

System Request Detection in Conversation Based on Acoustic and Speaker Alternation Features

Tomoyuki YAMAGATA¹, Atsushi SAKO², Tetsuya TAKIGUCHI¹, Yasuo ARIKI¹

¹Department of Computer and System Engineering
Kobe University, 1-1 Rokkodai, Nada, Kobe, 657-8501, JAPAN

²Department of Informatics and Electronics
Kobe University, 1-1 Rokkodai, Nada, Kobe, 657-8501, JAPAN

{yamagata, sakoats}@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

For a hands-free speech interface, it is important to detect commands in spontaneous utterances. To discriminate commands from human-human conversations by acoustic features, it is efficient to consider the head and the tail of an utterance. The different characteristics of system requests and spontaneous utterances appear on these parts of an utterance. Experiment shows that by separating the head and the tail of an utterance, the accuracy of detection was improved. And also, considering the alternation of speakers using two channel microphones improved the performance. Although detecting system requests using linguistic features shows high accuracy, combining acoustic and turn-taking features lift up the performance.

Index Terms: system request detection, utterance verification, SVM, speech recognition, turn-taking

1. Introduction

Recently, speech interfaces are usually applied to the equipment which users cannot operate by hands; car navigation, robot. However, these interfaces have a problem that they cannot discriminate system requests - utterances which users talk to a system - from human-human conversations. Therefore, a speech interface of a car navigation today requires a physical button which on and off the microphone input. If there is no button for a car navigation, all conversations are recognized as commands for the system. The button spoils the merit of speech interfaces which users do not need to operate by the hand.

Speech spotter[1] is one of the solution to the problem. However, speech spotter requires users to change the style of utterance consciously. Concerning this issue, there are researches on discriminating system requests from human-human conversation by acoustic features calculated from each utterance [4]. And also, there are discrimination techniques using linguistic features. Keyword or key-phrase spotting based methods[2, 3] 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 paper, firstly we propose an advanced method of discrimination using only acoustic features. The difference of system requests and spontaneous utterances usually appears on the head and the tail of the utterance. By separating the utterance

section and calculating acoustic features from each section, the accuracy of discrimination was improved. Secondary, we introduce the consideration of the alternation of speakers. Considering turn-taking before and after the utterance, the performance was improved. Finally, we take linguistic features into account. Though the accuracy of discrimination using linguistic features is good itself, combining acoustic and turn-taking features lift up the performance.

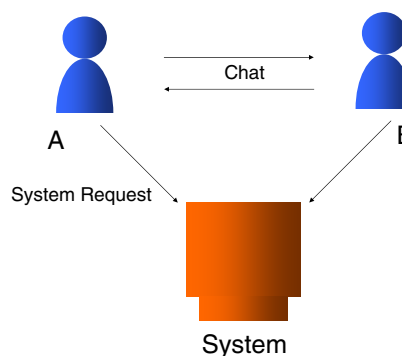


Figure 1: Two person + one system dialog.

2. Recording Conditions and Details of Corpus

The corpus for evaluation is recorded under the situation where two people and a system in a same place (Figure.1). Two people talk each other and sometimes make request to the system. This situation is quite common. For example, two people is in a car and operates a car navigation. In this paper, we used a mobile robot as a system, because recording in a real car causes noise problems. Our task is to detect system requests from various spontaneous utterances.

The whole picture of the robot is Figure 2. It consists of two microphones (those are different from recording microphones), two omni cameras (upper view and lower view), a laptop computer to control, a gripper to place a bottle, wheels and motors (advancement, retreat, rotation). The functions of the robot are shown in Table 1. Generally, we operate the robot by speaking a command a few meters away from it.

The recording microphones were set up on the breast of each person. The length of the recording time is 30 minutes.

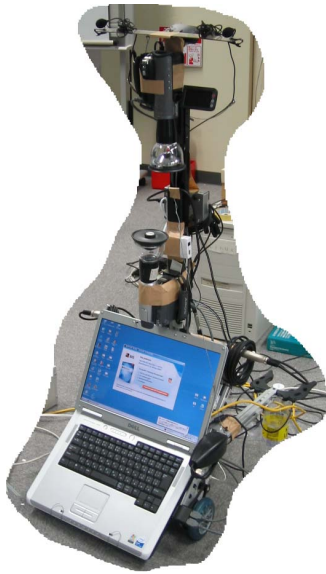


Figure 2: Picture of mobile robot.

Table 1: Function list of the mobile robot.

Functions	Sound source direction presumption based on CSP
	Move toward/backward sound source
	Obstacle avoidance
	Place a bottle by the gripper
Command examples	Take a face picture
	"Kotchi ni kite. (Come here.)"
	Mukou he itte. (Go to the other side.)"
	"Shashin wo totte. (Take my picture.)"
	"Watashi ni tsuite kite. (Come with me.)"
	Bottle wo oite. (Place the bottle.)"

We did not show them the list of commands that the robot can accept. One reason is to increase the variation of system request commands. The other reason is that we are going to develop a speech interfaces which accept not only specified commands but also various expressions. Therefore, they could speak commands that might be acceptable to the robot. We labeled those utterances as system requests manually. Table 2 shows the result of cutting out utterances from the record by power and zero-crossing.

3. Utterance verification in Spontaneous Speeches

We propose the system requests detecting method based on SVM. The overview of the system is shown in Figure 3. We describe acoustic parameters first. It is possible to detect system requests reasonably with acoustic features, because it does not need to reconstruct the discriminator when the system requests are added or changed. Calculated acoustic parameters are 8 di-

Table 2: The numbers of utterances and system requests.

Total utterance	System request
330	49

mensions shown in Table 3, but we calculate them from three sections described in 3.1. Thus, the acoustic features are 24 dimensions. Then, we describe turn-taking parameters. Turn-taking parameters are 3 dimensions calculated from the three sections. Considering the speaker's alternation, the distinction accuracy was improved. And, it is also effective to use linguistic features which based on term frequencies of each utterance. It can be said that the linguistic parameters are consists of frequencies of the system request words and garbage words. We describe linguistic parameters at the end.

3.1. 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 [4]. In this paper, we focus on the different characteristics of commands and human-human conversations which usually appear on the head and the tail of the utterance. For example, Figure 4 is the wave form of a command utterance, and Figure 5 is that of a spontaneous 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.

Utterance sections are detected by power and zero-crossing. But the method can detect only clear utterance sections. To detect whole utterances in spontaneous speeches, it is easy to put margins before and after the detected utterance sections. However, these margins contain some problems written above. Therefore, we do not join these margins to the detected utterance section, but calculate acoustic parameters (Table 3) also from each margin separately.

The power is computed by Root Mean Square (RMS). The pitch is calculated by LPC residual correlation. Table 4 shows the conditions of pitch estimation.

Table 3: Acoustic Parameters.

Power	Ave.	S.D.	Max.	Max. - Min.
Pitch	Ave.	S.D.	Max.	Max. - Min.

Table 4: Conditions of pitch estimation.

Sampling rate	16 kHz	Window type	Hamming
Frame length	25 ms	Max. pitch freq.	300 Hz
Frame shift	16 ms	Min. pitch freq.	70 Hz

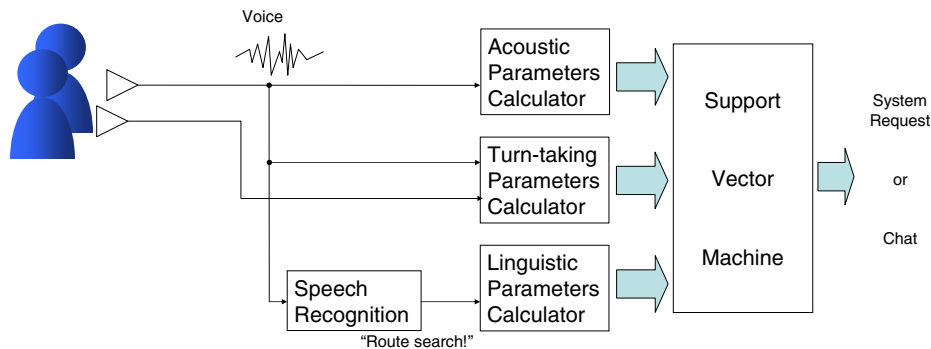


Figure 3: System overview of utterance verification.

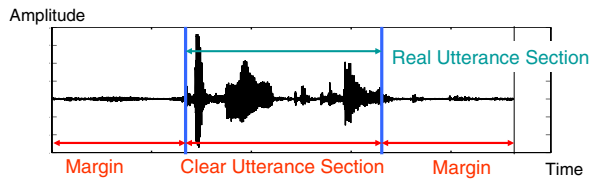


Figure 4: A sample of system request.

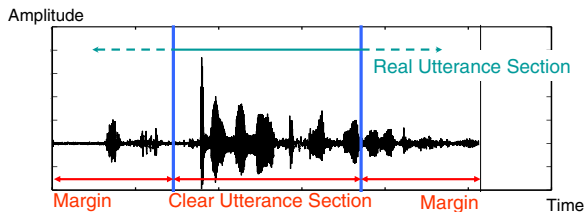


Figure 5: A sample of spontaneous utterance.

3.2. Turn-taking Parameters

The sounds in the head and tail margins sometimes contains 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 paper, we calculate these turn-taking parameters by crosspower-spectrum phase (CSP) [5]. 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 normalize 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 $s_i[t]$ received by the i -th microphone at time t : $CS(n; \omega) = X_i(n; \omega) X_j^*(n; \omega)$, where $*$ denotes the complex conjugate, n is the frame number, and ω is the spectral frequency. Then the normalized crosspower-spectrum is computed by the following:

$$\phi(n; \omega) = \frac{X_i(n; \omega) X_j^*(n; \omega)}{|X_i(n; \omega)| |X_j(n; \omega)|} \quad (1)$$

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

$$C(n; l) = \mathcal{F}^{-1} \phi(n; \omega) \quad (2)$$

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:

$$C(\hat{l}) = \underset{l}{\operatorname{argmax}} \left\{ \sum_{n=1}^N C(n; l) \right\} \quad (3)$$

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

$$D = \begin{cases} C(\hat{l}) & (0 \leq \hat{l} < \frac{N-1}{2}) \\ -C(\hat{l}) & (\frac{N-1}{2} \leq \hat{l} < N-1) \end{cases} \quad (4)$$

3.3. Linguistic Parameters

Linguistic features are term frequencies calculated from the results of speech recognition. Feature vectors consist of frequencies of the system request words and chatter words. Verifying utterances by linguistic features works accurately on the situation where the domain of the task is limited.

To verify utterances by linguistic features, we need to recognize speech first. In this subsection, we account for the conditions of speech recognition first. Then, we describe the method to calculate linguistic parameters.

To create a baseline of the acoustic model, we use about 200,000 Japanese sentences (200 hours) spoken by 200 males recorded in Corpus of Spontaneous Japanese (CSJ) [6]. Table 5 shows the conditions of acoustic analysis and the specification of HMM. To improve the speech recognition accuracy, the acoustic model adaptation by MLLR+MAP [7] was performed as closed after the construction of a baseline model. The adaptation data from our corpus is almost 10 minutes. The language model is made from the manually recognized data. In order to build the language model to be open condition for speaker-A, we use the transcriptions of speaker-B. Feed the corpus into Julius [8] - large vocabulary continuous speech recognition software - under these conditions, the results was obtained with 42.1% word accuracy.

From the results, linguistic parameters are computed. Term frequency vectors in each utterance are calculated. Then, the vectors are employed as linguistic parameters.

Table 5: Conditions Automatic Speech Recognition.

Acoustic Analysis	Sampling rate	16 kHz
	Feature parameters	MFCC (25 dim.)
	Frame length	20 ms
	Frame shift	10 ms
	Window type	Hamming
HMM	Type	244 Syllables
	Mixture	32 mix
	Vowel(V)	5 states 3 loops
	Consonant+Vowel(CV)	7 states 5 loops

4. Experiments

Experiments were performed to test the utterance verification using the proposed parameters. We used SVM^{light} [9] for support vector machine with RBF (Gaussian) kernel. When more than two kinds of parameters are used at the same time, we combined parameters as follows:

$$U = [\alpha P_1 \beta P_2], \quad (5)$$

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

Table 6 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.

Using only linguistic features, the result was 0.94. Because the domain of the commands the robot can accept is not so big, verifying utterances by linguistic parameters shows good result. Then, adding acoustic parameters, the f-measure became 0.95. And also considering turn-taking features, it reaches 0.96.

Table 6: Result of Utterance verification.

	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
Linguistic	0.94	0.94	0.94
Linguistic + Acoustic (24 dim.)	0.94	0.96	0.95
Linguistic + Acoustic (24 dim.) + Turn-taking	0.98	0.94	0.96

5. Conclusions

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 progresses the performance also. It shows fairly high accuracy of detecting system requests using linguistic features, but combining acoustic and turn-taking features increase the accuracy even more.

Future work includes evaluation under the situation where the system accept many kinds of commands and enlarge the amount of corpus. The improvement of detecting utterance sections and the consideration of new kinds of features are also the assignments.

6. References

- [1] Masataka Goto, Koji Kitayama, Katunobu Itou, and Tetsumori Kobayashi, "Speech Spotter: On-demand Speech Recognition in Human-Human Conversation on the Telephone or in Face-to-Face Situations", Proceedings of ICSLP-2004, pp.1533-1536, October 2004.
- [2] Tatsuya Kawahara, Kentaro Ishizuka, Shuji Doshita, and Chin-Hui Lee, "Speaking-style Dependent Lexicalized Filler Model for Key-phrase Detection and Verification", Proceedings of ICSLP98, pp.3253-3259, 1998.
- [3] Jeanrenaud, P. Siu, M. Rohlicek, J.R. Meteer, M. Gish, H., "Spotting events in continuous speech", Proceedings of ICASSP-94, Vol.1, pp.381-384, Apr., 1994.
- [4] Shinya Yamada, Toshihiko Itoh and Kenji Araki, "Linguistic and Acoustic Features Depending on Different Situations - The Experiments Considering Speech Recognition Rate", Proceeding of Interspeech 2005, pp.3393-3396, Sep. 4-8, 2005.
- [5] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," Proc. ICASSP, 1996, pp. 921-924.
- [6] S. Furui, K. Maekawa, H. Isahara: "Spontaneous Speech: Corpus and Processing Technology", The Corpus of Spontaneous Japanese, pp.1-6, 2002-2.
- [7] E. Thelen, X. Aubert, "Speaker adaptation in the Philipps system for large vocabulary continuous speech recognition," Proc. of ICASSP-97, 1997, vol. 2, pp. 1035-1038.
- [8] A.Lee, T.Kawahara, and K.Shikano. "Julius - an open source real-time large vocabulary recognition engine," In Proc. EUROSPEECH, pp.1691-1694, 2001.
- [9] "SVM-Light Support Vector Machine," <http://svmlight.joachims.org/>