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

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

BOOK CITATION INDEX
CLARIVATE ANALYTICS
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

System Request Utterance Detection Based on Acoustic and Linguistic Features

T. Takiguchi, A. Sako, T. Yamagata and Y. Ariki
Kobe University
Japan

1. Introduction

Robots are now being designed to become a part of the lives of ordinary people in social and home environments, such as a service robot at the office, or a robot serving people at a party (H. G. Okuno, et al., 2002) (J. Miura, et al., 2003). One of the key issues for practical use is the development of technologies that allow for user-friendly interfaces. This is because many robots that will be designed to serve people in living rooms or party rooms will be operated by non-expert users, who might not even be capable of operating a computer keyboard. Much research has also been done on the issues of human-robot interaction. For example, in (S. Waldherr, et al., 2000), the gesture interface has been described for the control of a mobile robot, where a camera is used to track a person, and gestures involving arm motions are recognized and used in operating the mobile robot.

Speech recognition is one of our most effective communication tools when it comes to a hands-free (human-robot) interface. Most current speech recognition systems are capable of achieving good performance in clean acoustic environments. However, these systems require the user to turn the microphone on/off to capture voices only. Also, in hands-free environments, degradation in speech recognition performance increases significantly because the speech signal may be corrupted by a wide variety of sources, including background noise and reverberation. In order to achieve highly effective speech recognition, in (H. Asoh, et al., 1999), a spoken dialog interface of a mobile robot was introduced, where a microphone array system is used.

In actual noisy environments, a robust voice detection algorithm plays an especially important role in speech recognition, and so on because there is a wide variety of sound sources in our daily life, and because the mobile robot is requested to extract only the object signal from all kinds of sounds, including background noise. Most conventional systems use an energy- and zero-crossing-based voice detection system (R. Stiefelhagen, et al., 2004). However, the noise-power-based method causes degradation of the detection performance in actual noisy environments. In (T. Takiguchi, et al., 2007), a robust speech/non-speech detection algorithm using AdaBoost, which can achieve extremely high detection rates, has been described.

Also, for a hands-free speech interface, it is important to detect commands in spontaneous utterances. Most current speech recognition systems are not capable of discriminating system requests - utterances that users talk to a system - from human-human conversations.

Therefore, a speech interface today requires a physical button which on and off the microphone input. If there is no button for a speech interface, all conversations are recognized as commands for the system. The button spoils the merit of speech interfaces that users do not need to operate by the hand. Concerning this issue, there are researches on discriminating system requests from human-human conversation by acoustic features calculated from each utterance (S. Yamada, et al., 2005). And also, there are discrimination techniques using linguistic features. Keyword or key-phrase spotting based methods (T. Kawahara, et al., 1998) (P. Jeanrenaud, et al., 1994) have been proposed. However, using keyword spotting based method, it is difficult to distinguish system requests from explanations of system usage. It becomes a problem when both utterances contain a same "keywords." For example, the request speech is "come here" and the explanation speech is "if you say come here, the robot will come here." In addition, it costs to construct a network grammar to accept flexible expressions.

In this chapter, an advanced method of discrimination using acoustic features or linguistic features is described. The difference of system requests and spontaneous utterances usually appears on the head and the tail of the utterance (T. Yamagata, et al., 2007). By separating the utterance section and calculating acoustic features from each section, the accuracy of discrimination can be improved. The technique based on acoustic features is able to detect system requests reasonably because it will not be dependent on any task and it does not need to reconstruct the discriminator when the system requests are added or changed.

Also, consideration of the alternation of speakers is described in this chapter. Considering turn-taking before and after the utterance, the performance can be improved. Finally, we take linguistic features into account, where Boosting is employed as a discriminant method. Its output score is not a probability, though, so the Boosting output score is converted into pseudo-probability using a sigmoid function. Though the technique based on linguistic features is dependent on tasks and it will need to reconstruct the discriminator when the system requests are modified, the accuracy of discrimination using linguistic features is better than that of the technique based on acoustic features.

2. Utterance verification using acoustic features

We describe the system request detection based on acoustic features first, where SVM (Support Vector Machine) is used. The overview of the system is shown in Figure 1. The proposed method based on acoustic features is able to detect system requests reasonably, because it does not need to reconstruct the discriminator when the system requests are added or changed.
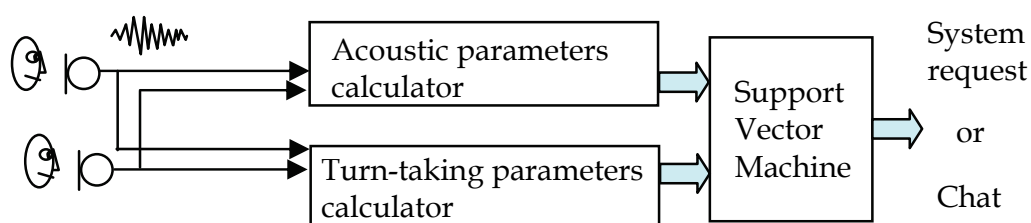

Fig. 1. System overview of utterance verification using acoustic features

2.1 Acoustic parameters

Even if we speak unconsciously, there are acoustic differences between utterances to equipments and those to humans under the condition the subject equipment is machinelike. In our work, we focus on the different characteristics of commands and human-human conversations which usually appear on the head and the tail of the utterance.

The start point and the end point of the utterance are indistinct in chatters while there are no sounds before and after the utterance in commands. There are mainly two reasons that make the start and the end point unclear. One reason is there are usually fillers and falters in chatters while there are short pauses on the head and the tail of utterances in commands. We usually put a short pause before a command to clarify and keep quiet until the system responds something. The other reason is the following person often begins to talk while the current person does not finish talking yet. In this Section, we deal with the former case. To put the former phenomenon to practical use, we calculate acoustic parameters not from the whole utterance section but from each three sections below.

To extract the head and the tail of the utterance, the power and zero-crossing are used in this work. Figure 2 is the wave form of a command utterance, and Figure 3 is that of a spontaneous utterance (chat). The head and tail of the utterance are indistinct in chatters while there are no sounds before and after the utterance in commands as described above. Therefore, as the head and tail of the utterance contain useful information written above, we do not join these margins to the detected utterance section, but calculate acoustic parameters (Table 1) also from each margin separately.

Calculated acoustic parameters are 8 dimensions shown in Table 1, but we calculate them from three sections described above. Thus, the acoustic features are 24 dimensions. The power is computed by Root Mean Square (RMS). The pitch is calculated by LPC residual correlation.
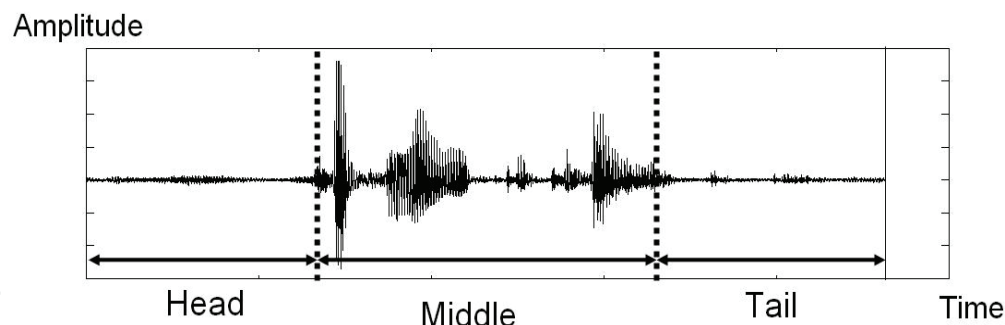

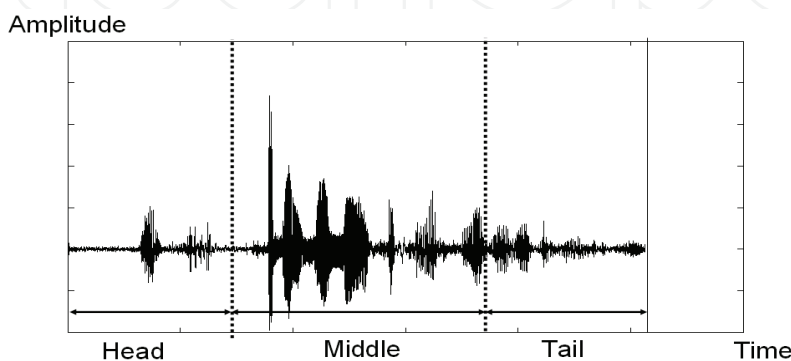Fig. 2. A sample of system request

Fig. 3. A sample of spontaneous utterance (chat)

• Power	Average	Standard deviation	Max.	Max. – Min.
• Pitch	Average	Standard deviation	Max.	Max. – Min.

Table 1. Acoustic parameters (Power and Pitch are used.)

2.2 Turn-taking parameters

The sounds in the head and tail margins sometimes contain a speech of the next person, though it is not so loud. Therefore, we should separate voices of the next person from fillers and flatters. Considering which person speaks in each utterance section improves the accuracy of utterance verification. For example, the utterance seems to be a chat if speakers changes like $B \rightarrow A \rightarrow B$ in each section. In this work, we calculate these turn-taking parameters by crosspower-spectrum phase (CSP) (M. Omologo and P. Svaizer, 1996). Under the condition two microphones are set up for each person, we can tell the speaker from which microphone receives the utterance first. Considering the time lag CSP shows the maximum value, we can tell which microphone receives first. Moreover, CSP considers only the phase of the wave by normalizing the crosspower. This feature fits the condition that the distance of two microphones changes, where the power ratio of two microphones changes.

The crosspower-spectrum is computed through the short-term Fourier transform applied to windowed segments of the signal $x_i[t]$ received by the i-th microphone at time t:

$$CS(n;\omega) = X_i(n;\omega)X_j^*(n;\omega) \tag{1}$$

where * denotes the complex conjugate, n is the frame number, and ω is the spectral frequency. Then the normalized crosspower-spectrum is computed by

$$\phi(n;\omega) = \frac{X_i(n;\omega)X_j^*(n;\omega)}{|X_i(n;\omega)||X_j(n;\omega)|} \tag{2}$$

that preserves only information about phase differences between x_i and x_j. Finally, the inverse Fourier transform is computed to obtain the time lag (delay).

$$C(n;l) = F^{-1}\phi(n;\omega) \tag{3}$$

If the sound source does not move (this means it does not move in an utterance), $C(n;l)$ should consist of a dominant straight line at the theoretical delay. Therefore, a lag is given as follows:

$$\hat{l} = \arg\max_l \left\{ \sum_{n=1}^{N} C(n;l) \right\} \tag{4}$$

In the situation that the microphones are set up for each person, the reliability of the lag is the matters. Thus, we calculate D from each section and make them turn-taking parameters.

$$D = \begin{cases} C(\hat{l}) & (0 \le \hat{l} < (N-1)/2) \\ -C(\hat{l}) & ((N-1)/2 \le \hat{l} < N-1) \end{cases} \tag{5}$$

3. Utterance verification using linguistic information

In this Section, we describe the proposed method that incorporates system request into a speech recognition system, where linguistic information in the system request task is used.

3.1 System request detection integrated with speech recognition

Speech recognition is formalized to find the most likely word sequence $W = \{w_1,...,w_N\}$ as well as the system request intention $s = \{Request, Chat\}$. Given the sequence of observed feature vectors O, speech recognition is formalized as follows:

$$
\begin{aligned}
(\hat{s}, \hat{W}) &= \arg\max_{s,W} P(s, W \mid O) \\
&= \arg\max_{s,W} \frac{P(s, W, O)}{P(O)}
\end{aligned}
\tag{6}
$$

The following Eq. (7) and (8) can be derived from the Bayesian theorem, where $P(O)$ is omitted due to independence from s and W.

$$
P(s, W, O) = P(s)P(W \mid s)P(O \mid W, s) \tag{7}
$$

$$
P(s, W, O) = P(W)P(O \mid W)P(s \mid W, O) \tag{8}
$$

Therefore, two scenarios (Eq. (7) and (8)) are considered in this work. First, Eq. (7) means that the acoustic model and the language model both depend on request intention s. In Eq. (7), we employ the request intention dependent language model and assume that the acoustic model is independent from request intention s. The N-gram which is dependent on the request intention is given by

$$
P(W \mid s) = \prod_i P(w_i \mid w_{i-1},..., w_{i-N+1}, s) \tag{9}
$$

$P(W \mid s = Request)$ and $P(W \mid s = Chat)$ are learned from the system request corpus and conversation corpus, respectively. After the recognition process using two language models, we find the request intention label having the maximum likelihood.

Next, the formulation of Eq. (8) consists of normal acoustic and language models. These models are the same as speech recognition models without request intention. In addition, Eq. (8) includes the model $P(s \mid W, O)$ that discriminates system requests based on word hypothesis **W** and observation **O** directly. $P(s \mid W, O)$ is a discrimination model such as Boosting or Support Vector Machines (SVM). Here, we employ a Boosting model due to computational costs, flexibility of expression and ease of combining various features. However, Boosting is not a probabilistic model. It is necessary to convert Boosting output $f(W, O)$ into pseudo-probability so that it can be incorporated into the probability-based speech recognition system. Consequently, Boosting output is converted into pseudo-probability using sigmoid function as shown in Figure 4. Sigmoid function can model close to the discriminative boundary in detail, and the range of values is 0 to 1. The parameters, a and b, are weighting factors of the sigmoid function, and they are estimated by the gradient method. Converting Boosting output $f(W)$ into pseudo-probability leads to the following derived equations:

$$P(s = \text{Request} \mid \text{W}, \text{O}) \approx \text{sigmoid}(f(\text{W}))$$
$$P(s = \text{Chat} \mid \text{W}, \text{O}) \approx 1 - \text{sigmoid}(f(\text{W})) \tag{10}$$

Here, language information only is used.

By integrating system request detection with speech recognition, system request detection can incorporate not only 1-best results but also hypotheses. In addition, it makes it possible to decide the hypothesis for request detection based on a probability framework. For example, there are two hypotheses, such as "Come here" and "You say come here." Here "Come here" is a system request and "You say come here" is a chat. In order to integrate these scores and speech recognition probabilities, these scores from AdaBoost are converted into pseudo-probabilities. After integration, the hypothesis with the best scores is selected as a result of system request detection. Even if the speech recognition probability, $P(\text{W})P(\text{O} \mid \text{W})$, of "You say come here." is larger than "Come here," when the boosting score of "Come here" is high enough, "Come here" will be selected as a final result.
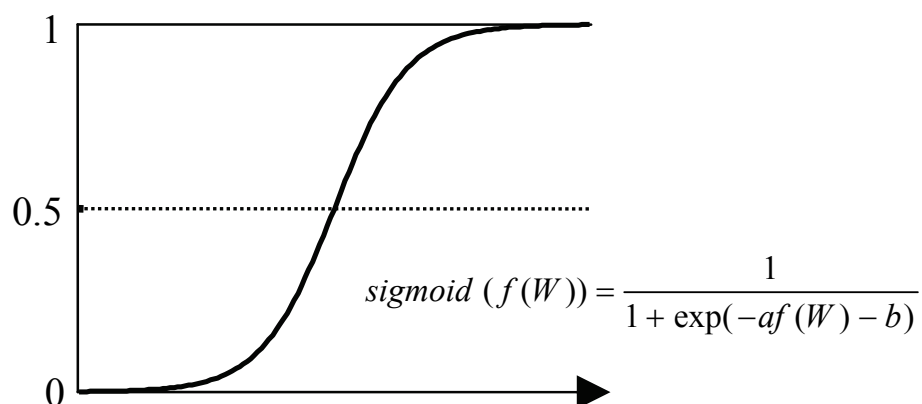

Fig. 4. Sigmoid function. Boosting output is converted into pseudo-probability using the sigmoid function.

3.2 Boosting

In this subsection, we describe a discrimination model based on Boosting in order to calculate $P(s \mid \text{W}, \text{O})$ in Eq. (10). AdaBoost is one of the ensemble learning methods that construct a strong classifier from weak classifiers (R. Schapire, et al., 1998). The AdaBoost algorithm uses a set of training data, $\{(W_1, Y_1), \ldots, (W_n, Y_n)\}$, where W_n is the n-th feature. In this work, the feature is a word (unigram) or a pair of words (N-gram). Y is a set of possible labels. For the system request detection, we consider just two possible labels, $Y = \{-1, 1\}$, where the label, 1, means "system request," and the label, -1, means "chat." For weak classifiers, single-level decision trees (also known as decision stumps) are used as the base classifiers (R. Schapire, et al., 2000). The weak learner generates a hypothesis $h_t : W \rightarrow \{-1, 1\}$ that has a small error. In the weak learner proposed by Schapire at el., the weak learners search all possible terms (unigram word or a pair of words) in training data and check for the presence or absence of a term in the given utterance. Once all terms have been searched, the weak hypothesis with the lowest score is selected and returned by the weak learner.

Next, AdaBoost sets a parameter α_t according to Eq. (13). Intuitively, α_t measures the importance that is assigned to h_t. Then the weight $z_{t+1}(i)$ is updated.

$$z_{t+1}(i) = \frac{z_t(i)\exp\{\alpha_t I(h_t(W_i) \neq Y_i)\}}{\sum\limits_{j=1}^{n} z_t(j)\exp\{\alpha_t I(h_t(W_j) \neq Y_j)\}} \tag{11}$$

The Eq. (11) leads to the increase of the weight for the data misclassified by h_t. Therefore, the weight tends to concentrate on "hard" data. After T-th iteration, the final hypothesis, $f(W)$, combines the outputs of the T weak hypotheses using a weighted mojority vote. The following shows the overview of the Adaboost.

Input: n examples $\{(W_1, Y_1), ..., (W_i, Y_i), ..., (W_n, Y_n)\}$

Initialize: $z_1(i) = 1/n, \quad i = 1, ..., n$

Do for $t = 1, ..., T$

1. Train a weak learner with respect to the weight z_t and obtain hypothesis $h_t : W \rightarrow \{-1, 1\}$

2. Calculate the training error e_t of h_t.

$$e_t = \sum\limits_{i=1}^{n} z_t(i)\frac{I(h_t(W_i) \neq Y_i) + 1}{2} \tag{12}$$

3. Set

$$\alpha_t = \log\frac{1 - e_t}{e_t} \tag{13}$$

4. Update the weight

Output: final hypothesis

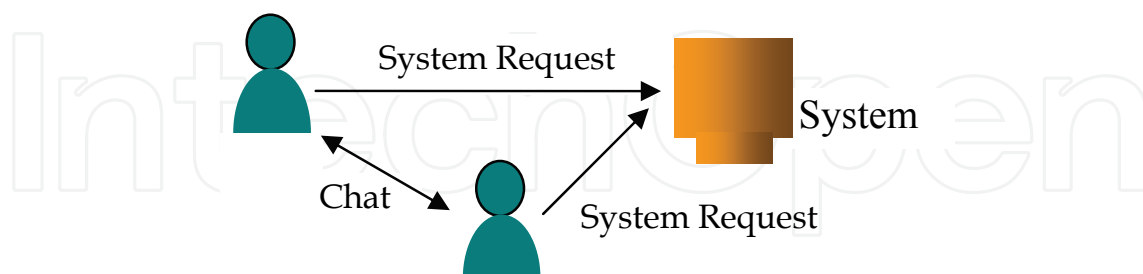$$f(W) = \frac{1}{|\alpha|}\sum\limits_{t=1}^{T} \alpha_t h_t(W) \tag{14}$$

Fig. 5. Two person + one system dialog

4. Experiments

4.1 Recording conditions and details of corpus
The overview of the recording condition is shown in Figure 5. The task has the following features.

- Two people are in proximity to the system concurrently.
- People talk with each other freely and make requests to the system at will.
- The system has several kinds of functions (Table 2).
- Commands and utterances are recorded through microphones clipped to the chest of each speaker.

Functions	Sound source direction estimation based on CSP
	Move toward/away from sound source
	Obstacle avoidance
	Place a bottle using the gripper
	Take a face photo
Command examples	*Kotchi ni kite.* (Come here.)
	Shashin wo totte. (Take my photo.)
	Mukoh he itte. (Go to the other side.)
	Watashi ni tsuite kite. (Come with me.)
	Bottle wo oite. (Place the bottle.)

Table 2. Functions of the robot

It is ordinary for two or more people to be in close proximity to the system at the same time. For example, a driver uses a car navigation system while talking with passengers, or someone controls a robot in the presence of an audience. In our experiment, we used the robot for the system as shown in Figure 6. The typical usages are to call the robot by saying, "Come here" and to have the robot take a picture by speaking "Take my picture." The robot

Fig. 6. Picture of mobile robot built in this work

can recognize the fixed commands shown in Table 1 at present. However, recorded speeches include many other command expressions such as "Come on," "Come rapidly," "Come, uhh … here," etc. These utterances are spoken to control the robot. However the robot cannot recognize these at present. We labeled these utterances as a system request since one of our purposes is to accept flexible expressions (we collected these utterances on purpose).

Non-request utterances consist of ordinary conversation statements. These utterances are spoken in spontaneous speaking style, and so it is too difficult to recognize accurately. In

addition, explanation utterances of the robot usage were included. For example, "You say, 'Come here,' and the robot will come," "Come here, go away and so on," etc. Note that these utterances include the same phrases that are found in the system requests.

The length of the recording time is 30 minutes. We labeled those utterances manually. Table 3 shows the result of cutting out utterances from the recorded speech data.

Total utterance	System request	Total vocabulary size
330	49	700

Table 3. Total number of utterances and system requests

4.2 Evaluation of utterance verification using acoustic features

First, experiments were performed to test the utterance verification (system request detection) using the acoustic features. In this work, we used SVM with RBF (Gaussian) kernel. When more than two kinds of parameters are used at the same time, we combined parameters as follows:

$$U = [\alpha P_1 \quad \beta P_2] \tag{15}$$

Here U is combined vector and the original feature vectors are P_1, P_2, α and β were given experimentally.

Table 4 shows the results of utterance verification evaluated by leave-one-out cross-validation. In this experiment, we set 0.7 seconds for both margins before and after the clear utterance sections. The results are the cases F-measure became the maximum values. The F-measure became 0.86 where acoustic parameters (24 dim.) are calculated from proposed three utterance sections, while that was 0.66 where the feature values (8 dim.) are calculated from a whole utterance. Then, adding turn-taking features, it turned out to be 0.89.

	Precision	Recall	F-measure
Acoustic (8 dim.)	0.71	0.61	0.66
Acoustic (24 dim.)	0.80	0.92	0.86
Acoustic (24 dim.) + Turn-taking	0.87	0.92	0.89

Table 4. Result of utterance verificat

4.3 Evaluation of utterance verification using linguistic information
4.3.1 Conditions of speech recognition

In the acoustic model, the baseline training data consisted of about 200,000 Japanese sentences (200 hours) spoken by 200 males in the Corpus of Spontaneous Japanese (S. Furui et al., 2002). Table 5 shows the conditions of acoustic analysis and the specification of HMM (left to right). To improve speech recognition accuracy, acoustic model adaptation was performed. Utterances for adaptation are different from those in the test set, but that speaker who recorded the utterances for adaptation was the same one used in the test set.

Language models were constructed using manual transcriptions of various utterances. Here, to meet open conditions, the language model for recognizing speaker A was constructed by transcriptions of speaker B. Note that the dictionary for speech recognition includes all words spoken by A and B. Thus, the out-of-vocabulary (OOV) rate was zero. For the multi

N-gram method (corresponding to Eq. (7)), language models were constructed for each speaker and each request intention (request and conversation). As a result of speech recognition, though word accuracy was 42.1%, F-measure of keywords was 0.67.

Sampling rate / Quantization	16 kHz / 16 bit
Feature vector	39-order MFCC
Window	Hamming
Frame size / shift	20 / 10 ms
# of phoneme categories	244 syllable
# of mixtures	32
# of states (vowel)	5 states and 3 loops
# of states (consonant + vowel)	7 states and 5 loops

Table 5. Experimental conditions of acoustic analysis and HMM

4.3.2 Results of system request detection

Experiments of request detection using speech recognition results were also performed using the 10-fold cross-validation method. Four experiments (Multi N-gram, sig-Boosting, Boosting, Confidence) were performed. Multi N-gram is based on Eq. (7). Sig-Boosting is based on Eq. (8). This method is system request detection integrated with speech recognition. Sig-Boosting incorporates not only 1-best results of speech recognition but also hypotheses. Boosting incorporates only 1-best results. In order to compare a conventional method, the experiment using the "confidence" method was performed. This method discriminates system requests based on confidence measures of speech recognition. If the average confidence measure of each word is larger than a threshold, an utterance is discriminated as a system request.

The experimental results are shown in Figure 7. We can see that sig-Boosting method achieved the best performance. Intrinsically, the Boosting method showed high
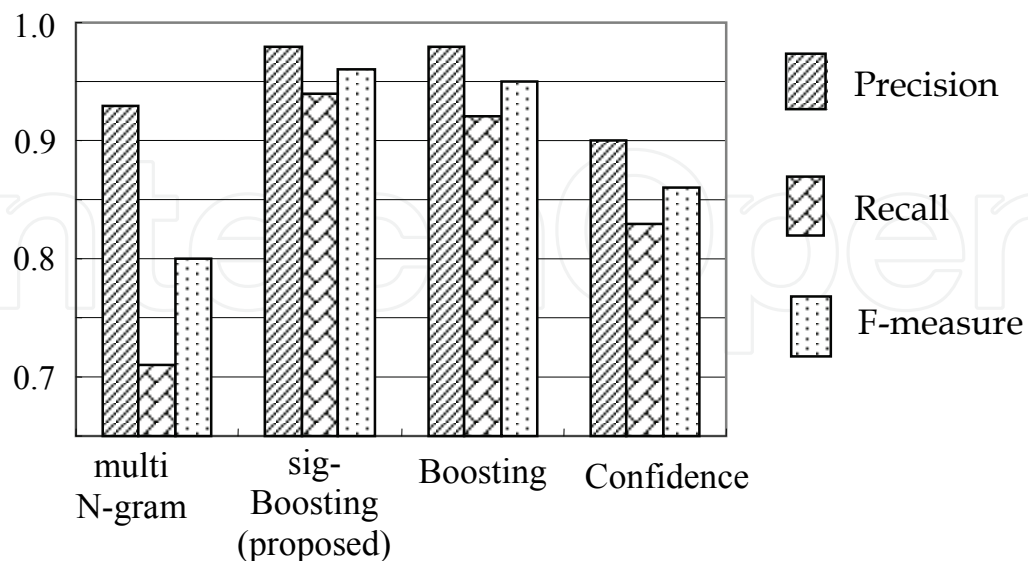

Fig. 7. Results of system request detection using linguistic information

performance. In addition, sig-Boosting recovered false-negative errors by incorporating speech recognition hypotheses. In the case where the 1-best results miss important

keywords, considering the hypotheses, the proposed method can recover the keywords from the hypotheses and improved the performance. On the other hand, the multi N-gram method and confidence method could not achieve performance as high as Boosting methods. Especially, these methods tend to mis-classify the utterances whose intention depends predominantly on one word: e.g., "toka" (meaning "etc.").

5. Conclusion

To facilitate natural interaction for a system such as mobile robot, a new system request utterance detection based on acoustic and linguistic features was employed in this chapter. To discriminate commands from human-human conversations by acoustic features, it is efficient to consider the head and tail of an utterance. The different characteristics of system requests and spontaneous utterances appear on these parts of an utterance. Separating the head and the tail of an utterance, the accuracy of discrimination was improved. Considering the alternation of speakers using two channel microphones also improved the performance.

Also we described the system request detection method integrated with a speech recognition system. Boosting was employed as a discriminant method. Its output score is not a probability, though, so the Boosting output score was converted into pseudo-probability using a sigmoid function. The experimental results showed that integration of system request detection and speech recognition improved the performance of request detection. Especially, in the case where 1-best results miss important keywords, the proposed method can recover the keywords from the hypotheses and improve the performance.

In the future, we plan to perform experiments using larger corpus and more difficult tasks. In addition, we will investigate a context-dependent approach for request detection. The consideration of new kinds of features is also the assignments.

6. References

H. G. Okuno, K. Nakadai & H. Kitano (2002). Social interaction of humanoid robot based on audio-visual tracking, *Proceedings of Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems,* LNAI2358, pp. 725-735, 2002.

J. Miura, et al. (2003). Development of a personal service robot with user-friendly interfaces, *Proceedings of Int. Conf. on Field and Service Robotics*, pp. 293-298, 2003.

S. Waldherr, R. Romero & S. Thrun (2000). A gesture based interface for human-robot interaction, *Autonomous Robots*, 9(2), pp. 151-173, 2000.

H. Asoh, et al. (1999). A spoken dialog system for a mobile robot, *In Proceedings of Eurospeech*, pp. 1139-1142, 1999.

R. Stiefelhagen, et al. (2004). Natural human-robot interaction using speech, head pose and gestures, *In Proceedings of Int. Conf. on Intelligent Robots and Systems.*, pp. 2422-2427, 2004.

T. Takiguchi, et al. (2007). Voice and Noise Detection with AdaBoost, *Chapter on Robust Speech Recognition and Understanding, Book edited by M. Grimm and K. Kroschel., I-Tech Education and Publishing,* pp. 67-74, 2007.

S. Yamada, et al. (2005). Linguistic and Acoustic Features Depending on Different Situations – The experiments considering speech recognition rate, *In Proceedings of Interspeech,* pp. 3393-3396, 2005.

T. Kawahara, et al. (1998). Speaking-style dependent lexicalized filler model for key-phrase detection and verification, *In Proceedings of ICSLP*, pp. 3253-3259, 1998.

P. Jeanrenaud, et al. (1994). Spotting events in continuous speech, *Proceedings of ICASSP*, pp. 381-384, 1994.

T. Yamagata, A. Sako, T. Takiguchi, and Y. Ariki (2007). System request detection in conversation based on acoustic and speaker alternation features, *In Proceedings of Interspeech*, pp. 2789-2792, 2007.

M. Omologo & P. Svaizer (1996). Acoustic source location in noisy and reverberant environment using CSP analysis, *Proceedings of ICASSP*, pp. 921-924, 1996.

R. Schapire, et al. (1998). Boosting the margin : A new explanation for the effectiveness of voting methods, *Annals of Statistics*, vol. 26, no. 5, pp. 1651-1686, 1998.

R. Schapire, et al. (2000). BoosTexter : A Boosting-based System for Text Categorization, *Machine Learning*, 39(2/3, pp. 135-168, 2000.

S. Furui, et al. (2002). BoosTexter : Spontaneous Speech : Corpus and Processing Technology, *The Corpus of Spontaneous Japanese*, pp. 1-6, 2002.

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:

T. Takiguchi, A. Sako, T. Yamagata and Y. Ariki (2008). System Request Utterance Detection Based on Acoustic and Linguistic Features, Speech Recognition, France Mihelic and Janez Zibert (Ed.), ISBN: 978-953-7619-29-9, InTech, Available from:

http://www.intechopen.com/books/speech_recognition/system_request_utterance_detection_based_on_acoustic_and_linguistic_features

INTECH
open science | open minds