1E1-06

Design for a System able to use Time-Critical Spoken Advice

Shunsuke Soeda [*1] Nigel Ward [*2]

[*1]Mech-Info Engineering, University of Tokyo [*2]Mech-Info Engineering, University of Tokyo

In collaborative action where a human is interacting with a semi-autonomous intelligent system, sub-second reponsiveness is often called for. This can be done, in part, by extracting the prosody of spoken language advice and using this information to bias the system's selection of its next action. This paper describes such a system.

1. Introduction

As computers get faster and better integrated with the real world, it is possible to envision robots which need to interact with people as they perform real-time tasks. For example, one can imagine a truck-driving system which is mostly autonomous but which sometimes needs directions and advice. The obvious man-machine interface for such a task is to give the user hands-on controls, such as a steering wheel, so he can directly take control whenever he wishes. However, doing so may defeat the purpose of automation, which is presumably to free-up the human to perform other tasks (such as navigating by looking at a map or communicating with other people).

Thus there may be situations where man-machine communication by voice is preferable, as it allows hands-free, eyes-free interaction. While the utility of voice communication is generally acknowledged for non-critical tasks, such as changing the radio station, where inaccurate and delayed responses are tolerable, for critical tasks such as guiding a moving object, the potential utility of voice has not been noted, much less demonstrated. However, people often communicate with each other by voice while performing such tasks; as when a driving instructor guides to a student, or when two men carry a refrigerator together.

This paper describes the first implementation of a system able to do this: able to take advice from a human while semi-autonomously performing a real-time control task.

2. Non-verbal Communication

Voice communication has two components: the content and the non-verbal aspects. In human communication, it is often important to listen to not only what was said, but how it was said. This includes utterance prosody, that is, the timing, energy levels, and pitch patterns. This is especially important in real-time cooperative work.

For example, one day when the second author was with his wife, supporting her as she was practicing driving on city streets, he said *slow down*, in a rather flat voice, as the speed was not truly dangerous, and he didn't want to be overbearing. The response was an immediate lane-change into the left-turn lane. It seemed that she had failed to recognize the words spoken, but had responded to the utterance nonetheless. Probably the problem was that the prosody of the utterance was ambiguous, and she interpreted it as a signal to 'act now', rather than the intended signal to 'be more careful'.

This example shows several things. First, it shows that people do pay attention to prosody in real-time tasks. Second, it shows that responses to prosody can be swift. This is something we have also observed in conversational interaction, where back-channels tend to occur about 350 milliseconds after a prosodic cue by the other speaker [Ward 99]. Third, it shows that prosody can be ambiguous, and is thus not likely to be a full solution to the problem of communication about real-time tasks. Prosody simply fails to distinguish between, say *turn left* and *slow down*, although it can generally distinguish between suggesting that the other person should 'act now' or 'pay more attention' or 'do something different' or 'keep doing what you're doing now' or 'get ready to act' and so on. Fourth, people performing real-time actions generally have their own plans, and spoken advice is interpreted relative to that plan; in this case, the *slow down* clearly only triggered execution of a plan already thought out. This phenomenon was also note by [Chapman 91], regarding advice to an agent (although his agent was performing a non-real-time task, and accepted only advice that was typed in, not spoken).

Another other interesting fact about prosody in communication is that it may require less effort on the part of the speaker, as revealed by utterances like *gyaa!* and *mm-mm-mmm*, which the speaker can produce without the effort, and time delay, involved in retrieving appropriate lexical items. It is also the case that prosody is faster to process. Today's speech recognition systems all require the user to finish a word before they can identify it (indeed, all but Dug-1 [Hirasawa 98] require him to finish an utterance); and even after this point there is a significant lag before the recognition results are available. However prosodic characteristics can easily be computed on-line, with partial results available even before the word is finished.

3. Task Domain

Video-game playing is a challenging real-time task, so we chose this as the domain for our agent. While there are video games where two players play together, we found these too complicated to tackle. Instead, we chose a very simple action game, `xlander`, which is a lunar landing simulation under the X window system. We added a cooperative dimension to this by dividing the control of the spacecraft:

Contact: Shunsuke Soeda, Mech-Info Engineering, University of Tokyo, shnsk@sanpo.t.u-tokyo.ac.jp

Table 1: Results of corpus analysis

Category	Meaning	Example (English translation)	Reaction of person pretending to be the agent			
			Try to rise	Try to rise slightly	Did not try to rise	Total
Act	Rise	"agatte" (go up)	**14**	8	3	25
	Fall	"sagatte" (go down)	3	1	**0**	4
Act slightly	Rise slightly	"chotto agatte" (go up a bit)	2	**8**	1	11
Keep	Keep the state	"sonomannma" (keep it that way)	**0**	**0**	**2**	2
	Nothing specific	"yoosi ganbaruzo" (Hmm, I'll do better this time)	1	0	**5**	6
Panic	Avoid crash	"ahhh!" (ahhh!!)	**2**	0	0	2
Total			22	17	11	50

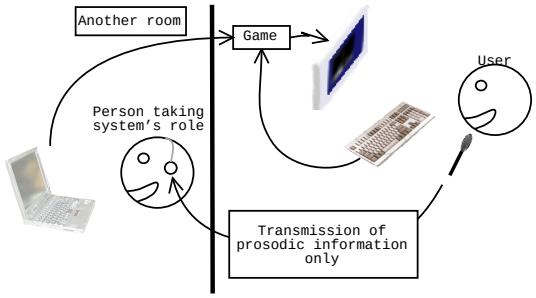

Figure 1: Human-human communication experiment

assigning the horizontal control — forward, backward, left and right — to the human player (user) and vertical control — thruster on/off — to the agent. In order to make communication essential, we gave access to key information only to the user; he therefore had to make decisions based on this information and convey those decisions to the agent. Specifically, the game screen information was provided only to the user, thus the agent had no direct access to information about the state of the spacecraft.

4. Analysis of Human-human Dialogs

We collected a corpus human-human communication in Japanese by having two subjects playing `xlander` together, one acting as the user, other pretending to be the agent (Fig. 1). We filtered the user's voice by Linear Predictive Coding in order to transmit only prosodic information, the pitch and the energy, of the user's voice. The person taking the agent's role had no other clues about the game than this non-verbal information. We collected 50 utterances in this way. Roughly 66% of them were interpreted correctly; thus using only non-verbal information, the person was often able to infer the other's intentions.

We then set out to analyze this data. Our first observation was that the prosody alone did not seem to convey which action was desired. Thus there did not seem to be any prosodic pattern meaning anything as specific as *turn thruster on*, for example. (There was a partial exception in that the typical pitch-patterns of the most common words, such as *agatte* and *sagatte*, were sometimes identifiable, but we chose not to use such lexical stress information.) What the prosodic patterns did seem to be conveying was at a more abstract level, pertaining mostly to the degree to which the other person was doing the right thing or not.

Thus we classified the utterances as seen in table 1. This classification is in some respects arbitrary, but it mostly succeeds in grouping together utterances with similar effects on the hearer, and it mostly succeeds in grouping together utterances with similar prosody.

5. The Agent as Implemented

Outline of the agent

Based on the analysis mentioned in the previous section, we build a semi-autonomous agent. The agent is made of two modules, the *sound analysis module* and the *decider module*. The sound analysis module takes raw sound data as input and sends the decider module one of four signals if it detects a prosodic feature matching a rule. The decider module takes the signals as input and controls the vertical thrust of the spacecraft. It is a finite state automaton with five states, changing its state in response to incoming signals and to time-outs.

The sound analysis module

The sound analysis module uses a window width of $180ms$ to calculate the slope of the energy and pitch every $10ms$. It first tries to match the "panic" rule by seeing if the slope of the energy is rising steeply and if the sound analyze module has not recognized any rule for $100ms$ (2). The next rule checked is the "act slightly" rule, which is recognized by a steep fall of the pitch. The third rule is the "act" rule, whose prosodic feature is a pitch rise. The "act slightly" rule and the "act" rule are only matched if $500ms$ has passed since the previous rule condition was recognized. If none of these rules are matched, but the sound analysis module recognized a voice for $400ms$, the sound analysis module interprets the voice as a "keep" rule.

When any of these four rules are recognized, the correspond-

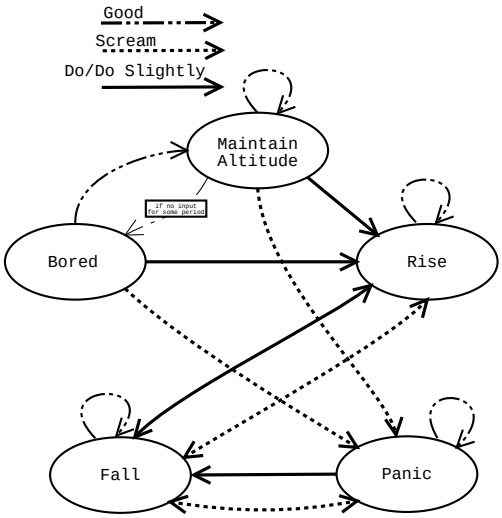

Figure 2: The decider module

ing signal is sent to the decider module.

The decider module

The decider module is a Finite State Automaton with five states(2).

- **Maintain altitude state** This is the default state. The decider module tries to control the vertical thrust to maintain the vertical speed (turns on the thruster for 11% of the time).
- **Bored state** The decider enters this state when the user gives no advice for $3000ms$ while in the "maintain altitude" state. The decider thrusts sparsely(2% of the time), and the spacecraft slowly accelerates downward. This state was inserted to encourage the user to give advice to the agent.
- **Rise state** This state is invoked by a "act" signal or a "act slightly" signal when the decider was in a "fall", "bored" or "maintain altitude" state. The decider thrusts often (33% of the time) so that the spacecraft accelerates upward.
- **Fall state** In this state there is no thrust at all, so spaceship is pulled down by gravity. The decider enters this state when it gets a "act", "act slightly" or "panic" signal in the "panic" or "rise" state.
- **Panic state** When the decider gets a "panic" signal and the state is not "rise" or "panic", it enters the "panic" state, and thrusts continuously while in this state.

The "panic" state has a duration of $500ms$, and transitions to the "maintain altitude" state after that. The "rise" state and the "fall" state have durations depending on how the decider entered that state. If it was on the "act slightly" signal, then the duration is $500ms$, and if it was on the "act"signal, the duration is $1500ms$. The "keep" signal is interpreted so that the duration of the entered state becomes $1500ms$; but a keep signal in the "bored" state, leads to the "maintain altitude" state.

Table 2: Prosodic features used by the sound analysis module

Rule name	Prosodic feature	Conditions of the rule
Panic	steep rise in the energy	$mEnergy > 8.0$
Act slightly	steep fall of pitch	$mlogF_0 < -0.002(Hz/ms)$
Act	rise in pitch	$mlogF_0 > 0.0008(Hz/ms)$
Keep	default rule	if no rule matches for $400ms$

6. Evaluation

We had three subjects (two male, one female) use the agent. They had no knowledge about how the system works. We collected 80 utterances (3), and analyzed them. For comparison, we also had them interact with a human being in the agent role, who was this time allowed to use verbal information.

The agent correctly responded to 40% of the utterances; which is to be compared to the 66%, achievable by humans using only non-verbal information. Compared to the human game-player, subjects considered the agent to be inferior in that it often failed to interpret orders, and in that its responses were too slow.

One problem subjects noted was the intrinsic difficulty of the task; landing spacecraft is a very hard task for most people, let alone those who have no access to altitude information, so our choice of `xlander` as a task for our experiments was probably not optimal.

Subjects also noted that the experiment set-up made them tense, and corpus analysis showed that their utterances contained less prosodic information than in the human-human dialogs. This was a reason that many utterances which were intended as commands to take action were interpreted as commands to keep in the same state.

Subjects also showed a tendency to give orders, rather than advice, to the system. Perhaps they did not recognize it as an autonomous intelligence. This led to inevitable failure, as in the case where the spacecraft kept rising even though the subject repeatedly told the agent to go down.

After the experiment was over, we explained the design of the agent to the subjects and let them try again. The generally did much better in this case. It is also worth noting that the first author was able, after some practice, to use the system to gently land a spacecraft. Clearly it is important in man-machine communication, just as in human-human communication, to have an understanding of the language abilities and reasoning abilities of the partner.

Table 3: Results of evaluation experiment

		Act rise	Act fall	Act slightly	Keep	Panic	Total
Act	Rise	16	0	4	4	0	24
	Fall	1	3	6	8	0	18
Act slightly		2	0	2	6	0	10
Keep		3	0	3	17	0	23
Panic		3	0	0	1	1	5
Total		25	3	15	36	1	80

7. Conclusion

Our system achieved a recognition rate of 49%, in a task where human can achieve 66%, suggesting that it is possible to make use of non-verbal information when understanding spoken advice in time-critical situations.

Ultimately a system like this should chose its next action based on (at least) 3 kinds of evidence: verbal inputs, non-verbal inputs, and its own judgment of what it should do next based on the context. The best way to combine all of these is, of course, to formulate treat each kind of evidence in probabilistic terms. Doing so would allow the system to select, at each moment, the hypothesis (action) which has the highest overall probability. In the models presented in this paper, neither the context model nor the use of non-verbal information is handled probabilistically; thus both need to refined before a fully integrated system can be built.

Acknowledgments

We thank the International Communication Foundation and the Japanese Ministry of Education for support.

References

[Chapman 91] Chapman, D.: *Vision, Instruction and Action*, Massachusetts Institute of Technology Press (1991).

[Hirasawa 98] Hirasawa, Jun-ichi M. N., Noboru Miyazaki and Kawabata, T.: Implementation of Coordinatice Nodding Behavior on Spoken Dialogue Systems, in *International Conference on Spoken Language Processing*, pp. 2347–2350 (1998).

[Ward 99] Ward, N. and Tsukahara, W.: A Responsive Dialog System, in Wilks, Y. ed., *Machine Conversations*, pp. 169–174, Kluwer (1999).