks are applied in the classification, prediction and recognition. The unsupervised neural-networks are mostly applied in the clustering (Zurada, 1992).

4.1. Back propagation neural network

Back propagation neural network (BPN), invented in 1957 is very typically implemented. BPN is used commonly in everyday life (Rumelhart, Hinton, & Williams, 1986). BPN consists of input layer X_i, hidden layer H_h and output layer Y_j. The input layer represents the input variable. The hidden layer represents how the input layer influences with output layer. The output layer represents the output variable. Every layer is connected by weighted values, but neurons in the same layer are not connected. Every layer is connected with different weights. Fig. 3 shows the framework of the back propagation neural network.

The back propagation neural network uses the steepest gradient descent method and minimizes the error function. BPN is an unsupervised learning network which is suitable for prediction, recognition and so on. Unsupervised learning must be given network training data which includes speech features X_i and its features, values represented by Y_j. Through the training, the network can calculate the weight. The testing data must be given first at the testing procedure, which includes the testing speech features X_i. The calculation process must add the previous weighting values and then it produces the prediction outcome T_k of speech recognition.

Through the training process, the Network calculates the new weighting value. The Network randomly initializes weight W between every layer and the input training data is calculated. Eqs. (8) and (9) calculate the hidden layer vector H_h.

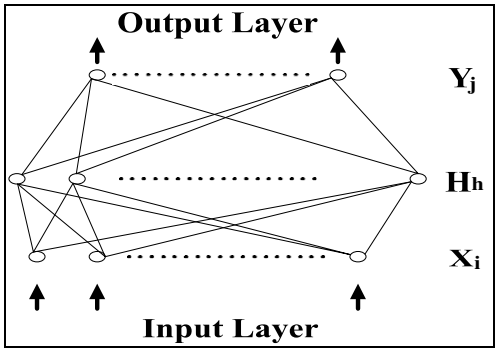

Fig. 3. Back propagation neural network structure.

$$Z_h = \sum_h W_{ih} * X_i - \theta_h \tag{8}$$

$$H_h = \frac{1}{1 + e^{-Z_h}} \tag{9}$$

where Eq. (9) calculates H_h and then insert Eq. (9) into Eq. (10) to get delta δ

$$\delta_h = H_h(1 - H_h) \sum_j W_{hj}\delta_j \tag{10}$$

where Eq. (10) gets delta δ_h and then insert Eq. (10) into Eq. (11) to get the hidden layer weighting delta ΔW_{ih}. The ΔW_{ih} adjusts the network weighting value between the input layer to the hidden layer.

$$\Delta W_{ih} = \eta \delta_h X_i \tag{11}$$

where Eq. (11) gets ΔW_{ih} and then insert Eq. (11) into Eq. (12) to get the new weight W_{ih}.

$$W_{ih} = W_{ih} + \Delta W_{ih} \tag{12}$$

From Eqs. (8) to (12) is the training cycle. In Eq. (10), δ_j and Y_j are the vectors to calculate the output layer vectors and the differences between the hidden layer and output layer. T_j is the target vector. The updated weighting values are the same from input layer to hidden layer. But the equation to calculate the delta Eq. (13) is different.

$$\delta_j = Y_j(1 - Y_j)(T_j - Y_j) \tag{13}$$

4.2. Self-organizing map

Self-organizing map (SOM) is an unsupervised learning network. The input is the series of values. The fundamental principle of SOM is to learn the clustering rules and apply them to the testing samples from the training samples. SOM consists of input layer X_i and output layer Y_j (Kohonen T, 1990). Fig. 4 shows the structure of self-organizing map.

The algorithm of SOM applies the Euclidean distance to calculate the output unit and network j topology and distance from the center. Eq. (14) D_j is to calculate the Euclidean distance of (X_j, Y_j) to C on the coordinate diagram (Linske, 1988).

$$D_j = \sqrt{(X_j - C_x)^2 - (Y_j - C_y)^2} \tag{14}$$

where (X_j, Y_j) are output j topology coordinates and C is the center of the topology coordinates.

4.3. Self-organizing map coordinate diagram

The Mandarin phonetic signal distribution on the radar map is using the cluster feature of the self-organizing map (SOM) neural

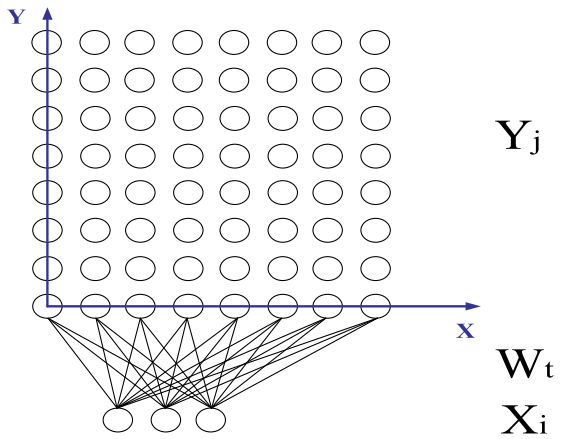

Fig. 4. Self-organizing maps structure.

network. SOM calculates the Euclidean distance of all of the Mandarin phonetic signals between each other and signals with similar characteristics move closer together. The similar Mandarin phonetic symbols cluster on the radar map. Each Mandarin phonetic symbol selects 20 features in the self-organizing map and trains 1000 times. The result shows on the corresponding position on the radar map. In Eq. (14), D_j is to calculate the Euclidean distance of N to M on the coordinate diagram.

4.4. Training

When the network begins to learn, the first step must set the network parameters, the input vector X, the hidden number H the output vector Y, the learning cycle, and the learning rate η. The Network randomly produces the weighting values including the input layer to the hidden layer Wxh_{ih}, the hidden layer to output layer Why_{hj}, the hidden layer's bias θh_h and the output layer's bias θy_j.

The values of Eq. (15) are inserted into Eq. (16) to get the hidden layer vector H.

$$\mathrm{net}_h = \sum_i Wxh_{ih}X_i - \theta h_h \tag{15}$$

$$H_h = f(\mathrm{net}_h) = \frac{1}{1+\mathrm{e}^{-\mathrm{net}_h}} \tag{16}$$

the values of Eq. (17) are inserted into Eq. (18) to get the output layer vector Y.

$$\mathrm{net}_j = \sum_h Why_{hj}H_h - \theta y_j \tag{17}$$

$$Y_j = f(\mathrm{net}_j) = \frac{1}{1+\mathrm{e}^{-net_j}} \tag{18}$$

Eqs. (16) and (18) calculate the H and Y, and individually are inserted into Eqs. (19) and (20) to get the value of delta δ.

$$\delta_j = Y_j(1-Y_j)(T_j - Y_j) \tag{19}$$

$$\delta_h = H_h(1-H_h)\sum_j Why_{hj}\delta_j \tag{20}$$

Eq. (19) gets delta δ_j and is inserted into Eqs. (21) and (22), and separately gets the output layer weight delta ΔWhy_{hj} and bias delta $\Delta\theta y_j$.

$$\Delta Why_{hj} = \eta\delta_h H_h \tag{21}$$

$$\Delta\theta y_j = -\eta\delta_j \tag{22}$$

Eq. (20) gets delta δ_h and is inserted into Eqs. (23) and (24), and individually gets the hidden layer weight delta ΔWxh_{ih} and bias delta $\Delta\theta h_h$

$$\Delta Wxh_{ih} = \eta\delta_h X_i \tag{23}$$

$$\Delta\theta h_h = -\eta\delta_h \tag{24}$$

Eqs. (21) and (22) get ΔWhy_{hj} and $\Delta\theta y_j$ and are inserted into Eqs. (25) and (26) to get the new output layer weighting values Why_{hj} and new output layer bias θy_j.

$$Why_{hj} = Why_{hj} + \Delta Why_{hj} \tag{25}$$

$$\theta y_j = \theta y_j + \Delta\theta y_j \tag{26}$$

Eqs. (23) and (24) calculate the value of ΔWhy_{hj} and $\Delta\theta y_j$. Eqs. (23) and (24) are inserted into Eqs. (27) and (28) to get the new hidden layer weighting values Wxh_{ih} and the new hidden layer bias θh_h.

$$Wxh_{ih} = Wxh_{ih} + \Delta Wxh_{ih} \tag{27}$$

$$\theta h_h = \theta h_h + \Delta\theta h_h \tag{28}$$

Eqs. (15)–(28) are the formulas of the training cycle and hidden layer. The learning rate η affects the network convergence speed.

4.5. Testing

When BPN is testing, all Cepstrum coefficient data is delivered into the network and starts the iteration by the training data of Why_{hj}, θh_h, θy_j, and Wxh_{ih}. The neural network is based on the connected weight and bias to adjust the construction from testing data to get the target vector T.

Eq. (29) is inserted into Eq. (30) to get the hidden layer vector H.

$$net_h = \sum_i Wxh_{ih}X_i - \theta h_h \tag{29}$$

$$H_h = f(\mathrm{net}_h) = \frac{1}{1+\mathrm{e}^{-\mathrm{net}_h}} \tag{30}$$

Eq. (31) is inserted into Eq. (32) to get the input layer X to target vector T

$$\mathrm{net}_k = \sum_h Wht_{hk}H_h - \theta t_k \tag{31}$$

$$T_k = f(\mathrm{net}_k) = \frac{1}{1+\mathrm{e}^{-\mathrm{net}_k}} \tag{32}$$

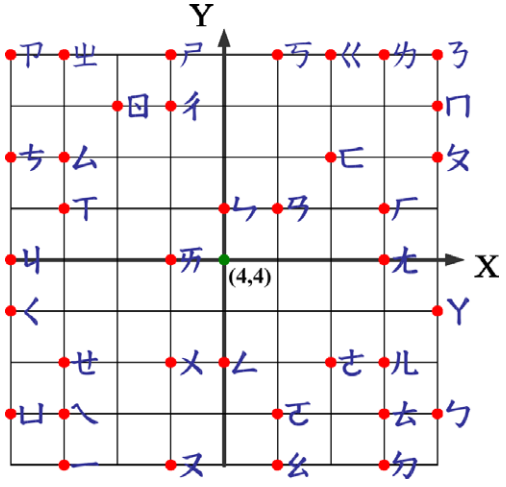

Fig. 5. Chinese phonetic symbol distribution on 2-dmensional coordinate graph.

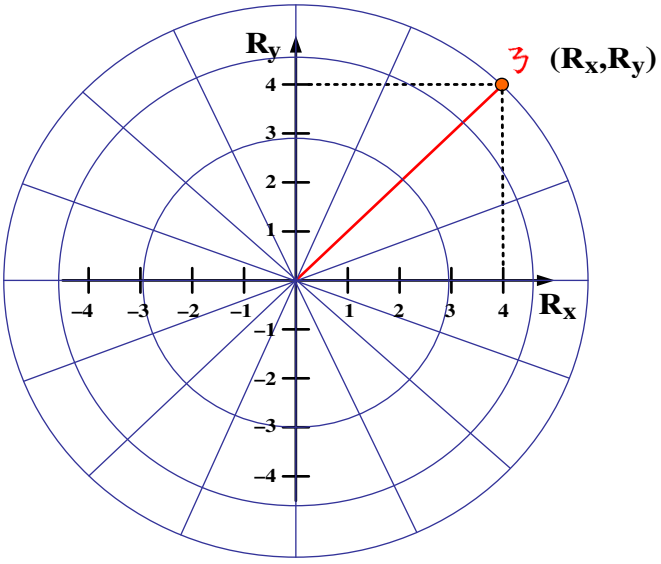

Fig. 6. 2-Dimensional coordinates converting to the radar map.

5. Radar charts

After the SOM training of the speech features, the network distributes every Mandarin phonetic symbol's signal on a 2-dimensional coordinate. The topology coordinate is an 8×8 2-dimensional array. Then, set the origin at (4,4) to get Mandarin phonetic symbols with radius 4 as shown in Fig. 5. The distance between each Mandarin phonetic symbol represents their similarity. The greater the similarity of the pronunciations, the closer the coordinates on the radar chart will be. This is the cluster characteristic of SOM.

The obtained 2-dimensional coordinates need to be converted into radar graph. The radius unit length of radar maps is set to 150. In Fig. 6, Mandarin phonetic symbol ㄋ coordinates are (R_x, R_y). After SOM training gets the coordinates (X, Y), then the (X, Y) are separately inserted into Eqs. (33) and (34) to get the radar map coordinates (R_x, R_y).

$$Rx = X^{*}26.52 \tag{33}$$

$$Ry = Y^{*}26.52 \tag{34}$$

Radius unit length of the radar maps is 150 with the equation of isosceles right triangles. An isosceles right triangle with two-equal sides, and their corresponding angles are 45°. The triangle length ratio is $1 : 1 : \sqrt{2}$ to get the radius unit length of 2-dimensional coordinates with length 106 units. The coordinate axis of radar charts of each scale is 26.52 units length which is one-forth of the axis length.

The results of each radar map are calculated from SOM training speech features. Each Mandarin symbol alphabet extracts 20 data features as training data. Through 1000-iteration training cycles, 2-dimensional coordinates are obtained.

6. Experiment results

This system is using BPN as the framework of speech recognition (Lu & Wang, 2006). Choosing the practice button of a phoneme signal on the operating interface, the computer screen shows the right position of the phoneme signal. Pressing the REC button starts the user's recording by microphone and the data is shown in the speech waveform window. When the speech waveform window shows the user's extracted speech, pressing the EXTRACT button converts the speech into its features. After finishing the last step, pressing the RECOGNITION button applies the BPN recognition test's speech. In final, pressing the SCORE button shows the recognition results on radar chart window, compare to the database.

The tester using the microphone emits the appropriate phoneme signal. The red line on the radar graph is the mark of the corrected pronunciation. The green line is the user's mark of the testing pronunciation. When the pronunciation is correct, the two lines overlap. If the pronunciation is not correct, the two lines are separated. The distance between the two lines is the reference to adjust the pronunciation. The system calculates the corrected pronunciation and similarity of the test and then the score is displayed on the screen, shown in Fig. 7.

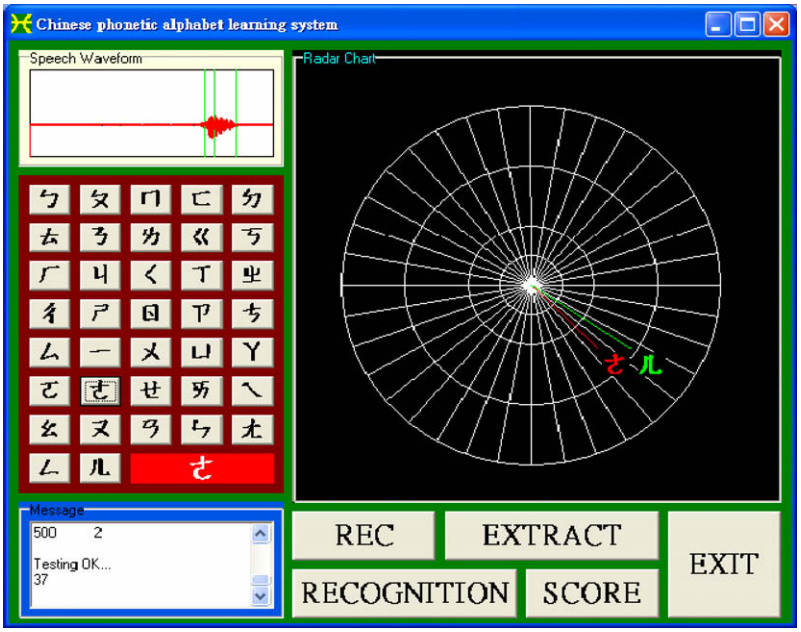

Fig. 7. The implementation of the graphic-displaying speech learning system.

Table 1
Mandarin phonetic alphabet recognition rate

Accuracy rate (%)	Chinese phonetic alphabet
100–90	ㄇㄈㄑㄒㄖㄙㄨㄚㄛㄝㄞㄠㄢㄤㄦ
90–80	ㄐㄧㄩㄣ
80–60	ㄋㄗㄎㄜㄟㄡ
<60	ㄅㄆㄉㄊㄌㄍㄎㄓㄔㄕㄗ
Average rate	72%

Twenty features from each phoneme were extracted for training in BPN. Through the adjustment of the experimental process, each phoneme trains 1000 times. After the adjusted weighting value, the recognition rate is appropriated. The recognition accuracy rates are shown in Table 1. The accuracy rate over 80% is good enough for hearing impaired to learn the phoneme pronunciation. All of the remaining phonemes under 80% recognition rates should not be used as the learning sets. Those low recognition rate phonemes are caused by the similarity in features which need more study to resolve.

7. Conclusion

The RDS improves the traditional language training method using radar maps to increase the hearing impaired's learning performance. When the hearing impaired start to learn languages, they can not judge whether their pronunciation is correct or not and then indirectly cause learning obstacles and decrease their learning intention. In general, the hearing impaired use assistive devices and special teaching methods to increase their learning efficiency. Auditory training is the most important issue in the hearing impaired education. The hearing impaired has residual hearing ability. If the hearing-impaired students were trained by lip language, they lose their original hearing ability to recognize the speech ability. They lose the chance to use speech as a communicating tool with other people (Stewart & Kluwin, 2000).

From the standpoint of an assistive device for the hearing impaired, how to develop a suitable graphic-displaying learning system for the pre-school children is a very critical issue for children's speech learning. Students can use the radar map on the system to modify their pronunciation and also to increase their learning performance. This system can help hearing-impaired people build up their speech learning foundation and reduce their language learning problems. However, owing to the characteristics of sound variability, speech is easily to be affected by external noise. This is why it is hard to get the optimal recognition rate. How to increase the recognition rate is a major technical problem. For advanced study, the recognition rate is one of the major issues to overcome in pronunciation, spelling, vocabulary, phrase and grammar to build up a comprehensive language learning system.

Acknowledgement

This work was supported by the National Science Council of Taiwan, ROC under Grant No. NSC-96-2416-H-167-006 and No. NSC 96-2221-E-167-026-MY3.

References

Johnson, James E., Christie, James F., & YawKey, Thomas D. (1998). *Play and early childhood development* (2nd ed.). Allyn and Bacon/Logman.
Kohonen, T. (1990). The self-organizing maps. *Proceedings of the IEEE, 78*(9), 1464–1480.
Linske, R. (1988). Self-organization in a perceptual network. *IEEE Computer Magazine, 21*, 105–117.
Lu, C. T., & Wang, H. C. (2006). C/V segmentation on Mandarin spontaneously spoken speech signals using SNR improvement and energy variation. *IEICE Transaction on Information and Systems, E89-D*, 363–366.
Newport, E. L. (1991). Contrasting conceptions of the critical period for language. In S. Carey & R. Gelman (Eds.), *The epigenesist of mind* (pp. 111–130). Hillsdale, NJ: Lawrence Erlbau Associates.
Patil, Pramod B. (1998). Multilayered network for LPC based speech recognition. *IEEE Transactions on Consumer Electronics, 44*(2), 435–438.
Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. New Jersey: Prentice Hall.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536.
Stewart, David A., & Kluwin, Thomas N. (2000). *Teaching deaf and heard of hearing students: Content, strategies, and curriculum* (1st ed.). Allyn and Bacon/Logman.
Zurada, J. M. (1992). *Introduction to artificial neural systems*. USA: West Publishing Company.