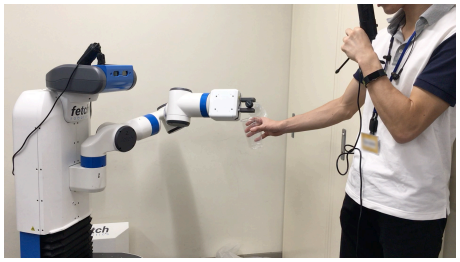

Figure 1: Fetch robot holding a bottle in its arm

Figure 2: Robot pouring a beverage

Extending Discrete Verbal Commands with Continuous Speech for Flexible Robot Control

Naoya Yoshimura[1,2]
Hironori Yoshida[1]
Fabrice Matulic[1]
1: Preferred Networks Inc., Tokyo
2: Osaka University
yoshimura.naoya@ist.osaka-u.ac.jp
{hyoshida,fmatulic}@preferred.jp

Takeo Igarashi
The University of Tokyo
takeo@acm.org

ABSTRACT

Speech is a direct and intuitive method to control a robot. While natural speech can capture a rich variety of commands, verbal input is poorly suited to finer grained and real-time control of continuous actions such as short and precise motion commands. For these types of operations, continuous non-verbal speech is more suitable, but it lacks the naturalness and vocabulary breadth of verbal speech. In this work, we propose to combine the two types of vocal input by extending the last vowel of a verbal command to support real-time and smooth control of robot actions. We demonstrate the effectiveness of this novel hybrid speech input method in a beverage-pouring task, where users instruct a robot arm to pour specific quantities of liquid into a cup. A user evaluation reveals that hybrid speech improves on simple verbal-only commands.

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312791

KEYWORDS
Human Robot Interaction; Continuous Control; Voice Inputs

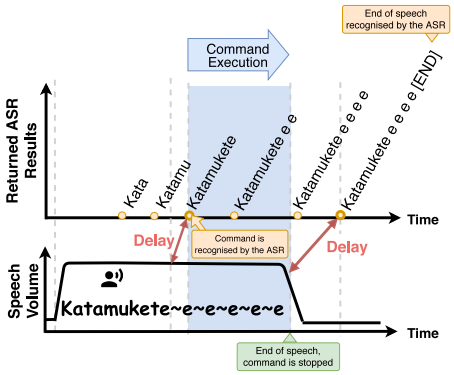

Figure 3: Example of a hybrid command. The command is started when a keyword is detected by the ASR and stopped when the volume filter detects the end of the speech. Delayed termination, as determined by the ASR, is thus avoided.

INTRODUCTION

Human voice has been widely used as a natural interface to control machines in a way that is similar to how people communicate with each other. The most common, and in most cases desirable, method to input commands is verbal speech, i.e. vocal commands consisting of natural words and phrases. This type of speech input requires the full sentence to be parsed and interpreted by a natural language processor for the command to be issued. As a result, verbal speech input is suitable for discrete operations, such as moving an object to a specific location, starting or stopping an action, etc. but does not lend itself to smooth real-time control of continuous actions, such as fine-tuned motion or the continuous adjustment of a parameter. To adequately support the latter cases using vocal input, techniques based on low-level sound features, like volume, pitch or frequency, have been proposed, e.g. to control the movement of a cursor [4] or a robot arm [3]. Those techniques allow smooth and fine-grained control, but they are less natural and limited in command variety.

In this late-breaking work, we propose to combine discrete verbal and continuous non-verbal speech input to enable both rich and fine real-time robot control. Our hybrid vocal input technique consists in extending the last vowel of a verbal command for as long as the command needs to be executed. For instance, a "move forward" command becomes "move forwaaaaaaaard", where the action is carried out until the utterance is stopped. As a preliminary proof of concept, we implement the recognition of hybrid speech commands by combining a standard speech recognition engine with detection of basic sound features. We comment on the technical challenges and limitations of this approach.

Our chosen application context to demonstrate the feasibility of our technique is a robot arm that pours a beverage in a cup held by a user (see Figure 1). Hybrid commands allow fine control of the pouring motion so that the desired quantity of beverage can be obtained. We evaluate our technique against simple verbal-only commands in a study with 12 participants. Results show that people perform better and feel more in control with hybrid speech than with plain verbal commands.

RELATED WORK

Research works and applications of speech-based input are numerous and here we just briefly review the literature on continuous vocal input.

Goto et al. [2] proposed a speech completion technique triggered by lengthening a vowel in the middle of a phrase. The detected filled pauses are only used to show completion candidates, not for task control. Igararashi and Hughes [4] used non-verbal features of voice, such as volume and pitch to support continuous control of sliders, cursors and game characters. In some examples, the non-verbal input is preceded by a phrase to specify which command should be executed. Similarly, the Vocal Joystick system allows people with motor impairments to control widgets on a screen [1] as

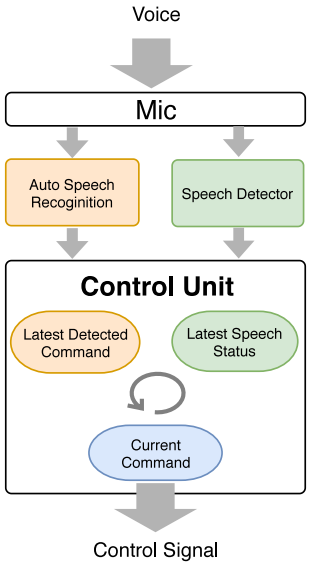

Figure 4: Flow Chart

Table 1: Commands in Japanese with corresponding action. The "teishi" (stop) command is only used for discrete input as speech stop signals the end of the command for Hybrid input.

No	Keyword	Action
1	katamukete (tilt)	rotate arm 7°/sec to pour
2	hayaku (fast)	rotate arm 12°/sec to pour
3	yukkuri (slow)	rotate arm 3°/sec to pour
4	modoshite (reverse)	rotate arm 7°/sec in reverse direction
5	teishi (stop)	stop
6	mouchotto (little more)	rotate arm 7°/sec to pour until end of speech then rotate arm in reverse direction 12°/sec

well as mechanical devices, such as robots [3], using continuous vocal input. Experiments showed that the optimal performance of the Vocal Joystick is comparable to a hand-based joystick for cursor manipulation. More recently, Takahashi and Mizuuchi [6] investigated how formants can be used to simultaneously control multiple degrees of freedom of a robot arm.

While low-level sound features such as pitch and formants allow for some degree of input differentiation, it cannot cover the breadth of commands that can be formulated with full words and phrases. Furthermore, it may be difficult to remember which vowel and sound attributes correspond to which operation. Igarashi and Hughes provide an example of a command word followed by a continuous vocal sound (a prolonged vowel "aaaa"), but the two are disconnected and thus this two-step sequence to issue a command may not feel very natural or comfortable to the user. Moreover, the technique was not evaluated.

HYBRID SPEECH INPUT

We introduce the concept of "hybrid speech input", which combines verbal speech and continuous non-verbal vocal input for natural and fine control. In this concept, the duration of a verbal command for a continuous action can be controlled by stretching out the last vowel of the command. For instance, a car that needs to be carefully backed into a parking space can be instructed to do so using the hybrid command "baaaaack". The car starts moving when the command is understood and stops when the speech stops. Note that this requires partial recognition of the command phrase as the action needs to be started before the phrase is completed. A straightforward method to realise this is to couple verbal command recognition using automatic speech recognition (ASR) with continuous low-level analysis of a desired sound property. In this first proof-of-concept implementation we use a standard speech recognition engine for the former and volume-based sound detection for the latter.

Speech Recognition

Our verbal speech recognition system is based on the free ASR Kaldi Speech Recognition Toolkit [5], which supports streaming speech input. The engine detects the start and stop of speech and in between continuously returns the partial phrase it has recognised so far. In this first step, we operate in a simplified setting, in which we recognise a set of specific task-relevant keywords and their synonyms contained in phrases. Our ASR further responds to keywords and phrases in Japanese, where all commands end with a vowel (see Table 1).

The partial results of the ASR are returned at irregular intervals and with some delay and thus cannot be relied upon to determine the duration of an extended command. While a command can only be started when it has been recognised by the ASR, we can avoid the longer delay to terminate it by monitoring instant low-level sound features such as volume. Our command processing pipeline is illustrated by means of an example in Figure 3 and Figure 4.

Figure 5: Small cup (left) and large cup (right) with target lines used in the experiments. A trial was considered successful if the poured water was within 1 cm above or below the target line.

We calculate the volume as a logarithmic ratio of effective sound pressure relative to a reference value. We use ambient sounds as the reference value. We also apply a high pass filter to remove frequency components below 375 Hz to reduce the influence of noise [4].

USER EVALUATION

There are many possible robotic applications for hybrid speech input. One such obvious task is navigation or precise positioning of a mobile robot. For our evaluation, we chose a beverage-pouring task as it requires precise real-time control. The user instructs the arm of a Fetch robot holding a bottle to pour a specified amount of liquid in a cup (see Figure 1). The user inputs vocal commands controlling the inclination of the arm and hence the speed at which the beverage is poured. In this experiment, we compare our hybrid technique with regular verbal speech.

Our goal was to assess two aspects of hybrid speech input: 1) overall feasibility, measured quantitatively by success rates (whether the user was able to pour the required quantity of liquid) and execution times; 2) subjective usability, evaluated by a questionnaire.

Experiment Protocol

Our study followed a within-subjects design in which participants commanded the robot to pour three different amounts of liquid in a plastic cup using, in turn, hybrid and verbal (discrete) speech input. The order was changed for each participants for counterbalancing. Participants stood in front of the robot arm with the glass extended just under the bottle. A container was placed below in case liquid spilled over (See Figure 2). Spoken commands were input via a headset microphone.

Participants were asked to perform two beverage-pouring tasks with two cups of different sizes (50ml and 400ml) twice, i.e. four tasks for each condition. The target quantities to be poured were shown by a line drawn on the cups (see Figure 5).

The robot was to be instructed using the command vocabulary of Table 1, which also included often used synonyms (not listed). For the hybrid condition, the length of speech determined command execution time as explained above. For the discrete condition, actions were performed until another command or a stop command was issued. A display showing the recognition results of the ASR was placed in front of participants for feedback.

12 participants (9 male and 3 female, aged from 22 to 39) were recruited for our experiments. Five had previous experience with manipulating robots.

After giving participants an overview of the study, they were given practice time for each of the techniques. First, they practised plain speech input with the ASR for 3 minutes in order to get a feel for how they should articulate commands to obtain good recognition results. Then, they were given 5 minutes to practise the actual pouring tasks with the robot. Pilot studies showed that people tended

Table 2: Average Task Success Rate

Method	All	Large	Small
Discrete	0.74	0.76	0.72
Hybrid	0.93	0.96	0.90

Table 3: Average Task Completion Time [Sec]

Method	All	Large	Small
Discrete	24.9	27.3	22.1
Hybrid	20.7	23.5	17.3

Table 4: The number of commands and their proportions (only commands used in successful trials are considered. Discrete: 29 trials, Hybrid: 40 trials.)

No	Command	Discrete	Hybrid
1	katamukete (tilt)	29 (16.3%)	28 (15.2%)
2	hayaku (quick)	9 (5.1%)	34 (18.5%)
3	yukkuri (slow)	34 (19.1%)	25 (13.6%)
4	modoshite (reverse)	28 (15.7%)	36 (19.6%)
5	teishi (stop)	29 (16.3%)	0 (0.0%)
6	mou chotto (bit more)	49 (27.5%)	61 (33.2%)
	Total	178 (100%)	184 (100%)

to speak faster when the cup was about to overflow, so participants were encouraged to control their diction and check the ASR feedback monitor.

Each session was video recorded. ASR results and robot movements were logged. Participants were given a questionnaire at the end of the experiments to provide feedback on aspects of speed, control and safety for the two techniques.

RESULTS

A session with a participant lasted roughly an hour.

Quantitative Results

We measured task success rates and execution times for each condition. A task was deemed successful if the poured liquid was within 10 mm of the target line. Execution times were measured from the time the first command was recognised until the final pouring action. The total latency of the command system as measured from the end of a spoken command to the start of its execution was at least 0.5 sec with slight variations depending on the recognition speed of the ASR. For hybrid commands, the delay between the end of speech indicating the end of the command and robot motion stop was consistently about 0.4 sec.

To mitigate the influence of ASR errors, we removed the results of trials, for which it was clear that failures were due to misrecognised commands (9 for Discrete and 5 for Hybrid). The results based on the remaining measurements are shown in Table 2 and Table 3 respectively. We provide those numbers only as reference without further statistical analysis, due to insufficient and unbalanced data as a result of the filtering. Nevertheless, the higher success rates and lower completion times for Hybrid correspond to our observations when commands were properly recognised. Table 4 shows the number and percentage of commands used in the successful trials.

Qualitative Assessment

In the questionnaire, we asked users to rate the following three aspects of the two techniques on a continuous scale from 0 to 1.

- **Usability** How much control did you have over the robot?
- **Safety** How safe did you feel when operating the robot?
- **Speed** How efficient were you, disregarding misrecognised commands?

Results are reported in Figure 6. Participants felt significantly safer and more in control with Hybrid input. Their perception of efficiency was generally also higher, but not significantly so.

Discussion

Despite training with the ASR, there were several cases of misrecognised commands and as a result, participants tended to be more careful and use slower commands to mitigate the impact, especially

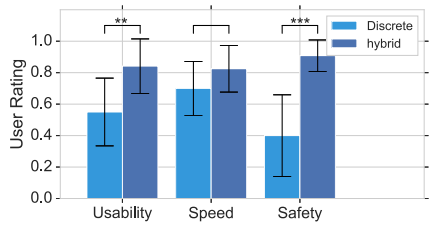

Figure 6: User Ratings for Usability, Speed and Safety. Significant differences, as determined by paired t-tests, are shown above the bars.

for discrete speech input. Interviews with participants revealed that hybrid commands were felt to be more adequate for small movements and actions requiring fine-tuning. On the other hand, for longer actions, discrete mode was said to be preferable as it does not require maintaining speech throughout the command. This hints towards a possible improvement to hybrid speech input, where both continuous and discrete types of commands could be supported based on a distinguishable trigger.

CONCLUSION AND FUTURE WORK

We proposed a novel voice command method for real-time robot control called hybrid speech input, which combines discrete verbal and continuous non-verbal speech. We presented, implemented and evaluated an example of hybrid speech, in which a command specified by a phrase is carried out for as long as its last vowel is extended. Our preliminary evaluation shows that this approach is promising.

Our future work includes improving our speech recognition implementation to increase accuracy and performance. After achieving a satisfactory level of robustness, we will consider integrating other basic sound features such as pitch to control further command parameters, similar to prior work on continuous non-verbal speech input. We would also like to investigate how hybrid speech could support both start-stop style commands and continuous input as identified in the discussion. Last but not least, we will evaluate our technique for other, more typical, robot tasks such as navigation and picking.

REFERENCES

[1] Jeff A Bilmes, Xiao Li, Jonathan Malkin, Kelley Kilanski, Richard Wright, Katrin Kirchhoff, Amarnag Subramanya, Susumu Harada, James A Landay, Patricia Dowden, et al. 2005. The Vocal Joystick: A voice-based human-computer interface for individuals with motor impairments. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 995–1002.

[2] Masataka Goto, Katunobu Itou, and Satoru Hayamizu. 2002. Speech completion: On-demand completion assistance using filled pauses for speech input interfaces. In *Seventh International Conference on Spoken Language Processing*.

[3] Brandi House, Jonathan Malkin, and Jeff Bilmes. 2009. The VoiceBot: a voice controlled robot arm. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 183–192.

[4] Takeo Igarashi and John F Hughes. 2001. Voice as sound: using non-verbal voice input for interactive control. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*. ACM, 155–156.

[5] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[6] Shizuka Takahashi and Ikuo Mizuuchi. 2017. Operating a robot by nonverbal voice based on ranges of formants. In *2017 3rd International Conference on Control, Automation and Robotics*. IEEE, 202–205.