Human-Robot Interface Using System Request Utterance Detection Based on Acoustic Features

Tetsuya Takiguchi
Organization of Advanced Science and Technology
Kobe University, Japan
takigu@kobe-u.ac.jp

Tomoyuki Yamagata
Graduate School of Science and Technology
Kobe University, Japan
yamagata@me.cs.scitec.kobe-u.ac.jp

Atsushi Sako
Graduate School of Science and Technology
Kobe University, Japan

Nobuyuki Miyake
Graduate School of Science and Technology
Kobe University, Japan

Jerome Revaud
Institut National des Sciences Appliquées de Lyon
69621 Villeurbanne Cedex, France

Yasuo Ariki
Organization of Advanced Science and Tech.
Kobe University, Japan

Abstract

For a mobile robot to serve people in actual environments, such as a living room or a party room, it must be easy to control because some users might not even be capable of operating a computer keyboard. For non-expert users, speech recognition is one of the most effective communication tools when it comes to a hands-free (human-robot) interface. This paper describes a new mobile robot with hands-free speech recognition. For a hands-free speech interface, it is important to detect commands for a robot in spontaneous utterances. Our system can understand whether user's utterances are commands for the robot or not, where commands are discriminated from human-human conversations by acoustic features. Then the robot can move according to the user's voice (command). In order to capture the user's voice only, a robust voice detection system with AdaBoost is also described.

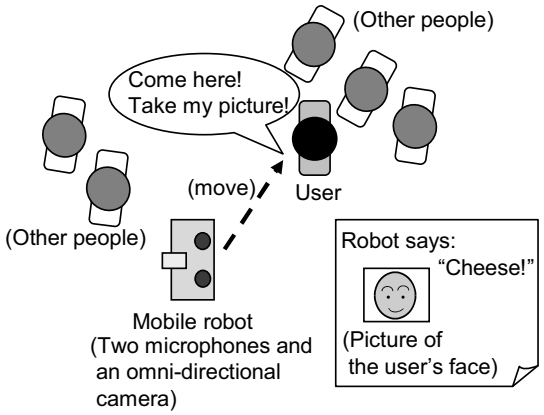

Figure 1. Scenario of mobile picture-taking robot.

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 [1][2]. 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 [3], 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 [4], 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 [5]. However, the noise-power-based method causes degradation of the detection performance in actual noisy environments.

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

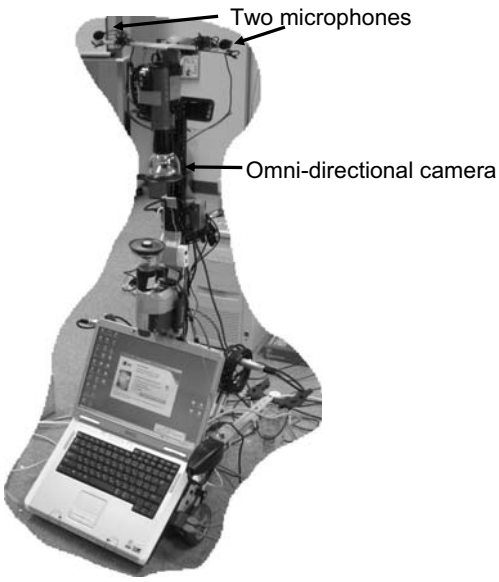

Figure 2. Picture of mobile robot built in this work.

acoustic features calculated from each utterance [6]. And also, there are discrimination techniques using linguistic features. Keyword or key-phrase spotting based methods [7, 8] 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, an advanced method of discrimination using only acoustic features is described. The difference of system requests and spontaneous utterances usually appears on the head and the tail of the utterance [9]. By separating the utterance section and calculating acoustic features from each section, the accuracy of discrimination was improved. In adition, a robust voice/non-voice detection algorithm using AdaBoost, which can achieve extremely high detection rates in noisy environments, is described in this paper [10].

Also, the user's direction estimation by CSP (Crosspower-Spectrum Phase) is implemented on the mobile robot. That enables the mobile robot to serve the user who calls to it from among other people. The two-channel noise reduction method is also implemented in order to improve the speech recognition performance. Using the user's direction estimated by the CSP method, the robot can move freely from its position to the user's position. After the mobile robot moves to the target position, it detects the user's face using the OpenCV library and takes the picture, which is integrated into the mobile robot's operating program.

2: Mobile Picture-Taking Robot

Figure 1 shows a scenario of the multi-modal robot that can take the user's picture (mobile picture-taking robot), and Figure 2 shows a picture of the mobile robot built in this work. The mobile robot can move intelligently in the user's direction by listening to the user's voice, and recognize what the user asks it. As shown in Figure 3, the robust speech recognition system on the mobile robot is composed of four steps. The first step is voice detection with AdaBoost, where