International Journal of Fuzzy Logic and Intelligent Systems, vol. 4, no. 2, September 2004. pp. 149-154

A Study on Design and Implementation of Speech Recognition System Using ART2 Algorithm

*Joeng Hoon Kim, *Dong Han Kim, *Won Il Jang, *Sang Bae Lee

* Department of Electronic Communication Eng, Korea Maritime University

Abstract

In this research, we selected the speech recognition to implement the electric wheelchair system as a method to control it by only using the speech and used DTW (Dynamic Time Warping), which is speaker-dependent and has a relatively high recognition rate among the speech recognitions. However, it has to have small memory and fast process speed performance under consideration of real-time. Thus, we introduced VQ (Vector Quantization) which is widely used as a compression algorithm of speaker-independent recognition, to secure fast recognition and small memory. However, we found that the recognition rate decreased after using VQ. To improve the recognition rate, we applied ART2 (Adaptive Reason Theory 2) algorithm as a post-process algorithm to obtain about 5% recognition rate improvement. To utilize ART2, we have to apply an error range. In case that the subtraction of the first distance from the second distance for each distance obtained to apply DTW is 20 or more, the error range is applied. Likewise, ART2 was applied and we could obtain fast process and high recognition rate. Moreover, since this system is a moving object, the system should be implemented as an embedded one. Thus, we selected TMS320C32 chip, which can process significantly many calculations relatively fast, to implement the embedded system. Considering that the memory is speech, we used 128kbyte-RAM and 64kbyte ROM to save large amount of data. In case of speech input, we used 16-bit stereo audio codec, securing relatively accurate data through high resolution capacity.

Key words : ART2, DSP(TMS320C32) , DTW, Speech Recognition

1. Introduction

'Human communicates with machine.' That is a form of HCI(Human Computer Interface) which has become indirectly influential along with the development of industry. Here, various biometric recognitions (face recognition, fingerprint recognition, iris recognition etc.) have been used, and speech recognition is one of them. The speech recognition technology means that a machine accepts human language and take proper actions for the language. This technology has been used in the whole areas of various industry, especially, information industry, digital communication, electric home appliances, multi-media etc. In this research, this technology is applied to a moving robot (electric wheelchair system) to provide more convenience to the physically handicapped whose hands or feet cannot move freely.

As for speech recognition method, there are DTW[1] where pattern matching is performed by distance; HMM(Hidden Markov Model)[2] where speech is recognized statistically; and NN(Neural Network) where the structure of brain is modeled. Depending on the recognition method, it can be divided into speaker-dependent recognition where it is applied to only one person and speaker-independent one where it is applied to multiple persons.

Considering that the electric wheelchair should be used by the physically handicapped, we designed the system as a speaker-independent type in this research and used DTW of which recognition rate is relatively suitable as the recognition algorithm for the first half part. To compensate the wrong recognition rate, the ART2 algorithm that was used as if a sorter since it was designed to maintain the patterns previously learned while not losing the flexibility necessary to learn new patterns was applied as the recognition algorithm for the second half part.

This paper consists of six chapters: in Chapter 2, we explained MFCC, which is the method to detect only features of large amount of speech data, and discussed how the speech sound section is detected and briefly mentioned about the theory of VQ for data compression.; in Chapter 3, we described the design of recognition algorithms including DTW and ART2 In Chapter 4, we explained the embedded-type speech recognition board using the algorithms described in previous chapters and how it was processed in aspect of software. In Chapter 5, we mentioned the actual experiments and corresponding results; in Chapter 6, we concluded this research

2. Feature Extraction

2.1 Feature Extraction & Vector Quantization

Figure 1 displays the MFCC process. Once the pre-emphasis

Manuscript received Aug. 20, 2004; revised Sep. 6, 2004

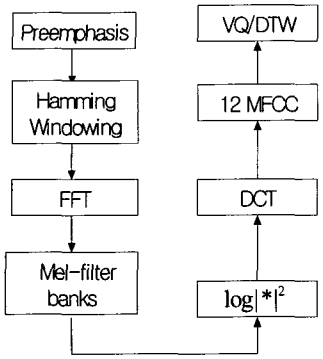

Figure 1. MFCC process procedure

is implemented to the speech signal, the Hamming window is applied. Then, FFT (Fast Fourier Transform) is used to transform it into the frequency domain. If the transformed frequency-domain values pass through the pre-setup filter bank, and DCT (Discrete Cosine Transform), the inverse-transform, is applied, and 12 coefficients are obtained per frame. MFCC is implemented with this filter bank.

2.2 Vector Quantization Algorithm

Feature values from each word were extracted by implementing speech detection and feature extraction procedures and sounding each words ten times. The feature values obtained here are combined into one file.

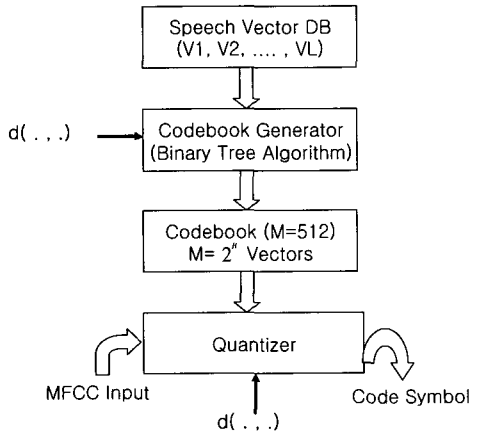

Figure 2. General block diagram of vector quantization

With this data, 512 codebooks were created using the vector codebook generation training algorithm on a PC. Figure 2 shows the general block diagram of the vector quantization procedure. For a new input, 12 feature vectors per frame are extracted and compared with the existing quantization table so as to produce the symbol with the closest match from the quantization table.[3]

3. Recognition Algorithm

The recognition algorithm used in this study, DTW, performs pattern matching between feature values of the word to be recognized the words saved in the reference. When feature values of a newly entered word are extracted, the distance value is obtained by comparing them to the features of the existing words. Among those obtained distances, the least value is determined as the recognition word. In this study, we used vector quantization to compress the feature values by 1/12 so as to decrease the size of the reference memory. Moreover, reduction in data per word, resulted in faster DTW distance calculation and a speed increase of approximately 4 times.

3.1 DTW Recognition Algorithm

DTW obtains similarity between the standard speech signal pattern and the inputted speech signal by using dynamic programming.

Namely, this method compensates for the difference on the time-axis. Let us say that the inputted speech pattern with length M is T=T(1),T(2),... ,T(M); the standard pattern with length N is R=R(1),R(2),R(3)..., R(N); then, the similarity between the two patterns can be described in (6):

$$D = \sum_{N}^{n=1} d(R(n), T(W(n))) \tag{6}$$

3.2 ART2 (Adaptive Reason Theory 2)

ART2 algorithm can learn not only binary input patterns but also analogue one. ART2 divides F1 layer into multiple sub-layers (W, X, U, V, P and Q) and does feed-back and feed-forward processes. ART2 has sub-layers and gain control part as shown in Figure 3. The parameters shown in Figure 3 can be explained as below: [4]

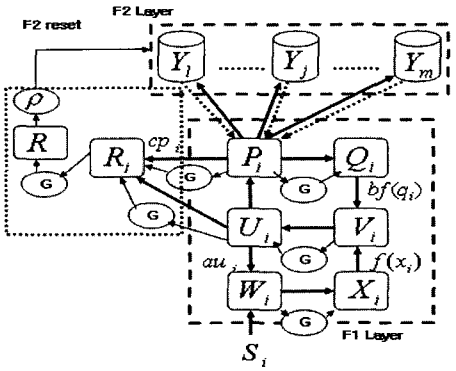

Figure 3. Architecture of ART2

- n : number of input units (F1 layer)
- m : number of cluster units (F2 layer)

- a, b : fixed weights in the F1 layer
- c : fixed weights used in testing for reset
- d : activation of winning F1 unit

e : a small parameter introduced to prevent division by zero when the norm of a vector is zero

ART2 algorithm consists of a total of 13 steps as below:

Step0) Initialize parameters
a, b, θ, c, d, e, α, ρ

Step1) Perform Step2) to Step12).

Step2) For each input vector s, do Step3) ~ Step11)

Step3) Update F1 unit activations & Update F1 unit activations again

Update F1 unit activations	Update F1 unit activations again
$u_i = 0$	$u_i = \frac{s_i}{e + \|v\|}$
$x_i = \frac{s_i}{e + \|s_i\|}$	$x_i = \frac{w_i}{e + \|w_i\|}$
$w_i = s_i$	$w_i = s_i + aw_i$
$q_i = 0$	$q_i = \frac{p}{e + \|p\|}$
$p_i = 0$	$p_i = u_i$
$v_i = f(x_i)$	$v_i = f(x_i) + bf(q_i)$

Step4) Compute signals to F2 units

$$y_j = \sum_{i=1}^{n} p_i b_{ji}$$

Step5) While reset is true, do Step6) ~ Step7)

Step6) Find F1 unit Y_j with largest signal
Define J such that $y_J \geq y_j$ for (j=1,2,….,m)

Step7) Check for reset

Conditions	$u_i = \frac{v_i}{e + \|v\|}$, $p_i = u_i + dz_{ij}$ $r_i = \frac{u_i + cp_i}{\|u\| + c\|p\|}$
If, $\|r\| < \rho - e$ then $y_J = -1$	
◊ Reset is true ; repeat step 5	
If $\|r\| < \rho - e$, then	
$w_i = s_i + au_i$, $x_i = \frac{w_i}{e + \|w_i\|}$, $q_i = \frac{p_i}{e + \|p\|}$,	
$v_i = f(x_i) + bf(q_i)$	

Step8) Do Step9) ~ Step11)

Step9) Update weights for winning unit J

$$t_{Ji} = adu_i + \{1 + ad(d-1)\}t_{Ji}$$
$$b_{iJ} = adu_i + \{1 + ad(d-1)\}b_{iJ}$$

Step10) Update F1 activations

$u_i = \frac{v_i}{e + \|v\|}$	$x_i = \frac{w_i}{e + \|w\|}$
$w_i = s_i + au_i$	$q_i = \frac{p_i}{e + \|p\|}$
$p_i = u_i + dz_{ij}$	$v_i = f(x_i) + bf(q_i)$

Step11) Test stopping condition for weight updates

Step12) Test stopping condition for number of epochs

Figure 4 shows the flowchart of ART2

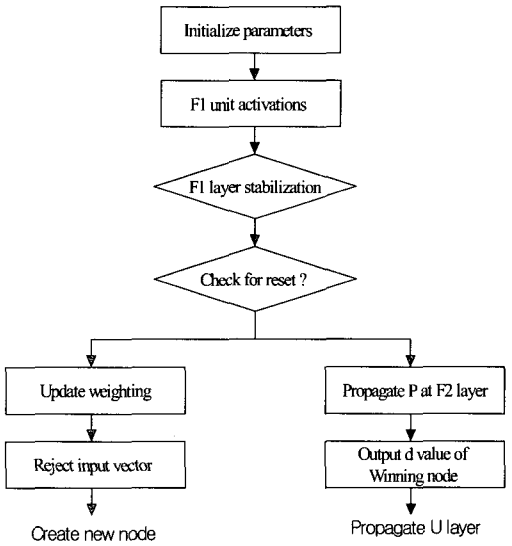

Figure 4. Flowchart of ART2

3.4. Suggested Hybrid Algorithm

This research used the recognition method applying DTW and ART2. Its structure is shown in Figure 5.

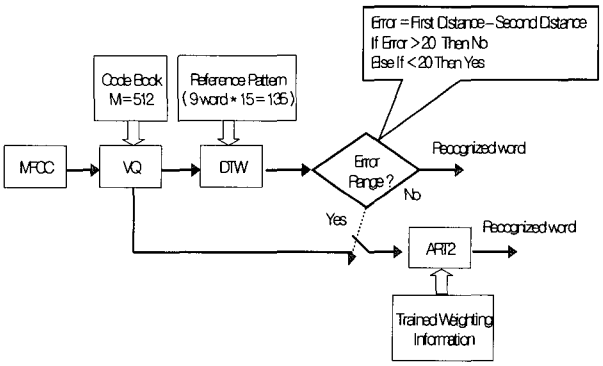

Figure 5. Suggested Hybrid Algorithm Block Diagram

Each distance is calculated through the recognition algorithm of the first half part (DTW). Through each distance calculation, the second distance value and the fist distance value are selected.

Here, we applied the error range. If the difference between the second distance and the first distance is within the range, it is converted to ART2. To have this range first, the precondition

that the first distance and the second distance are not same should be made.

Furthermore, if the difference between these two distances is more than 20, ART2 is applied. The applied data determines a word. If any word which is not in the reference is created, the recognition is stopped.

4. Embedded Speech Recognition Board

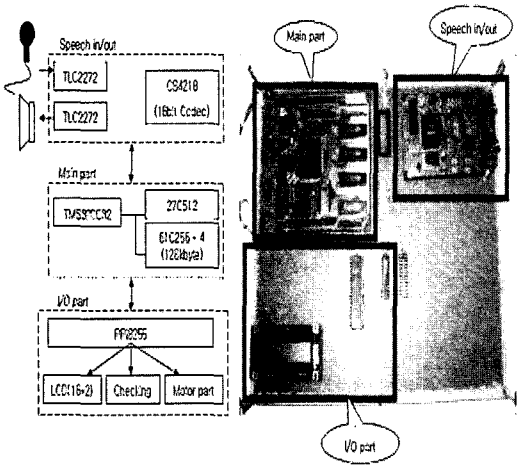

Figure 6. Real Speech Recognition Board

Speech has large amount of data by its own characteristics. In addition, to compress large amount of data and to find feature values, we need the chip that has fast calculation process capacity.

Considering these, we designed the speech recognition board as shown in Figure 6 in this research. [5][6]

To give a general account of this board, the speech is sent to TMS320C32, which is the DSP, in serial at 8kHz, through CS4218 with 16-bit resolution capacity. The delivered speech data is processed by the arithmetic operation process inside the DSP [7].

Considering that speech needs large amount of memory by its own characteristics, we used 128kbyte (61256*4) RAM and 64Kbyte (27C512) ROM. The recognized final values are sent to the LCD and the moving robot through the I/O interface (PPI8255).

5. Experiment and Study

In this chapter, we measured how the speech recognition changes and the recognition rate. In this measurement, fast data process and simulation are unavoidable on real-time, we used the data measured off-line. As for the recognition part, we compared the real-time processes at using DTW with those at using ART2 to see the recognition rate difference.

5.1. Experimental Condition

We collected a total of 135 speeches for the moving direction of the moving robot, and each pattern sequence is shown in Table 1.

Figure 7 shows the result of pre-processed patterns:

Table 1. Speech Pattern Sequence

A-Pe-Uro	1 ~ 15	Dwi-Lo	16 ~ 30
Oen-Jjok	31 ~ 45	O-Reun-Jjok	46 ~ 60
Cheon-Cheon-Hi	61 ~ 75	Ppal-Li	76 ~ 90
Jwa-Hoe-Jeon	91 ~ 105	U-Hoe-Jeon	106 ~ 120
Jeong-Ji	121 ~ 135		

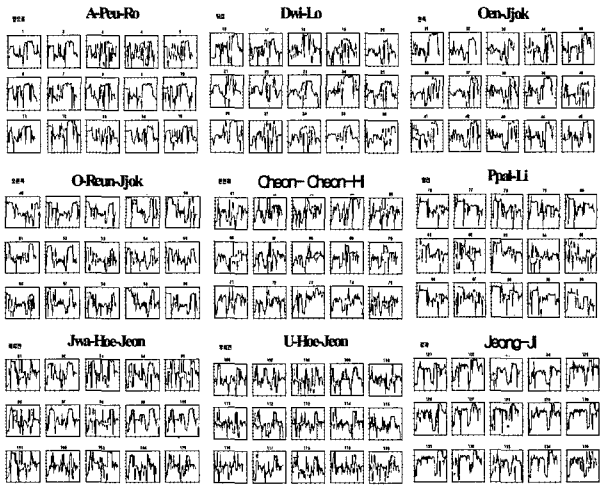

Figure 7. Preprocessing of Input Speech

5.2. ART2 Classification

Through the above patterns, we could obtain 135 (total speech number) x 80 (maximum frame number) input patterns. Table 2 (Appendix) shows the classification of these input patterns with ART2 algorithm.

ART2 parameter initialization is shown as below:

- Constant : a = 10, b = 10, c = 0.1, d = 0.995
- Training Factor : $\theta = 0.129$, $\eta = 0.2$, M = 50
- Error rate : e = 0.01

5.3. Speech Recognition

The recognition experiment was tested in real-time environment. Table 3 (Appendix) shows the result of classification on the DSP board using only ART2 and shows lower classification rate than off-line state. Table 4 (Appendix) shows the final recognition result of a) those using only DTW and b) those in hybrid of DTW+ART2 on the completed DSP recognition board. With the result, we can confirm that the hybrid gives about 5% increase of overall averages against using only DTW. Based on the above experiments, we can confirm that post-process recognition algorithm gives the improvement of recognition rate.

Table 2. Pattern Classification of ART2

Word	Winner Node	Classified Input Pattern	Word	Winner Node	Classified Input Pattern
A-Pe-Uro	1	1, 2, 3, 6, 7	Ppal-Li	19	76,77, 78,79 , 85,86, 88,90
	2	4, 8, 13		20	80, 81, 82,87, 89
	3	5, 9, 10, 15		21	83, 84
	4	11, 12, 14	Jwa-Hoe-Jeon	22	91, 92, 93
Dwi-Lo	5	16, 17, 27		23	94, 95, 96
	6	18, 19, 20, 21, 26, 29		24	97, 98, 99
	7	22, 23, 24, 25, 28		25	100, 101, 104
	8	30		26	102
Oen-Jjok	9	31, 32, 33, 35, 37, 39, 44, 45		27	103, 105
	10	34, 36, 38, 41	U-Hoe-Jeon	28	106, 108
	11	40, 42, 43		29	107, 119, 120
O-Reun-Jjok	12	46, 47, 48, 49, 50, 56, 57		30	109, 110
	13	51, 52, 54		31	111, 114, 118
	14	53, 55		32	112, 117
	15	58, 59, 60		33	113
Cheon-Cheon-Hi	16	61, 62, 68, 69, 71, 75		34	115, 116
	17	63, 64, 66, 67, 72	Jeong-Ji	35	121,122, 123,124, 125,126
	18	65, 70, 73, 74		36	127, 131, 132, 134
				37	128, 129, 130, 133, 135

Table 3. Recognition of ART2 (x) : Wrong Recognition

Word	1	2	3	4	5
A-Pe-Uro	(4)	(1)	(3)	(3)	**(x)**
Dwi-Lo	(6)	(5)	(7)	(5)	(3)
Oen-Jjok	(11)	(9)	(9)	(10)	(9)
O-ReunJjok	**(x)**	**(x)**	(13)	(14)	(13)
Cheon-Cheon-Hi	(16)	(17)	(16)	(18)	(18)
Ppal-Li	(19)	(19)	(20)	(21)	(19)
Jwa-Hoe-Jeon	(25)	(22)	(26)	(24)	**(x)**
U-Hoe-Jeon	(33)	(34)	**(x)**	(31)	**(x)**
Jeong-Ji	(35)	(36)	(37)	(37)	(35)

Table 4. Result of Speech Recognition Test

Word	Count	a) DTW		b) DTW+ART2	
		Recognition Number	Recognition Rate	Recognition Number	Recognition Rate
A-Pe-Uro	50	46	92%	49	98%
Dwi-Lo	50	45	90%	49	98%
Oen-Jjok	50	46	92%	48	96%
O-Reun-Jjok	50	47	94%	49	98%
Cheon-Cheon-Hi	50	45	90%	48	96%
Ppal-Li	50	45	90%	48	96%
Jwa-Hoe-Jeon	50	44	88%	46	92%
U-Hoe-Jeon	50	44	88%	47	94%
Jeong-Ji	50	47	94%	49	98%

6. Conclusion

In this research, we used speaker-dependent recognition among speech recognition methods while considering that the plant is an electric wheelchair and implemented the speech recognition program on the embedded system using DTW, which gives a relatively high recognition rate. In addition, this research introduced the concept of real-time, and we intended to sense only speeches out of speeches inputted continuously as time passes, to process them fast and to recognize them so that they could be applied to the operation of electric wheelchair. Accordingly, we utilized the concept of vector quantization and compressed the speech vectors up to 1/12 to reduce the time for recognition process significantly. However, the recognition rate was somewhat deteriorated in the event of using the vector quantization, so we applied the hybrid of DTW+ATR2 in this research to compensate it, and as a result, we improved the recognition rate. Consequently, we obtained about 5% recognition improvement.

For future research directions, we need to study more about the application of continuous speech and speaker-independent speech recognition to the embedded system and the implementation of the above speech recognition system on other application fields. Also, we will need to develop the recognition system strong against noisy environment.

Reference

[1] L.R.Rabiner, B.H.Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993

[2] Lawrence Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition", Proc. IEEE, Vol.77, No. 2, February 1989.

[3] Lawrence Rabiner, "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent Isolated Word Recognition", Bell System Technical Joural, Vol. 62, No.4, April 1983.

[4] Carpenter, G.A., Grossberg S., "ART2: Self-organization of stable category recognition codes for analog input patterns", Applied Optics, Vol. 26, No. 23, pp. 4919-4930, 1987.

[5] Joeng Hoon Kim, "A study on Design and Implementation of Embeded System for Speech Recognition Process", Journal of Fuzzy Logic and Intelligent System, Vol. 14, No. 2, April 2004.

[6] Joeng Hoon Kim, "A study on Deveolpment of Embeded System for speech Recognition using Multi-layer Recurrent Neural Prediction Models & HMM", Journal of Fuzzy Logic and Intelligent System, Vol. 14, No. 3, June 2004.

[7] Ji Hong Lee & Seoil DSP Technology Research, "Applications of DSP chip", Seoil DSP Co., Ltd.

Joeng Hoon Kim
received the B.Sc degree of Information Communication from Dong- Myong University of Information Techno- logy in 2001, and M.Sc degree from Korea Maritime University of Electronic Communication in 2003. Presently, he is currently pursuing his Ph.D degree in Korea Maritime University, and working as an affiliate professor in Dong-Myong University of Information Technology. His research interests include neural Network, fuzzy, speech recognition and visual recognition.

Dong-han Kim
received the B.Sc degree of Electric Engineering from Chin-ju National University in 2001, and M.Sc degree from Korea Maritime University of Electronic Communication in 2003. Presently, he is currently pursuing his Ph.D degree in Korea Maritime University. His research interests include Embedded System., Digital Image Processing, Aritificial Intelligence.

Won IL Jang
received the B.Sc degree of Information Communication from Dong-Myong University of Information Technology in 2002. He is currently pursuing his M.Sc degree in Korea Maritime University. His research interests include Visual Recognition, Image Processing, Embedded System

Sang Bae Lee
acquired B.Sc degree from Dong-A University in 1981, and M.Sc degree and Ph.D degree from Korea University in 1983 and 1989 respectively. He was researcher at Canada Saskatchewan University (ISRL) from 1993 to 1994. Presently, he is a full professor of Korea Maritime University. His research interests include fuzzy-control system, fuzzy-neural control and biometric