MMSDS: Ubiquitous Computing and WWW-based Multi-Modal Sentential Dialog System

Jung-Hyun Kim and Kwang-Seok Hong

School of Information and Communication Engineering, Sungkyunkwan University, 300, Chunchun-dong, Jangan-gu, Suwon, KyungKi-do, 440-746, Korea
kjh0328@skku.edu, kshong@skku.ac.kr
http://hci.skku.ac.kr

Abstract. In this study, we suggest and implement Multi-Modal Sentential Dialog System (MMSDS) integrating 2 sensory channels with speech and haptic information based on ubiquitous computing and WWW for clear communication. The importance and necessity of MMSDS for HCI as following: 1) it can allow more interactive and natural communication functions between the hearing-impaired and hearing person without special learning and education, 2) according as it recognizes a sentential Korean Standard Sign Language (KSSL) which is represented with speech and haptics and then translates recognition results into a synthetic speech and visual illustration in real-time, it may provide a wider range of personalized and differentiated information more effectively to them, and 3) above all things, a user need not be constrained by the limitations of a particular interaction mode at any given moment because it can guarantee mobility of WPS (Wearable Personal Station for the post PC) with a built-in sentential sign language recognizer. In experiment results, while an average recognition rate of uni-modal recognizer using KSSL only is 93.1% and speech only is 95.5%, advanced MMSDS deduced an average recognition rate of 96.1% for 32 sentential KSSL recognition models.

1 Introduction

The multi-modal interfaces integrating various sensory channels such as speech, vision and haptic can increase the bandwidth and application of human-computer interaction and it may improve the interactive properties and functions of the system. The functionality of a multi-modal interface can be fairly extensive, and it may find their applications in a number of fields. To name only a few examples, they proved to be a viable aid for visually impaired users, an alternative to WIMP (Windows, Icon, Menu, and Pointer based interfaces) interfaces in mobile computing, an entertaining extension of computer games [1]. In recent years, there have been a lot of innovations and evolutions in the areas of human-computer interaction, multi-modal interface and speech and sign language recognition as a part of natural language understanding. As a case study on a multi-modal interaction and sign language, we can find and introduce the following examples. Siska Fitrianie et al. described a computer model for a multi-modal communication system based on the famous Eliza question-answering

system [2], Wu jiangqin et al. implemented 26 word-level sign language recognition system using neural network and HMM hybrid method [3] and a multi-agent modal language for concurrency with non-communicating agents is introduced by Stefano Borgo [4]. Also, Silvia Berti and Fabio Paternò developed multi-modal interfaces for different platforms starting with logical user interface descriptions in multi-device Environments [5] and Mark Barnard et.al introduced on multi-modal audio-visual event recognition for analysis of football games [6]. However generally, according as most of traditional studies like this keep the accent on an implementation of uni-modal recognition and translation system that recognizes and represents one of various natural components based on wire communications net and a vision technology, they have not only several restrictions such as limitation of representation, conditionality on the space and limitation of the motion, but also some problems such as uncertainty of measurement and a necessity of complex computational algorithms. Nevertheless, users can issue requests using speech, gesture and so on, or dynamic fusions of the two based on ubiquitous computing.

Consequently, we suggest and implement advanced MMSDS integrating speech and KSSL gestures based on the VXML for WWW-based speech recognition (and synthesis) and ubiquitous computing-oriented WPS with a built-in KSSL recognizer. The advantages of our study are as following: 1) it improves efficiency of KSSL input module according to the wireless communication net and user need not be constrained by the limitations of a particular interaction mode at any given moment, 2) our study that allows dialog and communication functions between the hearing-impaired and hearing person is very important and essential, and 3) it recognizes and represents continuous KSSL with flexibility in real time and can provide a wider range of personalized and differentiated information more effectively.

In section 2, we describe an implementation of WPS-based embedded KSSL recognizer. Also, a synopsis of WWW-based VXML module and integration architecture of speech and KSSL for MMSDS are given in section 3. In section 4, we evaluate a performance of the system with recognition experiment results for sentential KSSL recognition models. Finally, this study is summarized in section 5 together with an outline of challenges and future directions.

2 Wearable Personal Station-Based Embedded KSSL Recognizer

2.1 Regulation and Components of the KSSL

Sign language is a language which uses manual communication instead of sound to convey meaning - simultaneously combining hand shapes, orientation and movement of the hands, arms or body, and facial expressions to fluidly express a speaker's thoughts [7]. However, not only a absolute natural learning and interpretation of such KSSL very difficult and takes a long time to represent and translate it fluently in hearing person, but also understanding and learning of spoken language in the hearing-impaired is impossible and uncertain. In other words, because the KSSL is very complicated and is consisted of considerable numerous gestures, motions and so on, it

are impossible that recognize all dialog components which are represented by the hearing-impaired. Therefore, we prescribe that this paper is a fundamental study for perfect dialog and communication between the hearing-impaired and hearing person, and selected 25 basic KSSL gestures connected with a travel information scenario according to the "Korean Standard Sign Language Tutor (KSSLT)[8]". And necessary 23 hand gestures for travel information - KSSL gestures are classified as hand's shapes, pitch and roll degree. Consequently, we constructed 32 sentential KSSL recognition models according to associability and presentation of hand gestures and basic KSSL.

2.2 Improved KSSL Input Module Using Wireless Haptic Devices

For an improved KSSL input module, we adopted blue-tooth module for wireless sensor network, 5DT company's wireless data (sensor) gloves and fastrak® which are one of popular input devices in the haptic application field. Wireless data gloves are basic gesture recognition equipment that can acquires and capture various haptic information (e.g. hand or finger's stooping degree, direction) using fiber-optic flex sensor. The structural motion information of each finger in data glove are captured by f1=thumb, f2=index, f3=middle, f4=ring and f5=little in regular sequence. Each flexure value has a decimal range of 0 to 255, with a low value indicating an inflexed finger, and a high value indicating a flexed finger. Also, the fastrak® is electromagnetic motion tracking system, a 3D digitizer and a quad receiver motion tracker. And it provides dynamic, real-time measurements of six degrees of freedom; position (X, Y, and Z Cartesian coordinates) and orientation (azimuth, elevation, and roll) [9], [10]. The architecture and composition of KSSL input module is shown in Fig. 1.

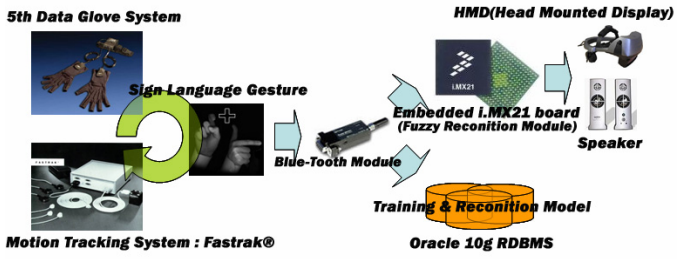
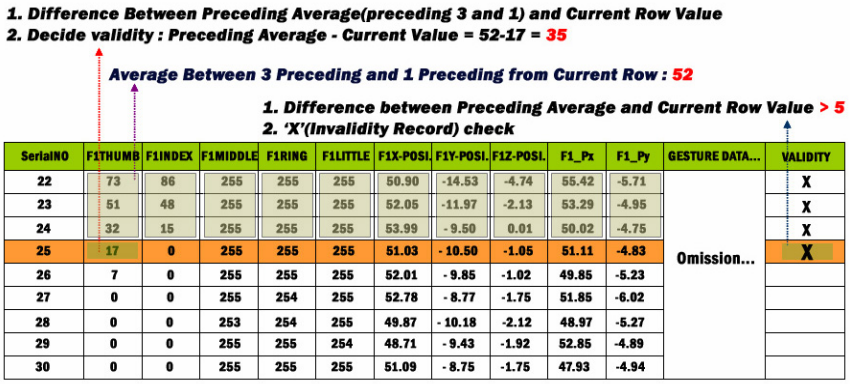


Fig. 1. The architecture and composition of KSSL input module.

2.3 Feature Extraction and Recognition Models using RDBMS

A statistical classification algorithms such as K-means clustering based on attributes into k partitions, QT (Quality Threshold) clustering that is an alternative method of partitioning data, fuzzy c-means clustering algorithm and Self-Organizing Map (SOM) based on a grid of artificial neurons whose weights are adapted to match input vectors in a training set had been applied universally in a traditional pattern recogni-

tion systems with unsupervised training, including machine training, data mining, pattern recognition, image analysis and bioinformatics [11], [12], [13]. However, such classification algorithms have some restrictions and problems such as the necessity of complicated mathematical computation according to multidimensional features, the difficulty of application in a distributed processing system, relativity of computation costs by patterns (data) size, minimization of memory swapping and assignment. Accordingly, for a clustering method for efficient feature extraction and a construction of training / recognition models based on a distributed computing, we suggest and introduce improved RDBMS (Relational Data-Base Management System) clustering module to resolve such a restrictions and problems. The RDBMS is database management system that maintains data records and indices in tables and their relationships may be created and maintained across and among the data and tables. Also, it has the capability to recombine the data items from different files, providing powerful tools for data usage [14]. The RDBMS-based analytic functions for substantial pattern clustering and segmentation are designed to address such problems as "calculate a running total", "find percentages within a group", "top-N queries", "compute a moving average" and many more [15]. A clustering rule to segment valid gesture record set and invalid record set in the RDBMS classification module is shown in Fig. 2.



SerialNO	F1THUMB	F1INDEX	F1MIDDLE	F1RING	F1LITTLE	F1X-POSI.	F1Y-POSI.	F1Z-POSI.	F1_Px	F1_Py	GESTURE DATA...	VALIDITY
22	73	86	255	255	255	50.90	-14.53	-4.74	55.42	-5.71		X
23	51	48	255	255	255	52.05	-11.97	-2.13	53.29	-4.95		X
24	32	15	255	255	255	53.99	- 9.50	0.01	50.02	-4.75		X
25	17	0	255	255	255	51.03	- 10.50	-1.05	51.11	-4.83	Omission...	X
26	7	0	255	255	255	52.01	- 9.85	-1.02	49.85	-5.23		
27	0	0	255	254	255	52.78	- 8.77	-1.75	51.85	-6.02		
28	0	0	253	254	255	49.87	- 10.18	-2.12	48.97	-5.27		
29	0	0	255	255	254	48.71	- 9.43	-1.92	52.85	-4.89		
30	0	0	255	255	255	51.09	- 8.75	-1.75	47.93	-4.94		

Fig. 2. The clustering rules to segment in the RDBMS classification module.

2.4 Fuzzy Max-Min Composition-based KSSL Recognition

Fuzzification and Membership Function. For a design of membership function, many types of curves can be used, but triangular or trapezoidal shaped membership functions are the most common because they are easier to represent in embedded-controllers [18]. So, we applied trapezoidal shaped membership functions for representation of fuzzy numbers-sets, and this shape is originated from the fact that there are several points whose membership degree is maximum. The proposed the fuzzy membership functions are shown in Fig. 3.

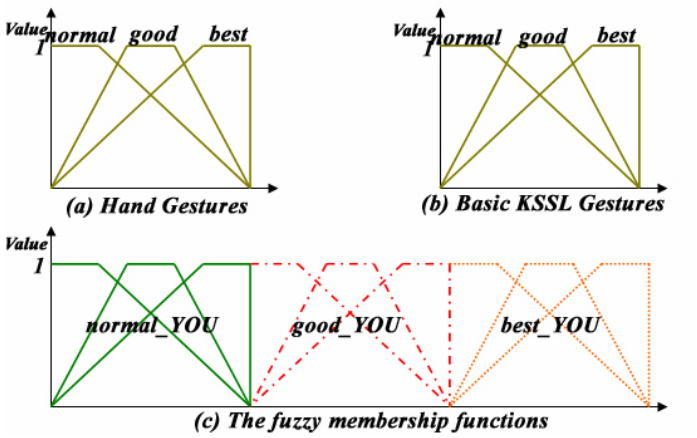


Fig. 3. The fuzzy membership functions for KSSL recognition. Because fuzzy numbers-sets according to KSSL recognition models are very various and so many, we represent membership functions partially: "YOU" in KSSL).

Fuzzy Logic: Max-Min Composition of Fuzzy Relation. In this paper, we utilized the fuzzy max-min composition to extend a crisp relation concept to relation concept with fuzzy proposition and to reason approximate conclusion by composition arithmetic of fuzzy relation. Two fuzzy relations R and S are defined on sets A, B and C (we prescribed the accuracy of hand gestures and basic KSSL gestures, object KSSL recognition models as the sets of events that are happened in KSSL recognition with the sets A, B and C). That is, $R \subseteq A \times B$, $S \subseteq B \times C$. The composition $S \bullet R = SR$ of two relations R and S is expressed by the relation from A to C, and this composition is defined in equation (2) [17].

$$\text{For } (x, y) \in A \times B,\ (y, z) \in B \times C, \qquad (2)$$

$$\mu_{S \bullet R}(x, z) = \max_{y} [\min(\mu_R(x, y), \mu_S(y, z))]$$

$S \bullet R$ from this elaboration is a subset of $A \times C$. That is, $S \bullet R \subseteq A \times C$. If the relations R and S are represented by matrices M_R and M_S, the matrix $M_{S \bullet R}$ corresponding to $S \bullet R$ is obtained from the product of M_R and M_S; $M_{S \bullet R} = M_R \bullet M_S$. That is, we can see the possibility of occurrence of B after A, and by S, that of C after B in Table 1, 2. For example, by matrices M_R, the possibility of "Best" $\in B$ after "Best" $\in A$ is 0.9.

Table 1. The matrices M_R for the relations R between the fuzzy set A and B

R	Accuracy of basic KSSL gestures				
Accuracy of hand gestures	Best	Good	Normal	Bad	Very_bad
Very_bad	0.0	0.1	0.2	0.6	0.9
Bad	0.0	0.2	0.3	0.8	0.6
Normal	0.2	0.3	0.6	0.4	0.3
Good	0.7	0.9	0.5	0.3	0.2
Best	0.9	0.7	0.5	0.2	0.1

Table 2. The matrices M_R for the relations R between the fuzzy set B and C

S	Accuracy of 25 basic KSSL gestures				
Accuracy of basic KSSL gestures	Insignificance	Bad_YOU	Normal_YOU	Good_YOU	Best_YOU
Best	0.1	0.2	0.4	0.6	0.9
Good	0.2	0.3	0.5	0.8	0.7
Normal	0.3	0.4	0.6	0.3	0.2
Bad	0.7	0.8	0.4	0.2	0.1
Very_bad	0.9	0.6	0.3	0.1	0.0

Table 3. The matrix $M_{S \cdot R}$ corresponding to the relations $S \bullet R$

S • R	KSSL recognition model : "YOU"				
Accuracy of basic KSSL gestures	Insignificance	Bad_YOU	Normal_YOU	Good_YOU	Best_YOU
Very_bad	0.9	0.6	0.3	0.2	0.1
Bad	0.6	0.7	0.4	0.2	0.2
Normal	0.3	0.4	0.4	0.3	0.3
Good	0.3	0.3	0.5	0.8	0.7
Best	0.2	0.3	0.4	0.6	0.9

Also, by matrices M_S, the possibility of occurrence of "Good_YOU" after "Best" is 0.6. Also, the matrix $M_{S \cdot R}$ in Table 3 represents max-min composition that reason and analyze the possibility of C when A is occurred and it is also given in Fig. 4.

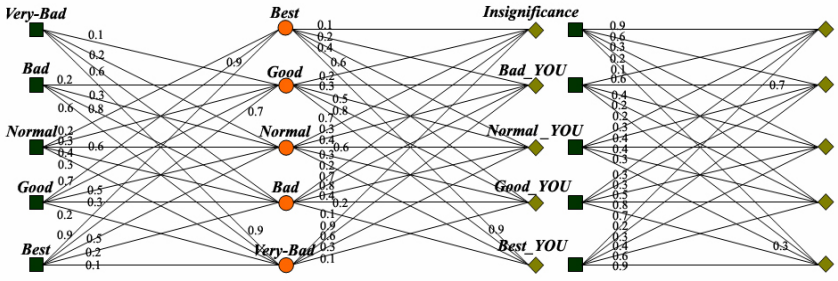


Fig. 4. Composition of fuzzy relation

3 Voice-XML for WWW-based Speech Recognition and Synthesis

3.1 Components and architecture of Voice-XML

VXML is the W3C's standard XML format for specifying interactive voice dialogues between a human and a computer [18]. A document server (e.g. a Web server) processes requests from a client application, the VXML Interpreter, through the VXML interpreter context. The server produces VXML documents in reply, which are processed by the VXML interpreter. The VXML interpreter context may monitor user inputs in parallel with the VXML interpreter. For example, one VXML interpreter context may always listen for a special escape phrase that takes the user to a high-level personal assistant, and another may listen for escape phrases that alter user preferences like volume or text-to-speech characteristics.

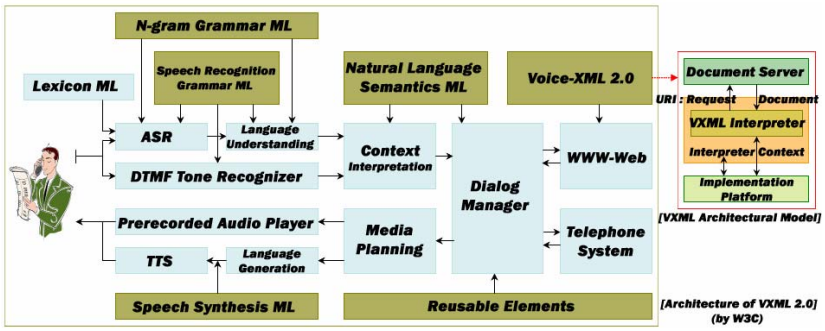


Fig. 5. The components of architectural model and architecture of W3C's VXML 2.0

The implementation platform is controlled by the VXML interpreter context and by the VXML interpreter. The components of architectural model and architecture of VXML 2.0 by W3C are shown in Fig. 5.

3.2 Integration of Speech and KSSL for MMSDS

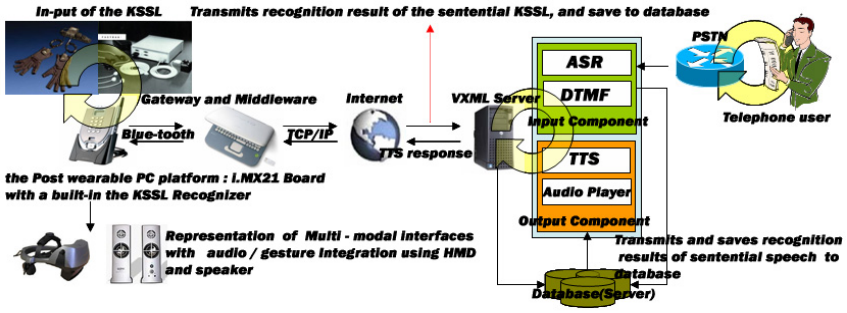


Fig. 6. The components and architecture of MMSDS using speech and sign language (gestures)

The user connects to VXML server through internet and PSTN using WPS based on wireless networks (middleware) and telephone terminal, and input prescribed speech and KSSL. The user's sentential speech data which is inputted into telephone terminal transmits to ASR-engine (we used 'HUVOIS-TTS' that is speech recognition and synthesis S/W for visually-impaired people and developed by KT Corp. in Korea), and saves sentential ASR results to MMI database. Also, user's KSSL data which is inputted into embedded WPS is recognized by sentential KSSL recognizer, transmits and saves sentential recognition results to VXML server using TCP/IP protocol based on middleware and wireless sensor networks. Sentential ASR and KSSL recognition results execute comparison arithmetic by internal SQL logic, and transmit arithmetic results to MMSDS. These arithmetic results are definite intention that user presents. Finally, user’s intention is provided to user through speech (TTS) and visualization. The KSSL recognition processes of MMSDS synchronize with the speech recognition

and synthesis using VXML. The suggested a scenario and architecture of MMSDS using speech and KSSL are shown in Fig. 6. And the flowchart of the MMSDS integrating VXML for WWW-based speech recognition and synthesis and ubiquitous-oriented sentential KSSL recognizer is shown in Fig. 7.

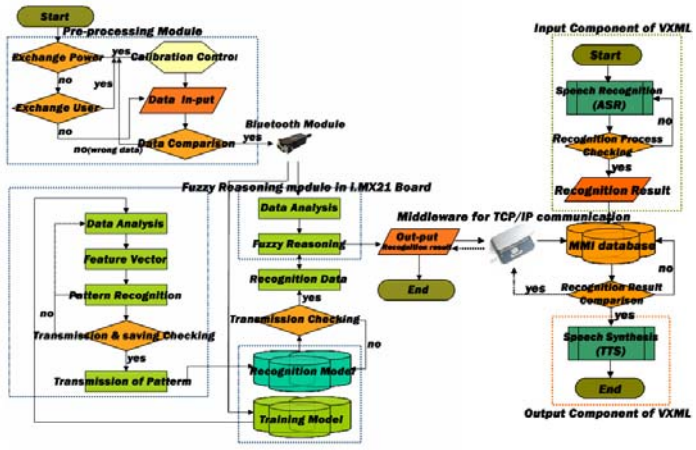


Fig. 7. The flowchart of the MMSDS integrating VXML and KSSL Recognizer

4 Experiments and Results

The distance between the KSSL input module and embedded WPS for processing of the KSSL recognition composed in about radius 10M's ellipse form. When user inputs the KSSL and speech, we move data gloves and receivers of motion tracker to prescribed position. For every 20 reagents, we repeat this action 15 times. While user inputs the KSSL using data gloves and motion tracker, speak using blue-tooth headset of telephone terminal. Experimental results, the uni-modal and MMSDS recognition rate for 32 sentential recognition models are shown in Table 4. Also, the comparison charts are given in Fig.8 respectively.

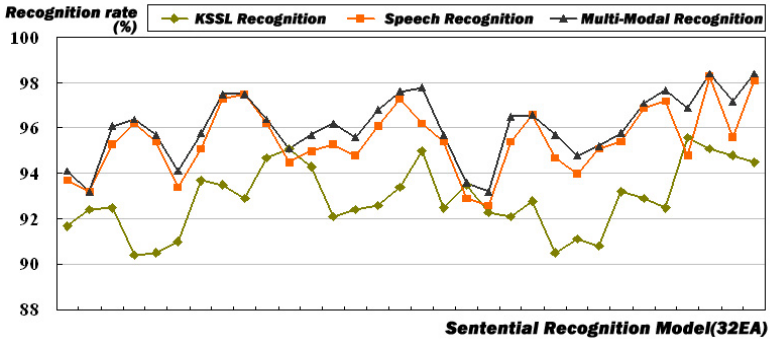


Fig. 8. An average recognition rate of the uni-modal and MMSDS

Table 4. An uni-modal and MMSDS recognition rate about 32 sentential recognition models

Sentential Recognition Model			Uni-modal recognition rate		MMSDS recognition rate
			The KSSL (%)	Speech (%)	The KSSL + Speech (%)
I	go	to museum	92.7	93.7	94.2
		to airport	92.4	93.7	93.9
		to station	92.5	95.8	96.1
	come	from museum	90.4	96.2	96.4
		from airport	92.5	95.4	95.7
		from station	92.4	93.4	94.1
	arrive	at museum	93.7	95.1	95.8
		at airport	94.5	96.9	97.4
		at station	92.9	96.5	96.5
	love	you	94.7	96.2	96.4
	lost	my passport	95.3	94.5	95.1
	am	sorry	94.3	95.0	95.7
		fine	92.1	95.3	96.2
		a Korean	92.4	94.8	95.6
		from Korea	92.6	96.1	96.8
	want	to Seoul	93.4	97.3	97.6
-	thank	you	95.0	96.2	97.8
Are	you	ok?	92.7	95.4	95.7
You	are	welcome	93.5	93.1	93.6
We	go	to museum	92.3	93.6	93.7
		to airport	93.1	95.4	96.5
		to station	92.8	96.6	96.6
	come	from museum	90.5	94.7	95.7
		from airport	92.3	95.1	95.2
		from station	90.8	95.1	95.2
	arrive	at museum	93.2	95.4	95.8
		at airport	92.9	96.9	97.1
		at station	92.5	97.2	97.2
-	good	morning	95.6	94.8	96.1
		afternoon	95.1	98.3	98.4
		evening	94.8	95.6	97.2
		night	94.5	98.1	98.4
An average recognition rate			93.1	95.5	96.1

5 Conclusions

Ubiquitous and wearable computing is an active topic of study, with areas of study including multi-modal user interface design, augmented reality, pattern recognition, using of wearable for specific applications or disabilities. As preliminary study for recognition and representation of KSSL, our researchers implemented hand gesture recognition system that recognize 19 hand gestures according to a shape and stoop degree of hand. Accordingly, with this preliminary study, we implemented WPS-based sentential MMSDS integrating VXML module (speech) and sentential KSSL recognizer (gesture). In experiment results, the MMSDS is more efficient and powerful than uni-modal recognition system that uses one in the KSSL (gesture) or speech. Especially, while the average recognition rate of uni-modal recognition system that use KSSL (gesture) only is 93.1%, the MMSDS deduced an average recognition rate of 96.1% and showed difference of an average recognition rate as much as about 3.0%. In conclusion, we clarify that this study is fundamental study for implementation of advanced multi modal recognizer integrating the human's five senses such as sight, hearing, touch, smell, and taste to take the place of traditional uni-modal for recognizer in natural and sign language processing.

Acknowledgement

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment)" (IITA-2005- C1090-0501-0019) and this work was supported by the Brain Korea 21 Project in 2006.

References

1. M. Fuchs, P. et al.: Architecture of Multi-modal Dialogue System. TSD2000. Lecture Notes in Artificial Intelligence, Vol. 1902. Springer-Verlag, Berlin Heidelberg New York (2000) 433–438
2. Siska Fitrianie. et al.: A Multi-modal Eliza Using Natural Language Processing and Emotion Recognition. TSD 2003. Lecture Notes in Artificial Intelligence, Vol. 2807. Springer-Verlag, Berlin Heidelberg New York (2003) 394-399
3. Wu jiangqin. et al.: A Simple Sign Language Recognition System Based on Data Glove.. ICSP98, IEEE International Conference Proceedings (1998) 1257-1260
4. Stefano Borg.: A Multi-agent Modal Language for Concurrency with Non-communicating Agents. CEEMAS 2003. Lecture Notes in Artificial Intelligence, Vol. 2691. Springer-Verlag, Berlin Heidelberg New York (2003) 40-50
5. Silvia Berti and Fabio Paternò.: Development of Multi-modal Interfaces in Multi-device Environments. INTERACT 2005. Lecture Notes in Computer Science, Vol. 3585. Springer-Verlag, Berlin Heidelberg New York (2005) 1067-1070
6. Mark Barnard. et al.: Multi-Modal Audio-Visual Event Recognition for Football Analysis. IEEE XI11 Workshop on Neural Networks for Signal Processing, IEEE Workshop Proceedings (2003) 469-478
7. Use of Signs in Hearing Communities.: http://en.wikipedia.org/wiki/Sign_language
8. S.-G.Kim.: Korean Standard Sign Language Tutor, 1st, Osung Publishing Company, Seoul (2000)
9. J.-H.Kim. et al.: Hand Gesture Recognition System using Fuzzy Algorithm and RDBMS for Post PC. FSKD2005. Lecture Notes in Artificial Intelligence, Vol. 3614. Springer-Verlag, Berlin Heidelberg New York (2005) 170-175
10. 5DT Data Glove 5 Manual and FASTRAK® Data Sheet.: http://www.5dt.com
11. Richard O. Duda, Peter E. Hart, David G. Stork.: Pattern Classification, 2nd, Wiley, New York (2001)
12. Dietrich Paulus and Joachim Hornegger.: Applied Pattern Recognition, 2nd, Vieweg (1998)
13. J. Schuermann.: Pattern Classification: A Unified View of Statistical and Neural Approaches, Wiley&Sons (1996)
14. Relational DataBase Management System.: http://www.auditmypc.com/acronym/RDBMS.asp
15. Oracle 10g DW Guide.: http://www.oracle.com
16. W. B. Vasantha kandasamy.: Smaranda Fuzzy Algebra. American Research Press, Seattle (2003)
17. Scott McGlashan et al.: Voice Extensible Markup Language (VoiceXML) Version 2.0. W3C Recommendation, http://www.w3.org (1992)