We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

6,900
Open access books available

186,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Voice Activated Appliances for Severely Disabled Persons

Soo-young Suk and Hiroaki Kojima
Advanced Industrial Science and Technology
Japan

1. Introduction

People with severe speech and motor impairment due to cerebral palsy are great difficult to move independently and also cannot control home electric devices. Computer has much to offer people with disability, but the standard human-machine interface (e.g. keyboard and mouse) is inaccessible to this population. In this chapter, we describe a speech recognition interface for the control of powered wheelchair and home automation systems via severely disabled person's voices. In particular, we consider that our system can be operated by inarticulate speech produced by persons with severe cerebral palsy or quadriplegia in real-environment.

The aim of our research is divided two targets. One is easy to control of various home appliances by voice, and the other is to enable severely disabled person's movement independently using voice activated powered wheelchair. At first, Home automation system product for intelligent home is increasingly getting very common by the help of intelligent home technologies that increased easy, safety and comfort. Moreover, home automation is an absolute benefit and can improve the quality of life for the user. Home automation houses have been developed to apply new technologies in real environment, such as Welfare Techno Houses (Tamura et al., 2007), Intelligent Sweet Home (Park et al., 2007), Smart House (West et al., 2005). Interfaces based on gestures or voices have been widely used for home automation. However, gesture recognition based on vision technologies depends critically on the external illumination conditions. And gesture recognition is difficult or impossible for people suffering from severe motor impairments, such as paraplegia and tremors. Recently, a voice-activated system using commercial voice-recognition hardware in a low-noise environment has been developed for disabled persons capable of clear speech (Ding & Cooper, 2005).

The next, powered wheelchairs provide unique mobility for the disabled and elderly with motor impairments. Sometimes, the joystick is a useless manipulation tool because the severely disabled cannot operate it smoothly. Using natural voice commands, like ``move forward'' or ``move left'' relieves the user from precise motion control of the wheelchair. Voice activated powered wheelchair is required safety manipulation with high speech recognition accuracy because the accident can occur by a misrecognition. Although current speech recognition technology has reported high performance, it is not sufficient for safe voice-controlled powered wheelchair movement by inarticulate speech affected by severe

cerebral palsy or quadriplegia, for instance. To cope with the pronunciation variation of inarticulate speech, we adopted a lexicon building approach based on Hidden Markov Model and data mining (Sadohara at al., 2005), in addition to acoustic-modeling-based speaker adaptation (Suk at al., 2005). We also developed noise-canceling methods, which reduce mechanical noise and environmental sounds for practical use on the street (Sasou at al., 2004). However, though our voice command system has improved recognition performance by various methods, the system requires a guarantee of safety for wheelchair users in two additional conditions.

- To move only in response to the disabled person's own voice.
- To reject non-voice command input.

The first problem is to prevent operation of the wheelchair by unauthorized persons near the disabled user. A speaker verification method can be applied to solve this problem, but it is difficult to verify when using short word commands. Therefore, we are now developing a speaker position detection system using a microphone array (Jonson at al., 1993; Sasou & Kojima, 2006). The second problem is that a lot of other noise is input when the voice command system is being used. Also, a voice-activated control system must therefore reject noise and non-voice commands such as coughing and breathing, and spark-like mechanical noise in the preprocessing stage. A general rejection method has achieved a confidence measure using a likelihood ratio in a post-processing step. However, this confidence measure is hard to use as a non-command rejection method because of the inaccuracy of likelihood when speech recognition deals with unclear voice and non-voice sounds. Thus, a non-voice rejection algorithm that classifies Voice/Non-Voice (V/NV) in a Voice Activity Detection (VAD) step is useful for realizing a highly reliable voice-activated powered wheelchair system.

The chapter first presents the F_0 estimator and the non-voice rejection algorithm. Next, the inarticulate speech recognition is described in Section 3. In Section 4, we present a developed voice activated control system. And we evaluate the performance of our system in Section 5. Lastly, we offer our conclusions in Section 6.

2. Non-voice rejection using V/NV classification

The general VAD uses short time energy and/or ZCR for start and end point detection in a real-time voice command system with low complexity. However, VAD has a problem because various sounds are determined as voice sounds. Previous V/NV classification algorithms have generally adopted statistical analyses of F_0, the Zero-Crossing Rate (ZCR), and the energy of short-time segments. A method for voicing decision within a pitch-detection algorithm is presented in (Rouat at al., 1997). A combination of these methods, a cepstrum-based F_0 extractor, has been proposed (Ahmadi & Andreas, 1999). An auditory-based method for voicing decision within a pitch-tracking algorithm appears in (Mousset at al., 1996). For the purpose of non-voice rejection, we propose a V/NV classification using a reliable F_0 estimator.

2.1 YIN: fundamental frequency estimator

V/NV classification using F_0 information has been strongly tied to the problem of a pitch detection algorithm (PDA). A PDA can be formulated as an average magnitude difference function, average squared difference function, or similar autocorrelation methods in the time domain. In addition, cepstrum analysis is possible in the frequency domain by applying

the harmonic product spectrum algorithm. Among these F_0 extraction methods, we use the well known auto-correlation method based on YIN that has a number of modifications to reduce estimation errors (de Cheveigné, 2002). This method has the merit of not requiring fine tuning and uses fewer parameters. The name YIN (from "Yin" and "Yang" of oriental philosophy) alludes to the interplay between autocorrelation and the cancellation that it involves. The autocorrelation function of a discrete signal x_t may be defined as

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau} \tag{1}$$

where $r_t(\tau)$ is the autocorrelation function of lag τ at time index t, and W is the integration window size. YIN achieves a difference function instead of an autocorrelation function that is influenced in bias value.

$$d_t(\tau) = \sum_{j=t-\tau-W/2}^{t=\tau/2+W/2} (x_j - x_{j+\tau})^2 \tag{2}$$

Here, $d_t(\tau)$ is the difference function to search for the values of τ for which the function is zero. The window size shrinks with increasing values of τ, resulting in the envelope of the function decreasing as a function of lag as illustrated in Fig. 1(a). The difference function must choose a minimum dip that is not zero-lag. However, setting the search range is difficult because of imperfect periodicity. To solve this problem, the YIN method replaces the difference function with the cumulative mean normalized difference function of Eq. (3). This function is illustrated in Fig. 1(b).

$$d'_t(\tau) = \begin{cases} 1, & if \ \tau = 0, \\ d_t(\tau) \Big/ \left[(1/\tau) \sum_{j=1}^{\tau} d_t(j) \right] & otherwise \end{cases} \tag{3}$$

The cumulative mean normalized difference function not only reduces "too high" errors, but also eliminates the limit of the frequency search range, and no longer needs to avoid the zero-lag dip. One of the higher-order dips appears often in F_0 extraction, even when using
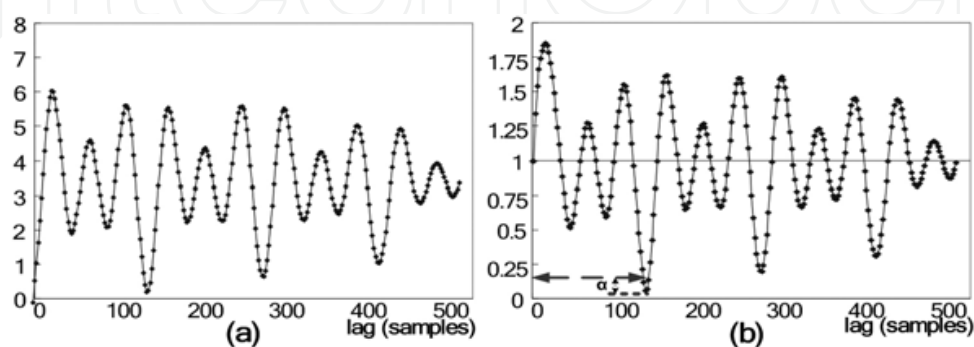

Fig. 1. (a) Example of difference function (b) Cumulative mean normalized difference function at same waveform

the modified function in Eq. (3). This error is called the sub-harmonic or octave error. To reduce the sub-harmonic error, the YIN method finds the smallest value of τ that gives a minimum of $d_t'(\tau)$ deeper than the threshold. Here, the threshold is decided by the value that adds a minimum of $d_t'(\tau)$ to the absolute threshold α in Fig. 1 (b). Absolute threshold is possible because of the achieved normalized processing in the previous step. In the final step, F_0 is extracted through the parabolic interpolation and best local estimation process.

2.2 V/NV classification

The general V/NV classification algorithm participates in the processing of each short-time speech segment. However, classification of a whole input segment is more important in reliable speech recognition in which non-voice rejection is possible. For this classification, the proposed algorithm decides V/NV from the ratio of the reliable F_0 contour over the whole input interval.

The function value $d_t'(\tau)$ defined by Eq. (3) is compared with the confidence threshold to decide the reliability of each F_0 frame. Here, the confidence threshold is selected such that the value is 0.05 to 0.2. Figures 2 and 3 depict examples of reliable F_0 contour extraction. A reliable F_0 contour using the cumulative mean normalized difference function is illustrated in Fig. 2 (b). When the confidence threshold of YIN-based F_0 is 0.1, only high reliability areas are selected, as illustrated in Fig. 2 (c).
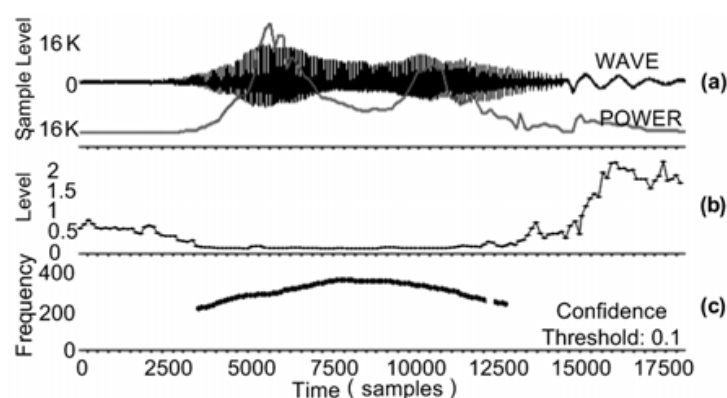

Fig. 2. (a) Example of a voice waveform (b) Cumulative mean normalized difference function calculated from the waveform in (a) (c) Reliable F0 contour in which the confidence threshold is applied

The conventional VAD method using energy and/or ZCR is detected noise as well as voice in Fig. 3 (a). However, you can see that reliable F_0 appears on only three frames because of the applied confidence threshold 0.1 in Fig. 3 (c). Furthermore, we can prove the performance by the examining that detected frequency is the inner voice frequency area. For V/NV classification from the extracted F_0 contour, we then compute the ratio of frames with the reliable F_0 as follows.

$$d = \frac{1}{M} \sum_{i=1}^{M} P_{th}(i) \tag{4}$$

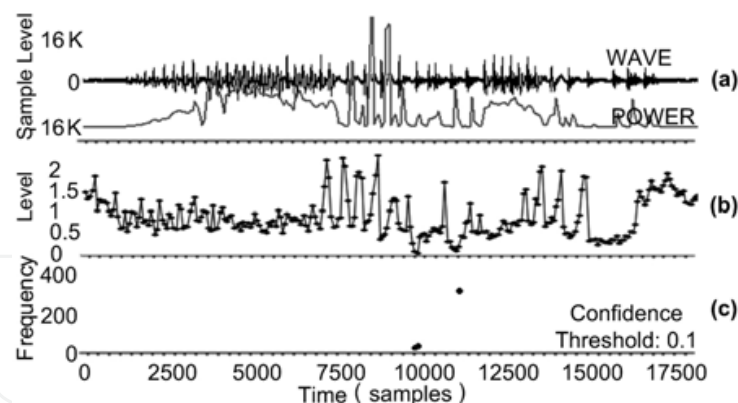$$P_{th}(i) = \begin{cases} 1 & if \ F_{\min} \le F_{oth} \le F_{\max} \\ 0 & otherwise \end{cases} \tag{5}$$

Fig. 3. (a) Example of a noise waveform (b) Cumulative mean normalized difference function calculated from the waveform in (a) (c) Reliable F0 contour where the confidence threshold is applied

Here, M indicates the total number of input frames, and F_{min}=60Hz and F_{max}=800Hz are experimentally chosen for a disabled person's voice. Finally, an input segment is classified as voice if d exceeds the V/NV threshold value. The cepstrum-based algorithm can also be used as the confidence threshold for extraction of the F_0 contour as indicated in Fig. 4. However, the F_0 extraction performance of a cepstrum-based algorithm is inferior to YIN, and it is difficult to determine a suitable threshold in various environments.
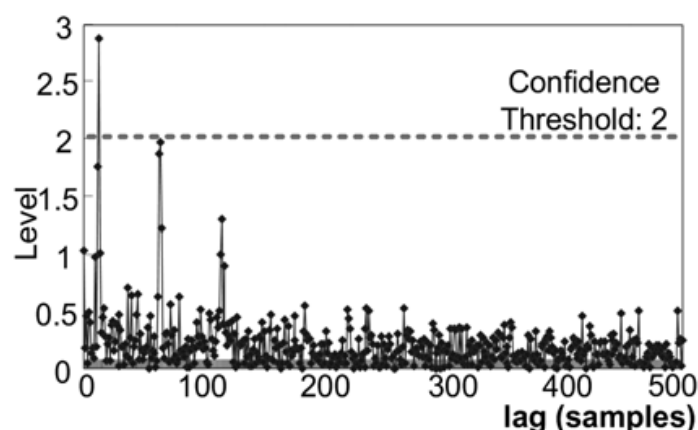
Fig. 4. Example of cepstrum signal to the applied confidence threshold

3. Inarticulate speech recognition

The severely disabled person has the problem of pronouncing even simple voice commands. Specially, our system is required high performance of speech recognition when controlling a powered wheelchair, because recognition performance is related to the disabled person's safety. Using the speech characteristics of each severely disabled person, the voice command input system needs to select words in which an utterance can be spoken easily and distinctly. Therefore, our system selected five of 12 word candidates using a test. However, as the disabled person has difficulty in speaking the command clearly, the utterance of "hi da ri" may be spoken as "hi hi hi da ri", "i a ri" etc., in spite of the selected voice command. For this reason, a single-template dictionary is unable to recognize unusual speech.

A multi-template dictionary was generated through analysis of the speech patterns of the disabled to solve this problem. To generate a multi-template dictionary, the initial dictionary

is prepared from the result of a phoneme-recognition experiment using a pre-recorded voice. The final dictionary is then generated by deleting unwanted candidates through a repetitive word-recognition experiment. The generated multi-template dictionary was comprised of the 27 templates of the "hidari" utterance, 15 templates for "migi," 13 templates for "mae," 4 templates for "koutai," and 5 templates for "ah."

Action	Command	Dictionary
Move Left	hidari hidari hidari	h i d a: r i: d a r i q h i h i i d a: r i ,…other 24
Move Right	migi	m i g i i m i z i p ,… other 13
Move Forward	mae	m a: a e m a e p i ,… other 11
Move Backward	koutai	k o u t a i ,… other 3
Stop	ah	a: ,… other 4

Table 1. Implemented set and multiple dictionaries for inarticulate speech recognition

Since speech recognition systems are known to demonstrate different results from reading speech and spontaneous speech, it is important for evaluation and modeling of voice command system. To obtain a sample of spontaneous speech affected by disability, which contains specific personal variations, we developed a voice operated toy robot and a graphical simulation demo system that uses the same recognition task such as in powered wheelchair operation.

For analysis of the input devices, each input device achieved speech detection through each recognition engine at the same time. Currently, we have collected more than 3000 unclear samples of speech affected by disability, using four types of microphones: headset(Audio-technica: AT810X), Bone conduction(Sony: ECM-TL1), Pin(PAVEC: MC-105), Bluetooth (Sonorix: OBH-0100). Transmission capacity of bluetooth microphone is limited by an 8 KHz sampling rate. Table 2 lists the recorded speech data used for the experimental evaluation. For the headset type, 579 inputs are collected. There are also 426 voice commands, 65 various noises, 76 other utterances that are not commands, and utterances of 12 another people. Therefore, V/NV classification is needed to satisfy voice activated powered wheelchair control requirements while maintaining high speech recognition accuracy.

Fig. 5. Speech recording environment (a) Voice operated toy robot system (b) Graphical simulation demo system

	Voice command	Noise	Other speech	Other people
Headset	426	65	76	12
Bone conduction	405	339	88	286
Pin	399	21	90	361
Bluetooth	337	22	64	62
Total	1567	447	318	721

Table 2. Analysis of the number of recorded data

4. Voice activated control system

The voice activated home appliances control system diagram is shown in Figure 6. In the diagram, speech input device is can be use not only headset but also mobile telephone via Skype Voice of IP (VOIP) module. At First, the microphone captures the speech signal. The incoming audio stream is then segmented for recognition using the VAD module. The stream is transmitted from speech interface to the recognition engine where the recognition procedure is carried out. Our recognition engine employed a Julian decoder (Lee at al., 2001) with mel frequency cepstral coefficient and adapted a speaker-dependent acoustic model. Finally the recognition result is executed when results are satisfied with voice using V/NV classification module.

By using the result of speech recognition for infrared remote control, system can be remote control powered wheelchair and home appliances including TV, radio, VCD/DVD player, lights, and fan. System is also offers hand-free management of telephone calls and direct calling to family through one voice command. Software is designed and developed under the visual studio platform and using visual C++ programming, which can be installed and run in Ultra Mobile PC (UMPC) under operating system of embedded windows.
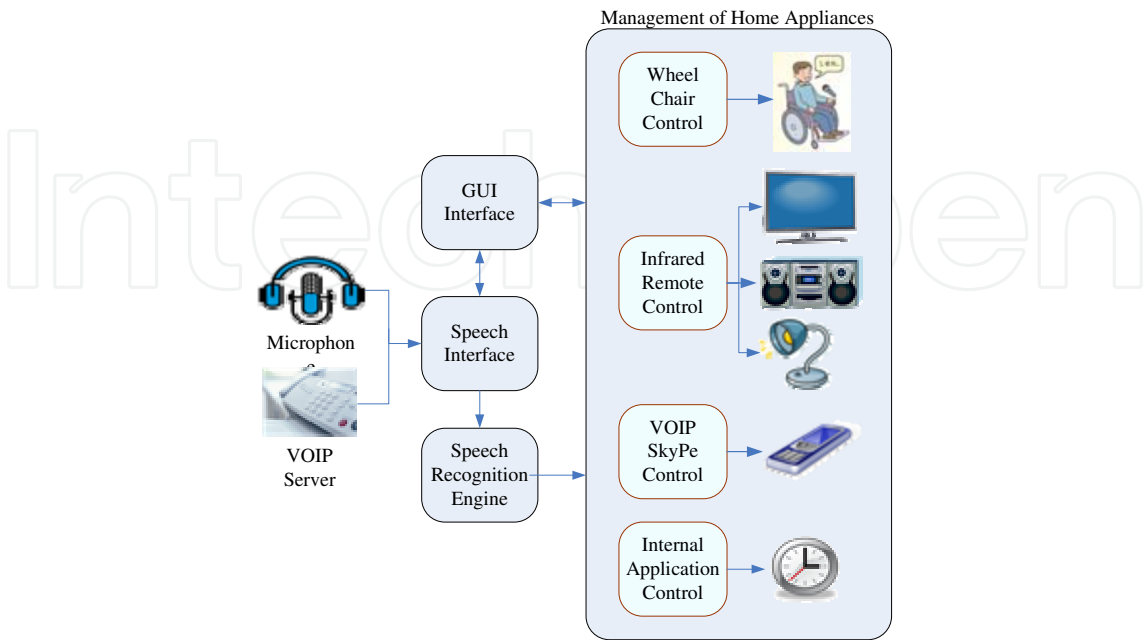

Fig. 6. Voice activated home appliance control system design

A voice controlled Graphic User Interface (GUI) is carefully designed for disabled person in Fig. 7. Click of the icon on the user's screen or voice command directly correspond to environmental commands (switching on the lamps, starting the Radio, calling the facilitator)

(a) Powered Wheelchair Mode   (b) Home Appliances Control Mode

(c) Sub Control Mode (TV, Audio, Phone)

Fig. 7. Example of system GUI design (a) Powered wheel chair mode (b) Home appliances control mode (c) Appliance control sub mode

The developed system consists of a headset, a Pentium M 1.2GHz UMPC, infrared transmitter for long distance transmission and a wheelchair controller, as depicted in Fig. 8. Also, wireless microphone or mobile phone can be used instead of wire headsets for user convenience.

(a) Prototype powered wheelchair

(b) infrared transmitter
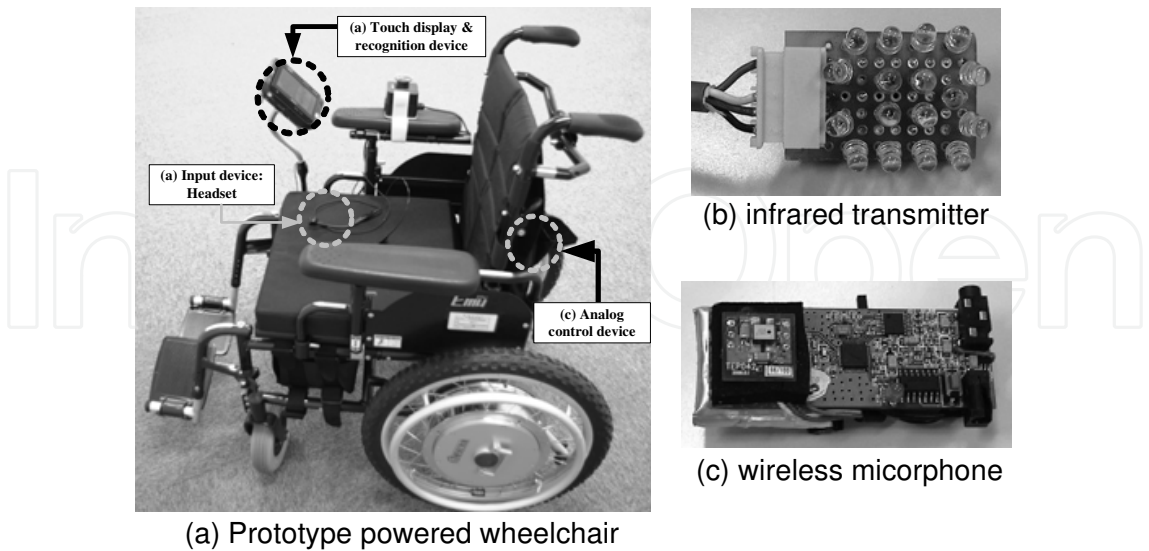
(c) wireless micorphone

Fig. 8. (a) Developed prototype wheelchair (b) Infrared transmitter (c) Wireless microphone

The voice commands to control wheelchair direction are not easy when use the minimum number of commands. So, our system use the state transition diagram for more free movement of voice activated powered wheelchair, as shown in Fig. 9.
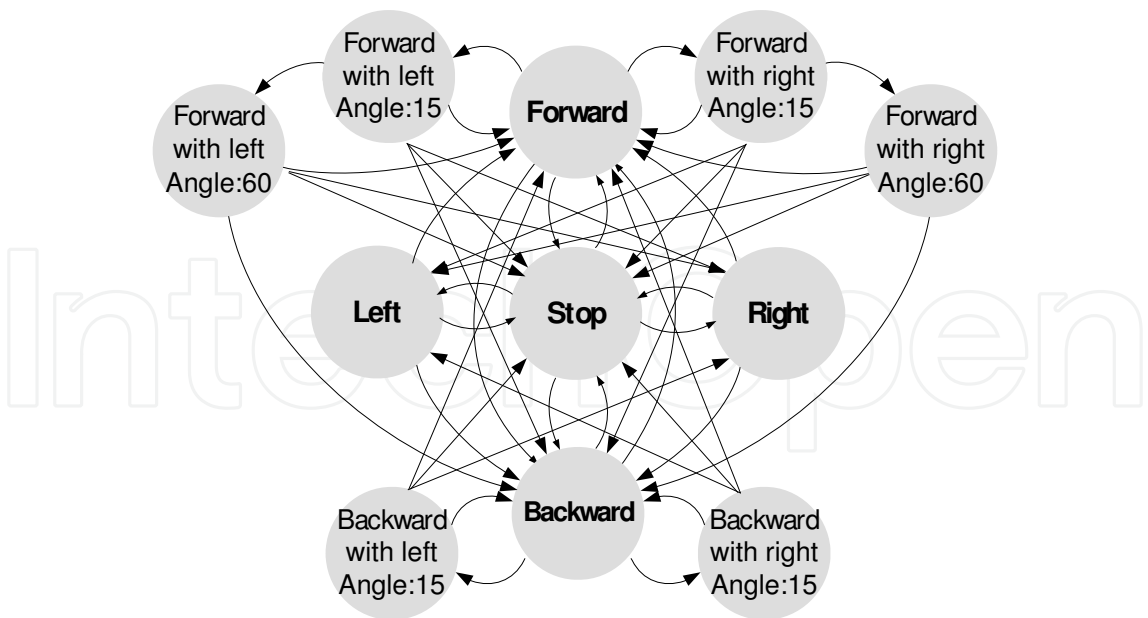
Fig. 9. State transition diagram by voice command

5. Experiment results

To evaluate the performance of our proposed method, we conducted V/NV classification experiments using the 1567 voice commands and 447 noises and employing YIN-based and cepstrum-based algorithms. The sampling frequency was 16kHz, the window size was 25ms, and the frame shift was 8ms.

Figure 10 depicts the V/NV classification performance and plots the recall-precision curves according to an individual confidence threshold. The results indicate that the YIN-based algorithm is superior to the cepstrum-based algorithm. When the confidence threshold of YIN is 0.08, the V/NV classification provides the best results with 0.97 and 0.99 rates for recall and precision. In other words, when the lowest threshold was selected for voice detection at a precision rate of 1, the miss-error rate of noise was only 4.9%.
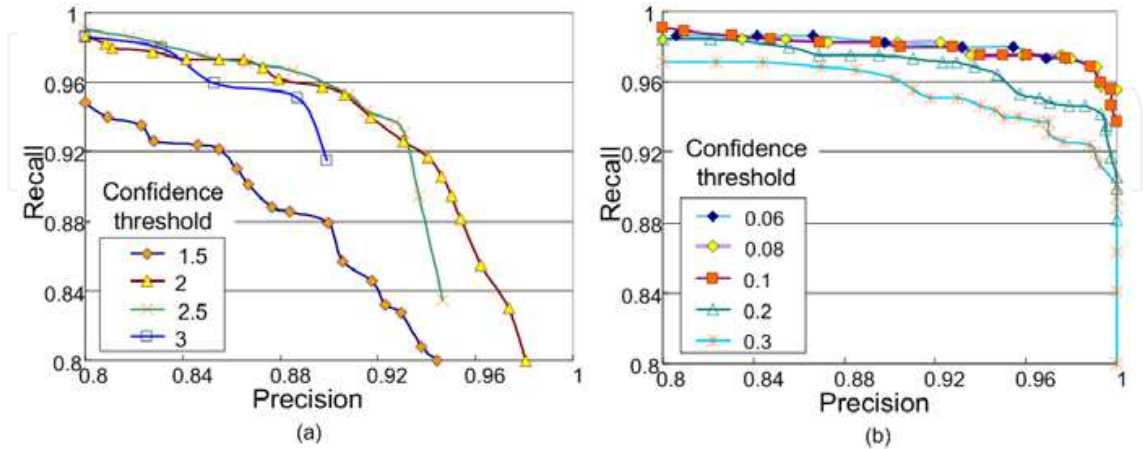

Fig. 10. Recall-precision curve of noise classification a) cepstrum b) YIN

Table 3 lists the best confidence threshold of each microphone with the best recall precision. When YIN uses the F_0 extraction method, the confidence threshold is stable at about 0.08.

Although the cepstrum algorithm can use the F_0 extraction method, it is difficult to decide on a suitable confidence threshold in each microphone environment.

	Headset	Bone conduction	Pin	Bluetooth
Cepstrum	3	2.5	1.5	2
YIN	0.05~0.1	0.06~0.08	0.07~0.1	0.08~0.1

Table 3. Confidence threshold analysis of four types of microphones with the best recall precision

A recognition experiment was performed in order to confirm the validity of the multi-candidate recognition dictionary and the adapted acoustic model for the basic performance of a voice-activated system. The recognition experiment used 2211 data elements recorded at an athletic meeting and outdoors with a noise background to evaluate the effectiveness of the proposed multi-template dictionary and adapted acoustic model. Ninety-six utterances were used for adaptation of the acoustic model, and the remaining 1334 utterances were used for evaluation.

Mix.	Baseline	Multi-Tem.	Adapt.	Multi-Tem. & Adapt
1	61.4	94.2	97.8	98.4
2	78.4	95.5	98.8	99.1
4	77.9	94.6	98.7	99.2
8	80.4	94.3	98.8	99.3
16	78.6	93.8	98.6	99.5
32	75.1	91.2	98.4	99.4

Table 4. Speech recognition accuracy with 2000-state HMnet model

The speaker-independent, 2000-state 16-mixture HMnet model was evaluated as the baseline. An average recognition rate of 78.6 was achieved, although there were five words in the dictionary because it did not consider the speech characteristics and variations of disabled persons. The average recognition rate was improved to 93.8% by applying the multi-template dictionary. The acoustic model that performed MAP adaptation achieved an average recognition rate of 98.6%. The average recognition rate was improved to 99.5% by applying the multi-template dictionary with the adapted acoustic model.

6. Conclusion

This chapter presented home appliances control system for independent life of the severely disabled person. In particular, the developed system can be operated by inarticulate speech and a non-voice rejection method for reliable VAD in a real environment with extraneous sounds such as coughing and breathing. The method classifies V/NV from the ratio of reliable F_0 contour over the whole input interval. We adopted the F_0 extraction method where YIN has the best performance among conventional methods. Our experiment results indicate that the false alarm rate is 4.9% with no miss-errors in which voice is determined to

be non-voice. And the average recognition rate was improved to 99.5% by applying the multi-template dictionary with the adapted acoustic model. Therefore, the speaker dependent acoustic model, dictionary and non-voice rejection algorithm can be helpful for realizing a highly reliable wheelchair control system.

7. Acknowledgement

8. References

Ahmadi, S. & Andreas S. S. (1999). Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. *IEEE Trans. Speech Audio Processing*, Vol. 7, No. 3, pp. 333-339.

de Cheveigné, A. & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustic Society of the America*, Vol. 111, pp. 1917-1930.

Ding, D. & Cooper, R.A. (2005). Electric powered wheelchairs. *IEEE Trans. Control Syst. Mag.*, Vol. 25, pp. 22–34.

Jonson, D. H. & Dudgeon, D.E. (1993). Array signal processing. *Prentice Hall*, Englewood Cliffs, NJ.

Lee, A.; Kawahara, T. & Shikano, K. (2001). Julius—an open source realtime large vocabulary recognition engine. *Proceeding of European Conference Speech Communication Technology*, pp. 1691–1694.

Mousset E., Ainsworth, W. A. & Fonollosa, J. A. R. (1996). A comparison of several recent methods of fundamental frequency and voicing decision estimation. *Proceeding of International Conference of Spoken Language Processing*, Vol. 2, pp. 1273–1276.

Park, K.; Bien, Z.; Lee. J.; Kim, B.; Lim, J.; Kim, J.; Lee, H.; Stefanov, D.H.; Kim, D.; Jung, J.; Do, J.; Seo, K.; Kim, C.; Song, W. & Lee, W. (2007). Robotic smart house to assist people with movement disabilities. *The Journal of Autonomous Robots*, Vol. 22, No. 2, pp. 183-198.

Rouat, J.; Liu, Y. C. & Morrisette, D. A. (1997). pitch determination and voiced/unvoiced decision algorithm for noisy speech. *The Journal of the Speech Communication*, Vol. 21.

Sadohara, K.; Lee, S.W. & Kojima, H. (2005). Topic Segmentation Using Kernel Principal Component Analysis for Sub-Phonetic Segments. *Technical Report of IEICE*, AI2004-77, pp. 37-41.

Sasou, A.; Asano, F.; Tanaka, K. & Nakamura, S. (2004). HMM-Based Feature Compensation Method: An Evaluation Using the AURORA2. *Proceeding International Conference Spoken Language Processing*, pp. 121-124.

Sasou, A. & Kojima, H. (2006). Multi-channel speech input system for a wheelchair. *Proceeding Mar Meeting of the Acoustical Society of Japan*, Vol 2006.

Suk, S.Y.; Lee, S.W; Kojima, H. & Makino, S. (2005). Multi-mixture based PDT-SSS Algorithm for Extension of HM-Net Structure. *Proceeding of September Meeting of the Acoustical Society of Japan*, Vol 2005, pp. 1-P-8.

Tamura, T.; Kawarada, A.; Nambu, M.; Tsukada, A.; Sasaki, K. & Yamakoshi, K. (2007). E-Healthcare at an Experimental Welfare Techno House in Japan. *The journal of Open Medical Informatics,*Vol. 1, No. 1, pp. 1-7.

West, G.; Newman, C. & Greenhill, S. (2005). *Using a camera to implement virtual sensors in a smart house.*, Smart Homes to Smart Care. IOS Press, pp. 83-90.

Speech Recognition

Edited by France Mihelic and Janez Zibert

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Soo-young Suk and Hiroaki Kojima (2008). Voice Activated Appliances for Severely Disabled Persons, Speech Recognition, France Mihelic and Janez Zibert (Ed.), ISBN: 978-953-7619-29-9, InTech, Available from: http://www.intechopen.com/books/speech_recognition/voice_activated_appliances_for_severely_disabled_per sons

INTECH
open science | open minds