IEICE TRANS. FUNDAMENTALS, VOL.E96–A, NO.11 NOVEMBER 2013

INVITED PAPER *Special Section on Smart Multimedia & Communication Systems*

A Robust Speech Communication into Smart Info-Media System

Yoshikazu MIYANAGA[†a)], ***Fellow***, **Wataru TAKAHASHI**[†], ***Student Member***, ***and*** **Shingo YOSHIZAWA**[††], ***Member***

SUMMARY This paper introduces our developed noise robust speech communication techniques and describes its implementation to a smart info-media system, i.e., a small robot. Our designed speech communication system consists of automatic speech detection, recognition, and rejection. By using automatic speech detection and recognition, an observed speech waveform can be recognized without a manual trigger. In addition, using speech rejection, this system only accepts registered speech phrases and rejects any other words. In other words, although an arbitrary input speech waveform can be fed into this system and recognized, the system responds only to the registered speech phrases. The developed noise robust speech processing can reduce various noises in many environments. In addition to the design of noise robust speech recognition, the LSI design of this system has been introduced. By using the design of speech recognition application specific IC (ASIC), we can simultaneously realize low power consumption and real-time processing. This paper describes the LSI architecture of this system and its performances in some field experiments. In terms of current speech recognition accuracy, the system can realize 85–99% under 0–20 dB SNR and echo environments.

key words: *smart info-media system, robust speech recognition, voice activity detection, speech rejection, ASIC, low power consumption design*

1. Introduction

Smart systems have been recognized as sophisticated, intelligent, and advanced systems. One of them, a smart human-interface system, is embedded with a high accuracy speech recognition module. Such a smart info-media system can support the next generation products in the near future.

Speech recognition modules have been embedded into smart home electronics, smart mobiles, smart cars, and smart multi-media entertainment systems. By using speech recognition, a sophisticated human interface can be realized in any system and has compensated for a certain number of digital divide problems. The development of speech analysis and recognition has a long history, and thus some high performance speech recognition technologies have already been implemented [1]–[7].

One current speech recognition system consists of a flame-based phonemes speech recognition and language model [2]–[7]. As another speech recognition approach, a phrase based speech recognition has been also explored [8]–[19]. Compared with the system with the flame-based phonemes speech recognition and language model, the phrase based speech recognition can provide higher recognition accuracy in various noise environments. On the other hand, it requires high calculation cost and many training databases for all target speech phrases. By using an efficient LSI design for a speech recognition system, the authors have been able to realize real-time speech recognition with low power consumption even although the total calculation cost becomes high [8], [11], [13], [17].

As advanced technology using the above phrase based speech recognition techniques, a speech communication system has been proposed, and this paper presents an overview of this system and its performance. The speech communication system consists of voice activity detection (VAD) under noisy conditions, noise robust speech recognition, and noise robust speech rejection mechanism. The embedded speech communication module recognizes all target speech phrases with high recognition accuracy where non target speeches are automatically rejected. The smart robot system can answer only for the target speech phrases. In terms of recognition accuracy, the system can realize 85–99% under 0–20 dB SNR and echo environments.

In addition to these methods, this paper presents also the architecture of this speech communication system. The LSI hardware has been implemented into a small robot system as a smart info-media robot system. This system adds the whole speech communication system above, a speech synthesizer, and some controllers to intelligent home electronics.

2. Speech Communication System

To design an automatic speech recognition (ASR) system suitable for real environments, we have to consider first the conditions under which the designed ASR may efficiently work. As one important condition, the types of noises and their amplitudes should be described. In addition to white noise, any various colored noises are considered, i.e., factory noise, car noise, home electronics noise, city sounds, street noises, speech noises, etc. Depending on such various noise characteristics and their amplitudes, the performance of the ASR drastically changes. For example, some ASRs can barely recognize any speech outdoors although they can almost perfectly recognize speech in all silent rooms.

Manuscript received March 5, 2013.
Manuscript revised June 5, 2013.

[†]The authors are with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060-0814 Japan.
[††]The author is with the Department of Electrical and Electronic Engineering, Kitami Institute of Technology, Kitami-shi, 090-8507 Japan.
a) E-mail: miya@ist.hokudai.ac.jp

DOI: 10.1587/transfun.E96.A.2074

The condition of the microphone should be also described carefully as prior information. The speech waveform recorded through an AD converter has different properties when the distance between speaker's mouse and the location of microphone is changed. In particular, some conventional ASRs can only recognize a speech waveform with an attached microphone whose distance may be around 5–10 cm. In other words, it is quite difficult for these ASR systems to recognize speech by using a short distance microphone whose distance is 10 cm–3 m and a long distance microphone whose distance is 3–5 m.

The conventional ASR assumes normally (1) 20 dB white noise, which indicates silent rooms, and (2) an attached microphone, which indicates short distance. However, when we consider an embedded ASR system, e.g., ROBOT with ASR, many poor surroundings and long distance conditions should be considered. In case of an exhibition room, out-door, and general house environments, the SNR should be always lower than 20 dB. In case of free-hands and general human communications, the 10 cm–5 m distances should be also considered. Our developed ASR can recognize speech under the above various conditions accurately. The concrete conditions used in our development are considered as (1) white noise, speech noise, and factory noise with 5 dB SNR, (2) an echo environment less than 150 ms, and (3) medium distance (< 5 m) from a microphone.

To keep the highest performance of ASR, our system only recognizes the set of selected phrases. In other words, any other words and phrases except target speech are rejected by our system. In Fig. 1, the overview of our speech communication system is described. The system consists of automatic speech waveform detection, i.e., automatic voice activity detection (VAD), automatic speech recognition (ASR), and automatic speech rejection. The total system has been designed with hardwire and it has been realized by an LSI system. Using this LSI system, we can realize our speech communication system with real-time processing and low power consumption.

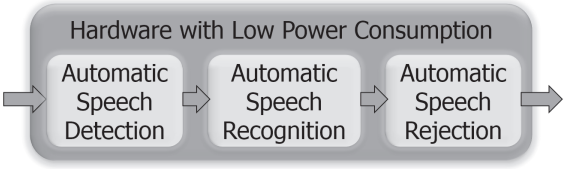

Fig. 1 Our speech communication system.

3. Noise Robust Processing

Several noise robust processes are implemented into automatic speech detection, recognition, and rejection. For the introduction of these techniques, a running spectrum domain should be explained first. Figure 2 shows the example of speech waveform and its mel-spectra on the running spectrum domain.

An observed speech waveform is divided into several frames. For each frame, Fourier power spectrum and/or cepstrum are calculated. In Fig. 2, mel-spectra are described. On the running spectrum domain, the time varying characteristics are well depicted, and thus these properties are used for speech processing. Note that the vertical axis indicates the index of frame and the horizontal axis indicates the frequency components of mel-spectra.

In the automatic speech detection, both band-pass filtering and fundamental frequency detection have been implemented. By using band-pass filtering, only the frequency band including high energy components of speech is selected. The noises whose frequency components are located out of the above frequency band are reduced. After BP filtering, the time locations including high-energy components of speech are selected. In addition, the estimates of fundamental frequencies are also used for this selection at the same time. Since the average energy at the fundamental frequency is normally higher than noise energy even under noisy conditions, the results on the fundamental frequency estimates contribute the high accuracy of automatic speech detection.

In ASR, we have to carefully consider how to reduce various noise effects and at the same time how to hold all important speech characteristics. In recent years, many robust speech recognition have been developed [20]–[32]. In some of these methods, the noise reduction may not be sufficiently obtained, and prior information is also required. To improve

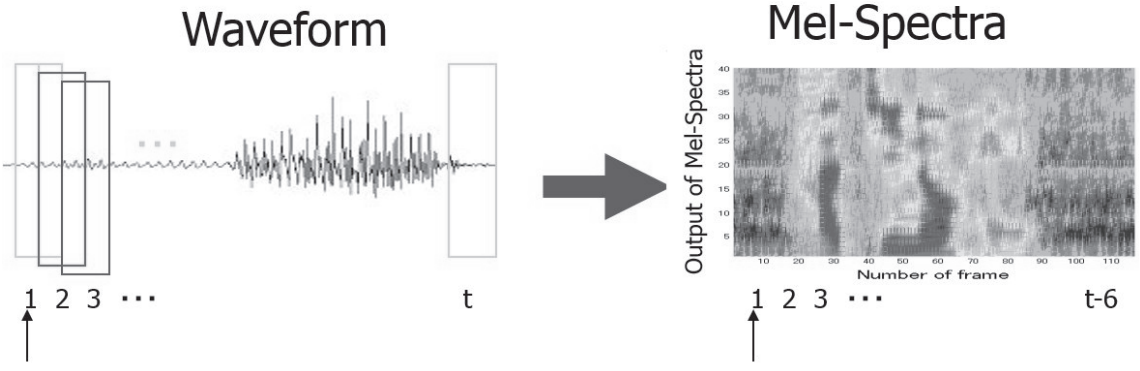

Fig. 2 Speech waveform covered by flames and its mel-cepstrum on running spectrum domain (RSD).

on the conventional methods, we have developed Running Spectrum Filtering (RSF) with Dynamic Range Adjustment (DRA) for our speech communication system [8]–[19].

As the post-processing module in the developed speech communication system, the robust noise/speech rejection technique has been implemented. The basic idea is based on that of many other conventional ones. In other words, the likelihood values calculated from all Hidden Markov Models (HMM) are evaluated for this rejection.

Figure 3 shows two simple examples. It shows the values from the maximum to the lower likelihoods given by HMMs. In the left-hand side graph, the maximum value is noticeably larger than others. This means the HMM whose indicates the maximum value definitely fits the Mel Frequency Cepstral Coefficient (MFCC) data given from an observed speech. The recognition result from the label of this HMM can be trusted. On the other hand, if the maximum likelihood value is similar to the next large value in Fig. 3(b), it may be difficult for this system to recognize whether the label of the HMM whose likelihood becomes the maximum is correct or not.

Although the basic idea above is simple and conventional, such an ideal case is almost never obtained under various noise conditions. The advanced criteria from the above consideration are depicted in Fig. 4.

The noise robust speech rejection method used in this system has several criteria. The criterion of "Tendency" has already been explained in Fig. 3. If the values from the maximum to the minimum are defined as M_1, M_2, M_3, ..., the criterion of "Tendency" calculates $M_1 - M_2$, and then its value is evaluated. In other words, if the value of $M_1 - M_2$ is larger than a threshold, it becomes valid. Otherwise, it is invalid. In addition, the criteria of "Maximum Score", "Ratio", "Square of Ratio", "MFCC Cluster", and "Group Evaluation of MFCC Cluster" show the following meanings:

1. Maximum Score: Evaluation of the maximum likelihood value, i.e., M_1. If the value is lower than a threshold, the result is rejected.
2. Ratio: Evaluation of the ratio between the largest and second largest values. The ratio is calculated as M_1/M_2.
3. Square of Ratio: Evaluation of $(M_1/M_2)^2$.
4. MFCC Cluster: Among all labels of HMMs, there are several similar pronunciations, e.g., "denki" and "genki", and the others come from different pronunciations. The cluster that has phrases with similar pronunciations should be evaluated by using levels on criteria different from others.
5. Group Evaluation of MFCC Cluster: Even if we evaluate the criterion of "MFCC Cluster", there are still slightly different properties between male and female among them. This criterion is applied when the above criterion of "MFCC Cluster" does not indicate clear difference.

For each HMM and its phrase label, the above criteria are applied, and all required thresholds are automatically trained by using all training speech databases.

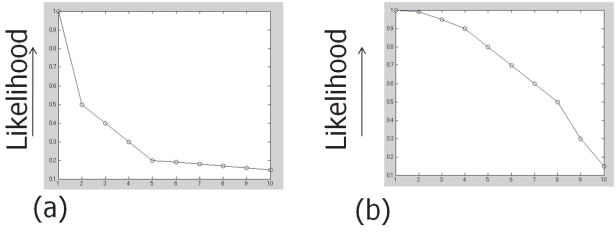

Fig. 3 Basic tendency on HMM likelihood values.

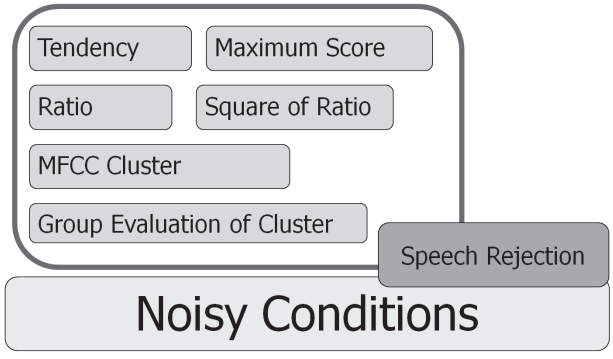

Fig. 4 Speech rejection method under noises.

4. Noise Robust Techniques in Speech Recognition

This section explains our proposed noise robust methods: Running Spectrum Filtering (RSF) and Dynamic Range Adjustment (DRA).

4.1 Running Spectrum Filtering (RSF)

Speech sounds usually change as time progresses. On the other hand, some noise components do not change radically. Therefore, when such time-varying components can be held and time invariant components are reduced, these noises are finally reduced.

If we use modulation spectrum domain (MSD), the above rhythm can be represented separately [9], [12], [19]. Figure 5 shows the relationship between running spectrum domain (RSD) and MSD. In the running spectrum domain,

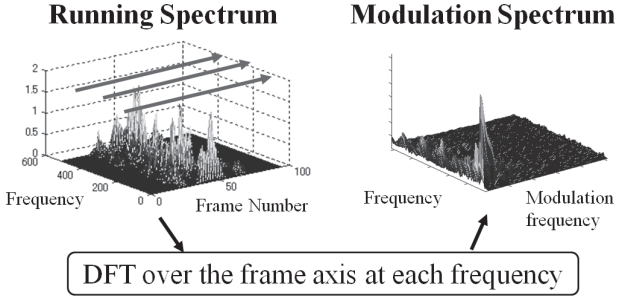

Fig. 5 Modulation spectrum domain.

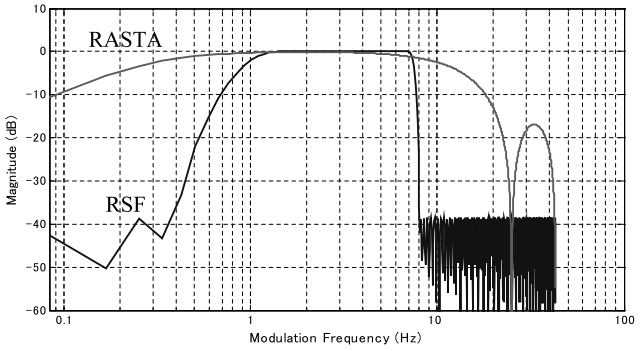

Fig. 6 RASTA and RSF filter in MSD.

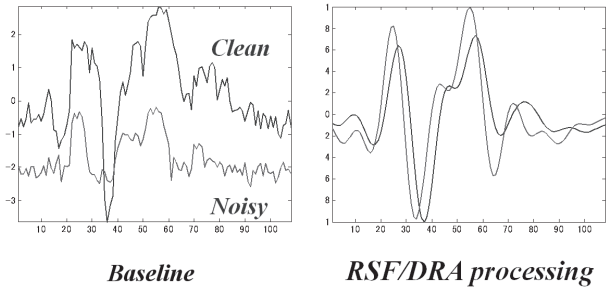

Fig. 7 Comparisons of MFCC time trajectory between clean and noisy speeches.

the time trajectory at a specific frequency is obtained by tracing its values at each flame. From the time trajectory, we can obtain a modulation spectrum by applying DFT to this trajectory.

It has been reported [9], [12], [19], [28] that speech components in MSD are dominant around 4 Hz, i.e., 1–6 Hz, and the range from 0 to 1 Hz and beyond 7 Hz can be regarded as noise. Therefore, speech components can be extracted by applying the band-pass filtering with 1–6 Hz on RSD.

Relative Spectral Transform (RASTA) is a well known method focusing on the modulation spectrum. RASTA employs IIR band-pass filters and removes noise components. However, IIR filters cause phase distortion. On the other hand, RSF employs FIR filters instead of IIR filters. This makes RSF stable and free from phase distortion. However, RSF requires high-order FIR filters to realize sharp modulation frequency cut-off, and such high order of FIR filters causes many delays. For example, to realize the modulation frequency properties of RSF shown in Fig. 6, 240 taps are required. Then, the required length of non-speech periods $l[sec]$ before and after the input speech is given by

$$l = \frac{\text{the number of taps} * \text{frame-shift}}{2 * \text{sampling rate}}. \quad (1)$$

4.2 Dynamic Range Adjustment (DRA)

The dynamic range of cepstrum indicates the difference between the maximum and the minimum values of time varying cepstral trajectory over frame axis. The dynamic range of speech energy drastically changes under additive noises. It causes the decrease of cepstral dynamic ranges. As a result, the speech recognition performance is seriously damaged under noise conditions since both the maximum and the minimum values play important roles as the characteristics of speech and are corrupted by noise. Figure 7 shows the distributions of the dynamic ranges and proves that dynamic range is usually reduced by additive noise even if RASTA or RSF is applied. The baseline in Fig. 7 indicates the difference between the dynamic range of clean speech and its noisy speech. On the other hand, the RFS/DRA processing in Fig. 7 shows how it compensates for these differences.

DRA normalizes amplitudes of a speech feature vector with the maximum amplitude. In DRA, the amplitude of a cepstrum is adjusted in proportion to its maximum amplitude as

$$\tilde{f}_i(n) = f_i(n) / \max_{j=1,\cdots,m} |f_j(n)| \quad (i = 1, \cdots, m), \quad (2)$$

where $f_i(n)$ denotes an element of the time varying cepstrum trajectory, m denotes the dimension and n denotes the frame number. By using (2), all amplitudes are adjusted into the range from -1 to 1.

With DRA/RSF, speech analysis is refined as shown in Fig. 7.

4.3 Multi-Conditions Based HMM

By using **4.1** and **4.2**, the noise robust technique effective against various noise sources can be obtained in our system. However, to realize a robust echo technique, an additional approach must be developed and implemented in the stage of HMM training. When we consider the differences between the features of speech sounds without echo and with any echo situations, their characteristics are changed. Accordingly, if HMM is well trained by using many speech databases but no echo speech database, it cannot recognize speech under echo conditions. However, if the HMM is trained by simultaneously using a speech database without echo conditions and another database with echo conditions, HMM can calculate accurately the likelihood even for an echo speech waveform.

Although there are many echo conditions (e.g., small echo (1–2 msec), medium echo (100–300 msec), and strong echo (500 msec and more)) the echo robust training can be sufficiently completed only when some typical echo conditions are applied in its training. In our developed system, an echo speech database is created from echo free speech data with 150 msec echo model. By using these echo free and echo speech databases, all HMM are trained in our system.

5. LSI Design of Speech Communication System

As another important property in human-machine interface,

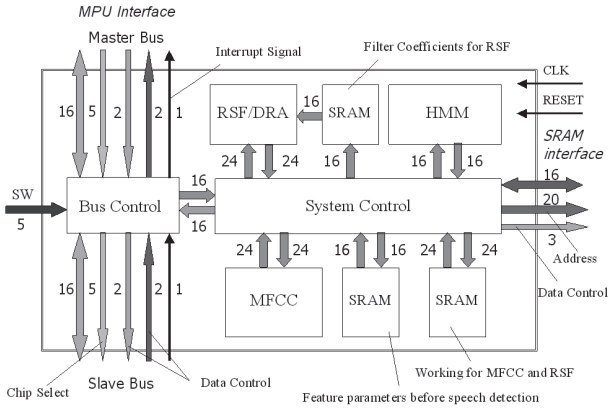

Fig. 8 Block diagram of speech communication system.

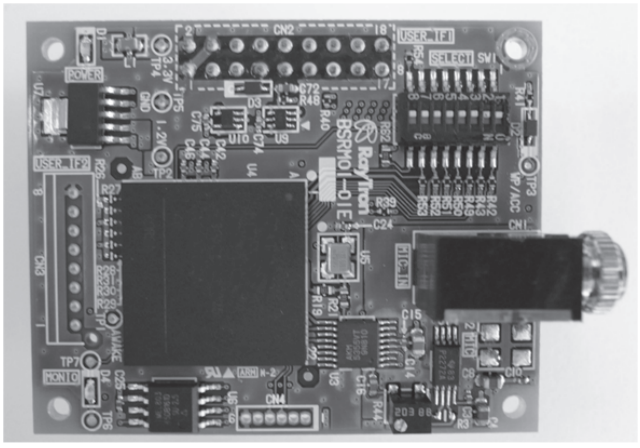

Fig. 9 Speech communication board.

a response time should be considered. When a general human interaction is observed, 100–200 msec responses can keep a real-time response. In other words, if the system can answer a user within 200 msec, a user feels its response is in real-time. Compared with other real-time processing (i.e., parallel/pipeline computer, wireless communication, and multi-media machine to machine communications), the time period of 200 msec seems to be long, and thus such systems are not difficult to design. However, a system with simultaneous low power consumption and real-time processing is complicated to design.

To realize both low power consumption and real-time processing, we have developed a full custom LSI for our speech communication system. Figure 8 shows a block diagram of it. The architecture consists of a HMM calculation module, a noise robust processing module, and a MFCC calculation module. In addition, the master and slave connections are given through BUS CONTROL, and thus several of the same LSIs can be connected in a system.

The first designed LSI was fabricated with Rohm CMOS 0.35 μm. It can respond with its recognition result within 0.180 msec/phrase, 10 MHz clock, and 93.2 mW power consumption where 1,000 phrases are recognized by using this LSI. When this LSI is applied to a product, the recognition results for 1,000 target phrases can be obtained in 180 msec. Real-time processing has been realized for 1,000 target phrases. By using parallel connections of these LSIs, the system can recognize more than 1,000 phrases within the same response time. The specifications are as follows:

1. Phrase ASR system where a maximum of 1,000 phrases can be registered.
2. FLASH memory of NOR type 32-bit and 16-bit ADC with at most 44.4-kHz sampling frequency are embedded.
3. The host connections are Serial to Peripheral Interface (SPI), Inter-Integrated Circuit (I2C), and Universal Asynchronous Receiver Transmitter (UART).
4. The maximum clock is 22.5792 MHz, and 3.3 v power voltage is supplied. The 180 msec response time can be realized with 10 MHz clock.
5. The small scale size is L: 55 mm × W: 44 mm × H: 12 mm.

Environment	Noise Level	Accuracy
Meeting Room	50 dB	96.4%
Elevator	50 dB	95.0%
Stairs	45 dB	85.1%
Car A (Idling, No-Moving)	50 dB	99.4%
Car B (High Speed, Open Window)	75 dB	93.3%
Car C (High Speed, Audio ON (FM))	75 dB	88.9%
Cruiser Board (Outside, high speed)	80 dB	82.7%

Fig. 10 Accuracy of speech communication system board.

Figure 9 shows the total system, which includes a speech communication system, several interfaces, and a microphone. The results of the field experiments using this evaluation board are shown in Fig. 10.

In these experiments, the microphone distances are 30 cm, 60 cm, and 90 cm where these distances are generally required from many users. The total number of phrases is 15. Its number seems to be quite small. However, some users who want to apply such developed ASR for a command speech input interface, request a small number of phrases. For the experiments on echo robustness, the conditions were a meeting room 6,400 mm × 3,200 mm × 2,400 mm, an elevator 1,800 mm × 1,500 mm × 2,300 mm, and the stairwell of the 12th floor of an office in a concrete building. For the experiments of noise robustness, the conditions were a parked four-door sedan car with closed windows, a four-door sedan car driven at 80 km/h with open windows (120 mm), and a four-door sedan car driven at 80 km/h driving with closed widows and the radio turned on (FM broadcasting). Note that the four speakers for a car radio system were located in the front and the rear. Figure 10 shows the performance of the embedded ASR system. For all conditions, the current system can show high accuracy.

For the total performance in echo and noise conditions,

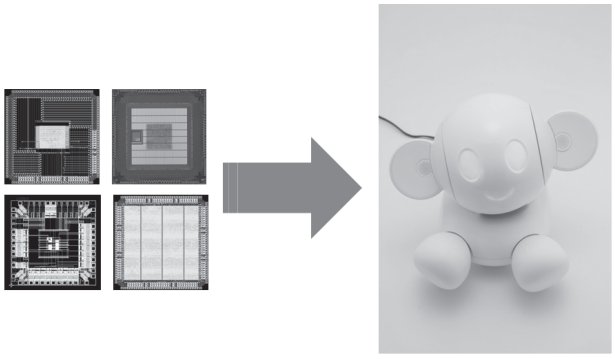

Fig. 11 Chapit.

the developed ASR system can realize 93.0% correctness (speech recognition accuracy and rejection accuracy). As an additional experiment, our system was also evaluated in the conditions of a moving small ship. The noisy level was around 80 dB, and it was difficult for us to talk even to each other under its noisy level. In this condition, 82.7% correctness was obtained. When we use this system in exhibitions, the system can recognize almost all words under the condition of 70–75 noisy level and 2–5 m microphone distance.

6. Speech Communication Module Embedded Smart Info-Media System

The speech communication module shown in Fig. 9 was embedded into the small robot shown in Fig. 11. The total system of this small robot was designed by Raytron [33]. The two microphones at the ears of this robot are integrated, and thus sounds from in front of this robot are mainly recoded and analyzed.

This robot is named Chapit, and its completion was announced in late 2007. It can currently recognize only 300 words. However, a user can say any words to Chapit within 30 cm–5 m distance and 5–30 dB noisy/echo conditions, and then it answers for the registered phrases. Its interface style seems to be quite similar to human-to-human communications. However, this interface is new to human-to-robot communications.

7. Conclusion

This paper introduced a speech communication system. It can provide high-accuracy speech recognition under noisy and echo conditions. The robust techniques are implemented into automatic speech detection, recognition, and rejection. In terms of current speech recognition accuracy, the system can realize 83–99% under 0–20 dB SNR and 150 msec echo environments.

The smart system has many meanings, and the advanced speech communication is regarded as a smart human interface system. The speech communication technology in this paper approaches human speech recognition ability from the viewpoint of the sound processing level. However, this technology needs to be integrated with intelligence processing and language processing levels. This will help to advance smart info-media systems.

Acknowledgement

The authors would like to thank RayTron, INC. and the VLSI Design Education and Research Center (VDEC), Tokyo University for fruitful discussions. This study is supported in parts by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (A1) (24240007), the Japan Science and Technology Agency for A-Step Program (AS2416901H) and KDDI Laboratories.

References

[1] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for mono-syllabic word recognition in continuously spoken sentences," IEEE Trans. Speech Signal Processing, vol.28, no.4, pp.357–366, 1980.
[2] S. Furui, "Speaker-Independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-34, no.1 pp.52–59, Feb. 1986.
[3] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol.77, no.2, pp.257–286, Feb. 1989.
[4] K.M. Knill, et al., "Use of Gaussian selection in large vocabulary continuous speech recognition using HMMs," Proc. ICSLP96, pp.470–473, 1996.
[5] X. Huang, Spoken Language Processing, Prentice Hall, 2001.
[6] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Tech. Rep., Sun Microsystems Inc., 2004.
[7] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp.131–137, Oct. 2009.
[8] S. Yoshizawa, Y. Miyanaga, N. Wada, and N. Yoshida, "A low-power LSI design of Japanese word recognition system," Proc. IEICE International Technical Conference on Circuits/Systems, Computers and Communications 2002, vol.1, no.1, pp.98–101, 2002.
[9] N. Wada, Y. Miyanaga, N. Yoshida, and S. Yoshizawa, "A consideration about an extraction of features for isolated word speech recognition in noisy environments," ISPACS2002, DSP2002-33, pp.19–22, Nov. 2002.
[10] N. Hayasa, Y. Miyanaga, and N. Wada, "Running spectrum filtering in speech recognition," SCIS Signal Processing and Communications with Soft Computing, pp.210–213, Oct. 2002.
[11] S. Yoshizawa, N. Wada, N. Hayasaka, and Y. Miyanaga, "Noise robust speech recognition focusing on time variation and dynamic range of speech feature parameters," Proc. International Symposium on Intelligent Signal Processing and Communication Systems 2003, vol.1, pp.484–487, Dec. 2003.
[12] N. Wada, S. Yoshizawa, and Y. Miyanaga, "A consideration about robust speech feature extraction for isolated word speech recognition," Proc. International Symposium on Intelligent Signal Processing and Communication Systems 2003, vol.1, pp.478–483, Dec. 2003.
[13] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, "Cepstral amplitude range normalization for noise robust speech recognition," Proc. IEICE Trans. Inf. & Syst., vol.E87-D, no.8, pp.2130–2137, Aug. 2004.
[14] S. Yoshizawa, N. Wada, N. Hayasaka, and Y. Miyanaga, "Scalable

architecture for word HMM-based speech recognition," Proc. IEEE ISCAS2004, pp.417–420, 2004.
[15] N. Hayasaka, S. Yoshizawa, N. Wada, Y. Miyanaga, and N. Hataoka, "A study of robust speech recognition system and its LSI design," Transactions on The Society of Instrument and Control Engineers, vol.41, no.5, pp.473–480, May 2005.
[16] S. Yoshizawa and Y. Miyanaga, "Robust recognition of noisy speech and its hardware design for real time processing," ECTI Transaction on Electrical Eng., Electronics, and Communications (EEC), vol.3, no.1, pp.36–43, Feb. 2005.
[17] S. Yoshizawa, N. Wada, N. Hayasaka, and Y. Miyanaga, "Scalable architecture for word HMM-based speech recognition and VLSI implementation in complete system," IEEE Trans. Circuits Syst. I, vol.53, no.1, pp.70–77, Jan. 2006.
[18] N. Hayasaka and Y. Miyanaga, "Spectrum filtering with FRM for robust speech recognition," Proc. 2006 IEEE International Symposium on Circuits and Systems, vol.1, pp.3285–3288, May 2006.
[19] N. Wada, N. Hayasaka, S. Yoshizawa, and Y. Miyanaga, "Direct control on modulation spectrum for noise-robust speech recognition and spectral subtraction," Proc. 2006 IEEE International Symposium on Circuits and Systems, vol.1, pp.2533–2536, May 2006.
[20] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, no.2, pp.113–120, 1979.
[21] J. Tierney, "A study of LPC analysis of speech in additive noise," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-28, no.4, pp.389–397, Aug. 1980.
[22] S.M. Kay, "Noise compensation for autoregressive spectral estimation," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-28, no.3, pp.292–303, March 1980.
[23] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise," Proc. IEEE ICASSP, pp.845–848, 1990.
[24] F. Martin, K. Shikano, Y. Minami, and Y. Okabe, "Recognition of noisy speech by composition of hidden markov models," IEICE Technical Report, SP92-96, Dec. 1992.
[25] M.J.F. Gales and S.J. Young, "Cepstral parameter compensation for HMM recognition in noise," Speech Communication, vol.12, no.3, pp.231–239, 1993.
[26] J.A.N. Flores and S.J. Young, "Continuous speech recognition in noise using spectral subtraction and HMM adaptation," Proc. ICASSP, vol.I, pp.409–412, 1994.
[27] M. Rahim and B.H. Juang, "Signal bias removal for robust telephone based speech recognition in adverse environments," Proc. ICASSP-94, pp.I-445–I-448, April 1994.
[28] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech Audio Process., vol.2, no.4, pp.578–579, Oct. 1994.
[29] K. Takagi, et al., "Rapid environmental adaptation for robust speech recognition," Proc. ICASSP95, pp.149–152, 1995.
[30] K. Aikawa, H. Hattori, H. Kawahara, and Y. Tohkura, "Cepstral representation of speech motivated by time-frequency masking: An application to speech recognition," J. Acoust. Soc. Am., vol.100, no.1, pp.603–614, July 1996.
[31] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," Proc. European Conference On Speech Communication And Technology, vol.3, pp.1079–1082, Sept. 1997.
[32] M. Shozakai, S. Nakamura, and K. Shikano, "An evaluation of speech enhancement approach ECMN/CSS for speech recognition in car environments," IEICE Trans. Inf.& Syst. (Japanese Edition), vol.J81-D-II, no.1, pp.1–9, Jan. 1998.
[33] Y. Miyazaki and Y. Miyanaga, "New development and enterprise of robust speech recognition systems," 2010 International Workshop on Information Communication Technology, no.1, WA-2-1 (CD-ROM), Aug. 2010.

Yoshikazu Miyanaga is a professor in Graduate School of Information Science and Technology, Hokkaido University. He is an associate editor of Journal of Signal Processing, RISP Japan (2005-present). He was a chair of Technical Group on Smart Info-Media System, IEICE (IEICE TGSIS) (2004–2006) and he is now a member of the advisory committee, IEICE TGSIS (2006-present). He is a vice-President, Asia-Pacific Signal and Information Processing Association (APSIPA). He was a distinguished lecture (DL) of IEEE CAS Society (2010–2011) and he is now a Board of Governor (BoG) of IEEE CAS Society (2011-present).

Wataru Takahashi received an M.E. degree from Hokkaido University, Japan in 2010. He is now a doctoral student in the Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan.

Shingo Yoshizawa received the B.E., M.E., and Ph.D. degrees from Hokkaido University, Japan in 2001, 2003 and 2005, respectively. He was an Assistant Processor in the Graduate School of Information Science and Technology, Hokkaido University from 2006 to 2012. He is currently an Associate Professor in Department of Electrical and Electronic Engineering, Kitami Institute of Technology. His research interests are speech processing, wireless communication, and VLSI architecture.