*Acoust. Sci. & Tech.* **28**, 3 (2007) 

**PAPER**

# Speech recognition interface system for digital TV control

Yoshihisa Nakatoh*, Hiroyasu Kuwano, Takeo Kanamori and Masakatsu Hoshimi

*Matsushita Electric Industrial Co., Ltd.,*
*1006 Kadoma, Kadoma, 571–8501 Japan*

(*Received 8 June 2006, Accepted for publication 24 November 2006*)

**Abstract:** In this paper, we describe a speech recognition interface system for digital TV (DTV) control. TV systems are currently undergoing digitalization and will become more multifunctional, leading to more complex TV operations. Thus, it is necessary for everyone to be able to use TVs easily, and a speech recognition interface is an important key technology. A speech recognition system, which is designed for home use, particularly for digital TV, must be simple and robust to environmental noises and speaker variations. To provide robustness to noise, we developed a noise reduction technique for house noise and an echo-canceling technique for TV sound. To achieve robustness to speaker variations, we developed new speaker adaptation techniques which are incorporated in the system. These of technologies results in a significant improvement in the recognition performance of the DTV.

**Keywords:** Digital TV, Speech recognition, Noise suppression, Speaker adaptation

**PACS number:** 43.72.Ne [doi:10.1250/ast.28.165]

## 1. INTRODUCTION

In recent years, speech recognition technology has been produced commercially in various products, such as videos [1], car navigation systems, and cellular phones. It is one of the key technologies that allows everyone to use appliances easily. We have developed the model voice method [2], which is a recognition system with a small vocabulary with a low computational cost for cellular phones. We have also developed the COMPATS (COntinuous Matching of PArameters of Time Spectral form) method [3,4], a recognition system with a the large vocabulary for car navigation.

In this paper, we describe speech recognition technology for a DTV [5]. Digital multi-channel TV broadcasting services in Japan, which started in December 2000, provide a wide range of information services such as a huge number of TV program channels in digital and analog forms, a TV program schedule distribution service and data broadcasting services (e.g., weather forecasts). Although the TV has become multifunctional, its operation has become complex. With further increases in the range as well as in the variety of information services, the TV set will increase its multifunctionality and its operation will become even more complex. Therefore, we proposed a speech recognition system as DTV interface and released a commercial product. Considering a speech recognition system for home application, automatic speech recognition (ASR) systems must be robust to environment noises and speaker variations. For many years, there have been many studies, the results of which have increased the robustness of ASR systems to noises. For example, AURORA [6] is famous as a study that targeted the cellular phone. In a real environment using speech recognition, the robustness to various noises, such as the sound outputted from a loudspeaker of DTV and various noises in the house, is important. We developed a noise reduction technique for noises in the house and an echo-canceling technique for TV sound. On the other hand, the diversity of speakers is caused by the difference between the characteristics of the vocal organs of children and aged persons. In recent years, there have been several studies on ASR systems for which the users can be any age, from children to aged persons [7,8]. We developed a speaker adaptation technique, a normalization technique with a frequency-warping procedure using an age-dependent acoustic model. With the implementation of these techniques and the design of a user-friendly interface, a DTV with speech recognition remote control was developed. In the following sections, we introduce its user interface, the ASR system, the noise reduction technique and the speaker adaptation technique and discuss their performance.

*e-mail: nakatoh.yoshihisa@jp.panasonic.com

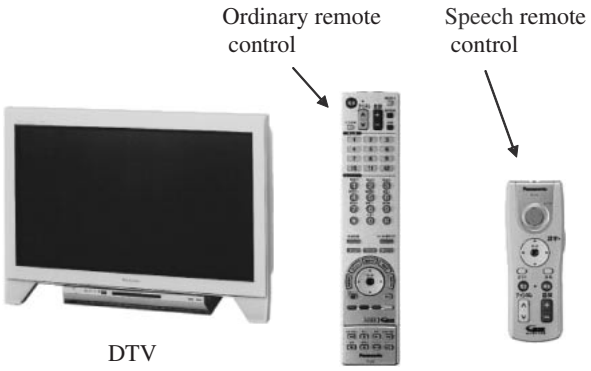


**Fig. 1** Speech remote control and the ordinary remote control.

## 2. USER INTERFACE

The main features of the DTV user interface with speech recognition are a speech-enabled remote control, a TV set equipped with special hardware, a shortcut speech command application and an interactive interface using eye-catching symbols. More details about these features are explained in the following sections.

### 2.1. Speech Remote Control

A speech remote control and an ordinary remote control are shown in Fig. 1. Because the remote control must be ready for immediate use, our proposed speech remote control has the following characteristics:

1) Small number of buttons

The ordinary remote control has more than 70 buttons. On the other hand, the speech-enabled remote control has only 14 buttons.

2) Comprehensive ability

The speech remote control is able to accept almost all commands that can be inputted by button operations using the ordinary remote control.

3) Easy-to-grip shape

To allow the users to hold the remote control close to their mouths, the speech remote control is designed to be easy to grip with a press-talk switch on its side.

4) Other features

Speech input is useful when inputting complex commands whereas button input is useful for simple operations such as changing the volume. On the basis of such a total-usability design idea, a minimum set of keys such as cursor keys, volume control keys and channel up and channel down keys are included on the speech remote control.

### 2.2. Shortcut Speech Commands

One remarkable feature of the speech remote control is a shortcut command consisting of a single utterance. The following functions can be controlled by means of the speech recognition modality:

1) Channel Selection:

Changes the channel by uttering either the nickname of the broadcaster or the channel number.

2) Category Search:

Locates a TV program category with an utterance, for example, "Soccer." An example of the category search operation scheme is shown in Fig. 2. The control operation layers, connected by solid lines, explain the scheme using the ordinary remote control. Without the speech remote control, the category search requires several steps through a category tree, such as "Program Control," "Category Lists," "Sports" and then "Soccer." Only one utterance of "Soccer" shows the soccer program list using the speech remote control, whereas five selections are required when using the ordinary remote control.

3) Quick Control:

Enables quick adjustment of what, including volume control, and recording and playing a video. The one-utterance control saves time and effort.

## 3. ASR SYSTEM

### 3.1. Hardware of Speech Remote Control

The diagram of the speech remote control system for a

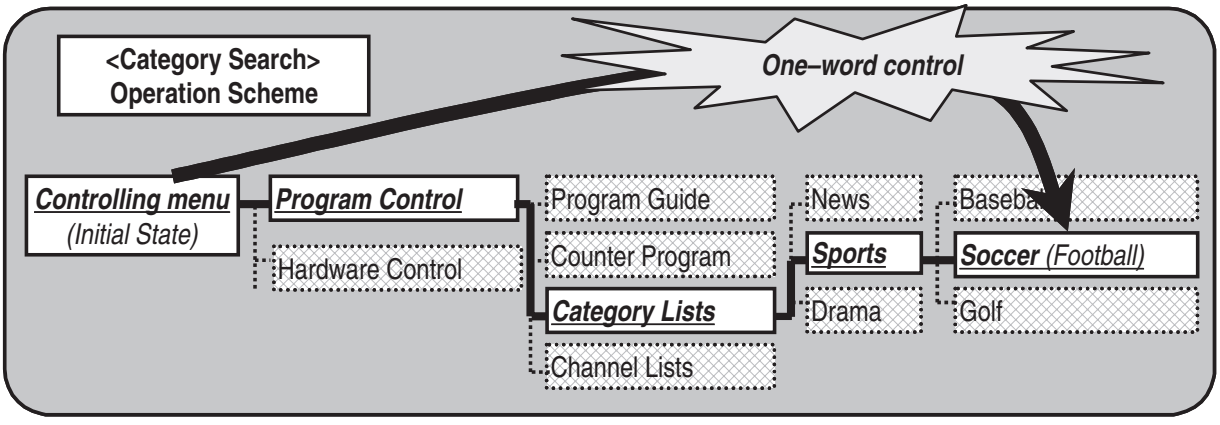


**Fig. 2** Category search operation scheme to search for programs.

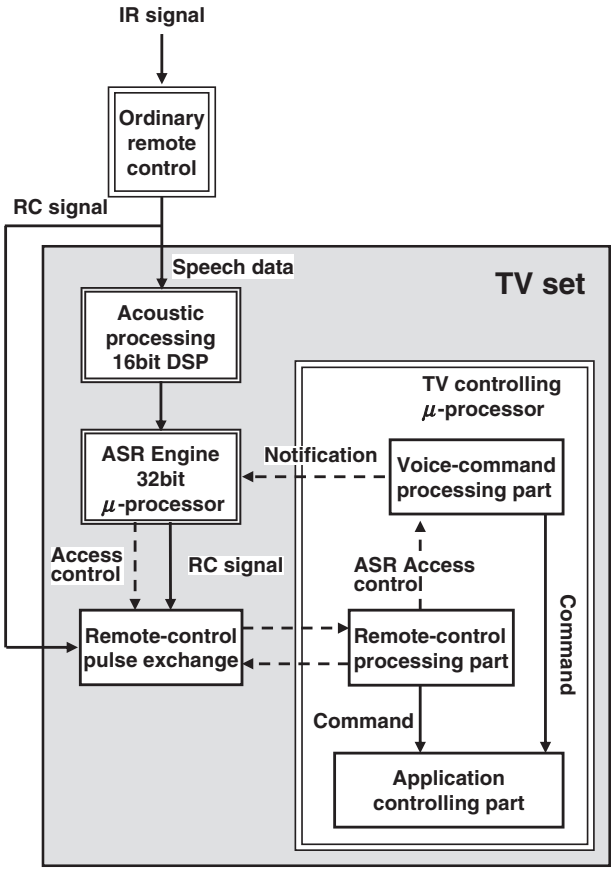


**Fig. 3** Diagram of the speech remote control system.

DTV is shown in Fig. 3. The TV speech control and ordinary remote control were can be simultaneously and priority for exclusive access control is given to the speech remote control. When the ordinary remote control is used and a button is pushed, a remote-control signal from the ordinary remote control is transmitted to the remote-control pulse exchange unit. On the other hand, when the speech remote control is used and the press-talk switch is pressed, the inputted voice is converted to infrared signals, which are transmitted to the DSP. The user pushes this switch for every utterance and continues pushing the switch while speaking until the user has finished talking. The reasons for using the press-talk switch are as follows: (1) to prevent recognition errors caused by noises and (2) to increase the battery life by reducing standby power.

In the DSP, acoustical processing, such as compensation for the frequency characteristic, noise reduction and TV sound cancellation, is executed. The signals, after acoustic processing in the DSP, are transmitted to the ASR engine. The ASR engine is executed in a microcomputer. Access control information and the remote-control signal from the ASR engine are sent to the remote-control pulse exchange unit. In both cases, using the ordinary remote control as well as the speech remote control, the remote-control code and access code, converted from their inputs, are sent to the remote-control processing part of the TV controlling microprocessor. The command from the remote-control processing part is sent to the application controlling part, which controls the TV operation. On the other hand, the ASR access code is forwarded to the voice-command processing part, from where a command is finally sent to the application controlling part. At the same time, notification information from the voice-command processing part is returned to the ASR engine.

For our TV product, since the restrictions on software and hardware are large, we have to realize signal processing and speech recognition using few resources. In our system, we use a 16 bit DSP to execute the signal processing and a 32 bit microcomputer (133 MIPS) to execute the speech recognition. Both devices are general-purpose products that have been introduced into various goods. In the ASR engine, it is important to realize using minimum resources. Our ASR system requires a memory of about 50 kBytes for the recognition program memory, about 330 kBytes for each acoustic model and 256 kBytes as working memory.

### 3.2. ASR Engine

The ASR system for the new DTV interface is our original speech recognition engine, COMPATS [3,4]. The specifications of the ASR engine are shown in Table 1. In our ASR system, a text-based word lexicon provides convenience in changing the word vocabulary for recognition. Regarding the recognition vocabulary, two or more words can be registered to one function or instruction. For example, up to six ways of paraphrasing one channel name are possible. The size of the vocabulary is about 400, and is about 1300 if it includes paraphrasing.

In addition to providing a user-friendly TV, the speech recognition system should have a good recognition performance for a variety of noises and speakers. First, to provid robustness to noises, we developed a noise

**Table 1** Specifications of ASR engine.

| | |
|---|---|
| Sampling freq. | 10 kHz |
| Window length | 30 [ms] Hamming |
| Window shift | 15 [ms] |
| Parameters in word matching | LPC-mel Cepstral parameters (7 dimensions), Normalized residual power, Delta log power |
| Sound unit | CV, VC |
| Acoustic models | 5–10 states, Single mixture, single covariance, age-dependent and gender-independent |

reduction technique for house noises and an echo-canceling technique for TV sound. Secondly, we developed new speaker adaptation techniques which are incorporated in the system to achieve robustness to speaker variations. Assuming that the speaking style is dependent on the speaker's generation, we developed age-dependent acoustic models (AD models). To normalize the characteristics of the vocal tract length (VTL) divergence, we employed a frequency-warping procedure. These of the AD model combined with the frequency-warping procedure results in a significant improvement in the recognition performance.

## 4. REDUCTION OF NOISE AND TV SOUND

The environment where the digital TV is used is difficult for speech recognition. In a house, there are various noises; for example, many appliances such as air conditioners and cleaners generate noise. On the other hand, the TV output includes sounds from the loudspeakers. Therefore, to achieve a good performance of speech recognition, we have developed the following: (1) a Winner filter for reduction of stationary noise and (2) an acoustic echo canceller to cancel of the TV sound. The diagram of the acoustical processing system is shown in Fig. 4. In the DSP, first, the noise is reduced. Secondly, the echo is cancelled.

### 4.1. Microphone for Speech Remote Control for DTV

In a system of speech recognition, it is important to correctly the kind of microphone and the position of the microphone. In particular, the relative position between the microphone and the mouth affects the recognition performance greatly, because it is a major factor that determines the SNR of an input sound. In an environment such as in a car, a speaker's (drive's) position is mostly fixed. Thus, the directional characteristics and sensitivity of the microphone can be adjusted appropriately, and surrounding noise can be estimated to some extent. However, for apparatus in the house, such as DTV, there are many variations of the relative position between the microphone and the mouth. Furthermore, depending on the position of the remote control and the direction of the face, the distance from the mouth to the microphone varies greatly. To stabilize the frequency characteristics and the sensitivity of sound, we chose an omnidirectional microphone for the system.

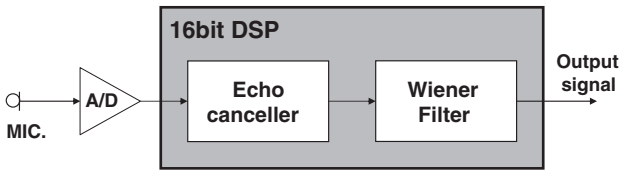


**Fig. 4** Diagram of the acoustical processing system in the DSP.

### 4.2. Reduction of House Noises

When using speech recognition in a real environment, it is necessary to obtain robustness to circumference noise. In a car, for example, the main noises are the sounds of the road and engine. The energy of such noises is mainly in the low-frequency band, and the spectrum of noises is also comparatively stationary. Therefore, it is possible to cope with the noises easily. However, in a house, in which we use apparatus such as DTV, various nonstationary noises occur. In our system, the Wiener filter, as shown in the following formula, is used to reduce the nonstationary noise.

$$y(t) = h(t) \otimes x(t) \tag{1}$$

$$H(f) = \frac{S(f) - \beta(SNR(f)) \cdot N(f)}{S(f)} \tag{2}$$

Where $x(t)$ and $y(t)$ are the signals before and after reduction, and $S(f)$ and $N(f)$ are the average power spectram of the voice and noise, respectively. The reduction coefficient $\beta$ for every filter bank is computed adaptively inaccordance with SNR, and then the transfer characteristics of the Wiener filter are controlled finely. $\beta$ is computed by the following method. First, the ratio of the minimum held value to the maximum held value in the amplitude of an input signal is calculated. Next, $\beta$ is computed by multiplying the ratio by the constant value. This process is based on the assumption that the amplitude of the noise corresponds to the minimum held value and the amplitude of the voice corresponds to the maximum held value. Consequently, the performance of noise reduction was increased. Table 2 shows the results of noise reduction using the Wiener filter. These experimental results show the reduction rate when white noise is outputted by a loudspeaker 1 m from the microphone. The speaker volume at the microphone is 67 dBA. When the reduction coefficient $\beta$ is fixed (before training), the reduction rate is only 4.8 dB. However, the reduction rate is 8 dB when $\beta$ is computed adaptively by training. The number of filter banks is limited to 2 bands in consideration of computational complexity. On the other hand, as a countermeasure against sudden noise, word spotting based on both-endpoint-free DP matching was used. In word spotting, detection of the speech is not necessary. Therefore, we can reduce the recognition error even when a sudden noise is inputted.

### 4.3. Cancellation of TV Sound

In an environment where digital TV is used, an

**Table 2** Results of noise reduction.

| | | |
|---|---|---|
| Reduction rate | Before training $\beta$ | 4.8 dB |
| | After training $\beta$ | 8.0 dB |

utterance of a user is mixed with the sound (TV sound) outputted from the loudspeaker of the TV. Therefore, TV sound becomes the major cause of recognition errors. In this system, this mixed TV sound is cancelled using echo-canceller technology based on the NLMS with step gain control. Echo-canceller technology has been introduced as a technology for hands-free calling in telephones and the teleconference systems. The TV sound can be cancelled using this technology, and the SNR of user speech is improved greatly.

In the DTV environment, the relative position between the remote control (microphone) and the TV changes frequently. Furthermore, the relative position between the remote control and the surrounding's reflective obstacles also change frequently. Therefore, it is necessary to stabilize and detect the double-talk state of the utterance of the user and the TV sound. Moreover, it is also necessary to increase the convergence speed of the adaptation filters.

In this system, a 1st adaptation filter for canceling TV sound, and a 2nd adaptation filter for rapidly detecting a double-talk state are introduced. Furthermore, the step gain of the 1st adaptation filter is controlled using the short-time input-and-output power ratio of the 1st adaptation filter to the 2nd adaptation filter [9]. Consequently, it becomes possible to improve the convergence speed. The updating method of the adaptation filter factor is shown in the formula

$$h(t+1) = h(t) + g(t) \cdot e(t) \cdot \frac{x(t)}{\|x(t)\|^2}, \qquad (3)$$

where $h(t+1)$ and $h(t)$ are the filter coefficients before and after cancellation, $g(t)$ is the step gain of the 1st adaptation filter, $e(t)$ is the error power of the 2nd adaptation filter and $x(t)$ is the power of the reference sound signal. The specifications of TV voice cancelling are shown in Table 3. This experimental result shows the cancellation rates when TV sounds are outputted by a loudspeaker 1 m from the microphone. The speaker volume at the microphone is 67 dBA. When the adaptation filter factor $h(t)$ is fixed (before adaptation), the cancellation rate is 12 dB. However, the cancellation rate is 20 dB when $h(t)$ is updated.

In the product, The power is switched off frequently to extend the battery life of the remote control. Therefore, a suitable filter factor is set up as an initial value in this system when the speech recognition system (power supply) is started.

**Table 3** Specifications of TV sound cancellation.

| | | |
|---|---|---|
| Cancellation time | | 100 ms |
| Cancellation rate | Before adaptation $h(t)$ | 12 dB |
| | After adaptation $h(t)$ | 20 dB |

## 5. SPEAKER ADAPTATION

Our new proposed adaptation method, which employs VTL normalization using age-dependent acoustic models, is expected to be very simple in practical use [10]. The method requires only the selection of the age-dependent acoustic model before it executes frequency warping according to the VTL estimated for each speaker. Therefore, this method requires a small amount of training utterance (only one word is enough) and is applicable to and suitable for the home application of ASR.

### 5.1. Age-Dependent Acoustic Models

Children and aged persons have distinct features in their pronunciation. As shown in Table 4, the characteristics of vocal organs differ among generations. These features are based on our observation of real speech data. Children are in the condition of articulation deficit because of their undeveloped organs. Aged persons are generally sluggish in their lip, tongue and teeth movement. In the case of aged persons, the unique features in their pronunciation are caused by the aging of their articulatory organs, loss of teeth, and so on. When listening to the utterances of children and aged persons, their pronunciation is slightly different from that of adult speakers (aged 20 to 60). These observed facts motivated us to use age-dependent acoustic models. Here, each age-dependent acoustic model is a gender-independent model with a single-mixture Gaussian distribution.

### 5.2. VTLN by Frequency Warping

To normalize VTL divergence caused by frequency warping, we employ the following Oppenheim's frequency-warping function [10,11] to obtain the frequency-warping coefficient (FWC). We estimate the FWC of a speaker by the following process. First, we ask the user to utter one indicated word. Then, using Eq. (4) we estimate the optimal FWC under a supervised condition using the indicated word, and cepstral domain feature parameters are frequency-warped using Oppenheim's method.

**Table 4** Vocal organ characteristics of each generation.

| Generation | Characteristics |
|---|---|
| Children (School children) | Undeveloped deficit in articulation organs |
| Aged persons | Sluggish movement of lips, tongue, and teeth |
| Adult | VTL is the predominant factor of interspeaker difference |

**Table 5** Effects of age-dependent acoustic models and frequency warping.

| | Category of test speakers | Children | Adult | Aged persons |
|---|---|---|---|---|
| | Age | 7–12 | 18–60 | 66–95 |
| | # of speakers | 25 | 50 | 30 |
| WER [%] | Baseline (without FW) | 11.0 | 4.0 | 8.4 |
| | Adult model with FW | 3.8 | 3.3 | 7.0 |
| | Age-dependent model with FW | 3.0 | 3.3 | 4.1 |

$$\hat{\alpha} = \arg\max_{\alpha} P(X^{\alpha} \mid \alpha, \theta) \tag{4}$$

In our recognition experiments, the complete training speech was frequency-warped using a default FWC. However, note that linear-cepstrum coefficients can be interpreted as mel-cepstrum coefficients using the same procedure as that used to normalize the VTL, at a 10 kHz sampling frequency.

### 5.3. Experiment on Age-Dependent Acoustic Models and Frequency Warping

The test conditions are shown in Table 1. The test vocabulary is a set of 100 phone-balanced city names of Japan. In the preparation for the test, a selected age-dependent acoustic model is loaded in the system, and the FWC is fixed in an effort to maximize the likelihood of an indicated one-word utterance and selected acoustic model. Some of our simulation test results are shown in Table 5. The test speakers consist of 25 primary school children, 50 male and female adults, and 30 male and female aged persons. All results are for word error rate (WER). The first results shown on the top of the table are the baseline results without using any speaker adaptation procedure and using an adult acoustic model. The next results are for frequency warping using an adult acoustic model. The last results are for our proposed method using an age-dependent acoustic model combined with frequency warping. These results show that frequency warping is effective for children. The combination of the age-dependent acoustic model and frequency warping is effective for both aged persons and children.

### 5.4. Speaker-Specific Templates for Unique Utterances

There are some unique utterances that are difficult to recognize even if we use the AD model combined with frequency warping. For such cases, additional trained speaker-dependent templates will be effective, as already reported [12]. We propose to add speaker-specific templates of the utterances themselves, and lexical templates of the words obtained using the original speaker-independent model. Note that the speaker-dependent template is trained using the series of phonetic units of the speaker-independent acoustic model.

**Table 6** Recognition performance utilizing all adaptation techniques.

| | Adapted | Baseline |
|---|---|---|
| WER [%] | 4.4 | 11.7 |

### 5.5. Recognition Performance Utilizing all Adaptation Techniques

We tested the recognition performance of the DTV prototype utilizing all the adaptation techniques described above, i.e., using AD models combined with frequency warping and using speaker-specific templates, when necessary. The test speakers were 6 adult males and 4 adult females. For the test, 100 control words were selected from the lexicon of the DTV speech recognition system. The results are shown in Table 6. The baseline average recognition WER for the 10 speakers was slightly more than 10%. Employing speaker adaptation techniques, the average WER was reduced to less than 5%, which we believe is a practical performance level.

## 6. CONCLUSION

We developed a speech recognition interface system for a DTV remote control and released a commercial product on the Japanese market in December 2001, the world's first commercial product of this type. After the product release, we evaluated the usability of this new user interface. In an experiment, a user operates functions for which the operating frequency is high, for example, channel tuning, program selection using shortcuts, program reservation, HDD video recording, and the speaker adaptation operation. We obtained the degree of successful task achievement and we interviewed the user. The feedback was classified into good points and points to be improved. The good points were as follows:

(1) The voice control is easier than button operation.
(2) Since the shortcut using speech recognition can skip procedures, it is very convenient.
(3) It is possible to operate the remote control without looking at a manual.

On the other hand, points to be improved were as follows:

(1) There is no information on the acceptable utterances.
(2) It is unclear what can be controlled by speech.
(3) The user cannot find the word corresponding to the function.

Thus, our next development target is to address the issues raised by these comments and to develop a TV with a speech dialogue interface.

In the future, we have to address the following several issues to extend speech recognition technology to a wide range of products and services.

(1) Robustness in the real environment
- Separation of speech and noise (including synthetic sounds)
- Rejection of neighbor's voice
- Separation of simultaneous utterances by two or more users

(2) Robustness to various users
- Regionality (dialect)
- Various expressions

(3) Dialog management for complicated use
- Nonregistered vocabulary
- User modeling

(4) Large-scale corpus (sound, text)
- Corpus collected in the environment according to application

On the other hand, it is also important that we continue appealing to users by producing the suitable applications.

## REFERENCES

[1] S. Hiraoka, K. Nomura, H. Kuwano, A. Ookumo, T. Watanabe, K. Niyada, E. Shuuda and Y. Nomura, "Evaluation of the speech recognizer for a voice programming VCR remote-controller," *Tech. Rep. IEICE*, SP91-8, pp. 17–24 (1991).

[2] M. Hoshimi, M. Yamada and K. Niyada, "A practical speech recognition method for unspecified speakers on a single DSP chip," *IEICE Trans.*, **J72-D-II**, 2096–2103 (1996).

[3] T. Kimura, H. Kuwano, A. Ishida, T. Watanabe and S. Hiraoka, "Compact-size speaker independent speech recognizer for large vocabulary using "COMPATS" method," *Proc. ICSLP 94*, pp. 1379–1382 (1994).

[4] T. Kimura, M. Endo, S. Hiraoka and K. Niyada, "Speaker independent word recognition using continuous matching of parameters in time-spectral form based on statistical measure," *Proc. ICSLP 92*, pp. 169–172 (1992).

[5] K. Fujita, H. Kuwano, T. Tsuzuki, Y. Ono and T. Ishihara, "A new digital TV interface employing speech recognition," *IEEE Trans. Consum. Electron.*, **49**, 765–769 (2003).

[6] AURORA http://portal.etsi.org/stq/kta/DSR/dsr.asp

[7] R. Nisimura, Y. Nishihara, R. Tsurumi, A. Lee, H. Saruwatari and K. Shikano, "Takemaru-kun: Speech oriented information system for real world research platform," *International Workshop on Language Understanding and Agents for Real World Interaction*, pp. 70–78 (2003).

[8] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," *Proc. ICASSP 03*, Vol. 2, pp. 137–140 (2003).

[9] H. Furukawa, J. Tagawa, T. Kanamori, S. Ibaraki and K. Kitajima, "Step gain control for acoustic echo cancellers," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 477–478 (1993).

[10] K. Fujita, Y. Ono and Y. Nakatoh, "A study of vocal tract length normalization with generation-dependent acoustic models," *Proc. ICSLP 2000*, pp. 706–709 (2000).

[11] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," *Proc. IEEE*, **60**, 681–691 (1972).

[12] J. Neena, C. Ronald and B. Etienne, "Creating speaker-specific phonetic templates with a speaker-independent phonetic recognizer: Implications for voice dialing," *Proc. ICASSP 96*, Vol. 2, pp. 881–884 (1996).