

# A Wearable Cross-language Communication Aid \*

Jani PATOKALLIO  
jani@sanpo.t.u-tokyo.ac.jp

Nigel WARD  
nigel@sanpo.t.u-tokyo.ac.jp

*HCI Laboratory, Mechano-Informatics, Engineering, University of Tokyo  
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan*

## Abstract

*This paper presents a wearable device, the Yak, that aids cross-language communication. The Yak produces utterances in the native's language at the user's command. The use of a heads-up display allows the user to operate the device while maintaining eye contact with the native, and while sending and receiving non-verbal signals. We believe this interaction paradigm will be valuable for many face-to-face encounters, for example as travelers deal with people in service roles, such as ticket agents, waiters, clerks. This paper describes requirements for a communication aid, the Yak's user interface, the Yak's current hardware configuration and scripting language, and the results of preliminary experiments.*

## 1 Introduction

Imagine that Beth has just arrived at the Mahale airport. Leaving the airport, she loads a Rutungu interaction module into her wearable computer, and goes over to the train ticket counter.

While waiting in line she clicks thru to reach the train-station phrase menu, then reviews the options available.

When she reaches the agent, she smiles and clicks to launch the standard introduction, the Rutungu equivalent of

“Hello. I do not speak Rutungu, so I will talk through this interpreting device. Is this OK?”

The ticket agent looks surprised for a second, then looks troubled. As he says something curt with an

abrupt hand gesture, Beth pushes forward her map and launches the next phrase.

“Thank you. I would like a round trip ticket to here”

says the machine, as Beth points to the city she has circled. The agent again says something curt and turns to his listing. Beth waits. In a moment he comes back and says something she which interprets as a question. Looking confused and apologetic, she launches:

“I'm sorry, my machine can only translate one way. Is a ticket available?”

The agent looks annoyed, then has an idea. He holds his fingers to his lips and makes a puffing motion, then raises his eyebrows and points at Beth. Beth clicks through two menus and launches:

“No-smoking, if available, please”

The agent says something as he turns away, but Beth decides not to pursue it; and in a moment he comes back with a ticket. He points to the price, and as Beth finishes paying, she launches:

“Which platform does the train leave from?”

The agent answers while gesturing a path with his finger.

“Could you please write that down for me?”

The agent scribbles the number 6 on the back of the ticket. Beth takes it, smiling her thanks as her machine says.

“Thank you very much.”

Studying the ticket, Beth notices she has 10 minutes before the train leaves. She finds what looks like a noodle stand in the corner, calls up the restaurant script and scans the menus to pre-load the phrases she will need to order curry without yak meat, and then walks over to have lunch.

---

\*This work was supported in part by the International Communications Foundation. We thank Yusuke Shinohara for his assistance. Some technical details are at <http://www.sanpo.t.u-tokyo.ac.jp/~jani/yak/>.

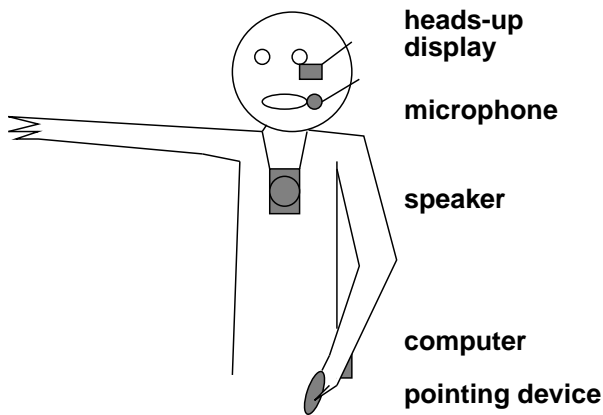


Figure 1: The Yak concept

## 2 Why a Heads-up Translation Device?

In normal human-to-human conversation, non-verbal cues are important. Gestures indicate where things are, facial expressions indicate degree of understanding, posture indicates attitude, tone of voice indicates invitation or query or request, and so on. In cross-language communication also, it is a common experience that gestures can get you a long way, even if no words are understood.

We believe the use of a wearable device with a heads-up display (Figure 1) can allow a person to communicate with someone in another language better than they could by using a translation device which they have to look down to use.

Moreover, given the limitations of speech recognition technology, there is no possibility of building a translation device capable of general bi-directional translation for at least a decade or two. Given this, it is essential to engage the non-verbal communication skills of the participants.

Thus the proposal is to only partially automate the communication process: to exploit the strengths of both man and machine to produce a hybrid solution to the problem of communicating across a language barrier. This was seen in the above scenario, where there was a division of labor between Beth and her translation device. The device actually output the sentences, but Beth was responsible for everything else: deciding when to launch each utterance, using smiles and gestures to elaborate on the utterances, and interpreting the agent's utterances, gestures and actions.

This paper describes the Yak, a wearable system built to explore usability issues in heads-up communication aids.

## 3 Related Research

The Diplomat/Tongues project at CMU [1, 2] was designed to “explore the feasibility of creating rapid-deployment, wearable bi-directional speech systems”. The focus of the effort was on the technical feasibility of speech-to-speech translation, specifically the three component technologies — machine translation, speech recognition, and speech synthesis. Questions of usability, have not yet, it seems been seriously considered. Rather, the inadequacy of current speech and language technologies led the system designers to cast the users in a supporting role: the users are enlisted in the task of preventing errorful translations (although this sort of “interactive editing” has not been well accepted in other machine translation applications [3]). Performing this role requires both users to have access to a GUI running on a laptop, requiring a style of interaction which diverges from the more common visions of wearable computer use.

Recently the LingWear “mobile tourist information system” at Karlsruhe has been reported as including a “translation module” with spoken output and perhaps input [4], but no details on the design have yet been published.

Hand-held translation aids are another area of research activity. Descended from the venerable phrase book (Berlitz, etc), and electronic equivalents such as the Canon WordTank, systems have recently appeared with speech output. More recently, a hand-held portable electronic dictionary with speech input was proposed by [5].

There are also systems which produce speech output for people who are unable to speak due to neurological or other reasons.

Other, non-wearable research in speech translation systems (including those in the CStar, ATR, Verbmobil [6] and CU Communicator [7] projects) seems to involve one of two usage scenarios. One is telephone conversations, which are, of course, a special case in that non-verbal communication is not present. (Although there have been preliminary investigations of the possibility of also transmitting facial expressions, while translating lip movements [8].) The other common scenario is business meetings, where the parties are assembled around a table in a conference room conducting negotiations, which is also a challenging scenario in that the topic of discussion is something abstract like a proposal or plan or schedule. Most research on both these scenarios assumes that all communication is done via the communication device.

Although not directly related to the current paper, another function that a wearable for travelers should



Figure 2: Our initial prototype was scary

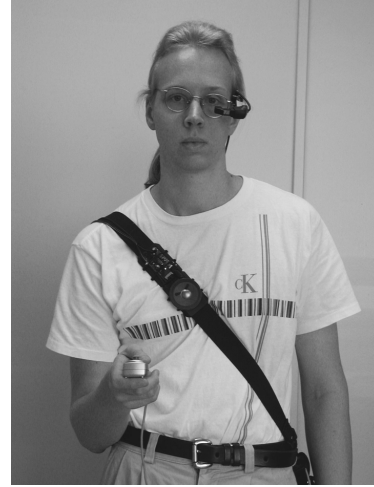


Figure 3: Wearing the current version

include is automatic sign translation [9].

Thus the possibilities for heads-up wearable translation have not previously been investigated.

## 4 Test Domain

As an initial domain for heads-up cross-language interaction, we are targeting dialogs in airports, train stations, restaurants, shops, hotels, and on the street. The simplicity of these dialogs makes it possible to build a working and usable system today. However we believe that the findings regarding usability will generalize to future systems which also support hybrid solutions to the problem of communicating across a language barrier.

These domains have two useful properties. The first one is that translation does not have to be bi-directional: the probable responses of the native are few enough in number that the user can generally classify a response based on non-verbal information alone. The second useful property is that the number of things the traveler will need to say are finite.

## 5 User Interface Design

This section raises some basic issues and explains how our design addresses them.

**Non-Threatening** The first consideration is, of course, that the system be not so scary that natives refuse to approach. We found this out with our first prototype (Figure 2), which used a Sony Glasstron PLM-A35 wearable display and a Sony VAIO C1 running Red Hat Linux (inside

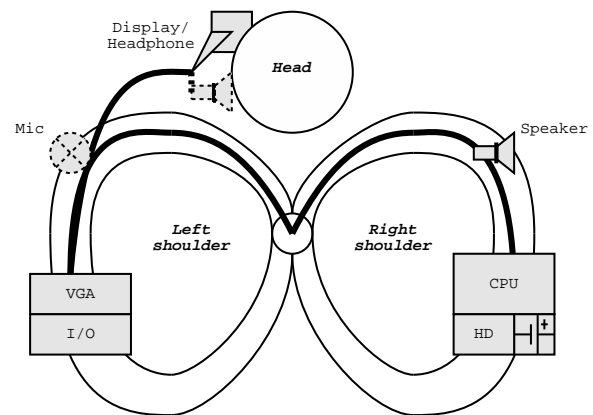


Figure 4: The harness-mounted Yak-2

the user's backpack). The stock Glasstron is an immersive display intended for movies and computer games and unsuitable for wearable use as is, so it was modified into a monocular display by removing the electronics for the right eye and cutting a hole in the cover. Apparently the futuristic metallic blue Glasstron was threatening: as seen in the picture above, even a fellow lab member kept a distance of about two meters while talking to the author wearing it.

**Easy to Wear** Initially we used a back-pack configuration, but then switched to having the system built into a shoulder bag. This streamlined the cables, provided a convenient pocket to store input devices while not in use and made the system much easier to take on and off.

In future, we want to get rid of the shoulder

bag and make the system lighter. The key will be to squeeze all electronics and connecting cables onto a holster (Figure 4), worn like a vest atop—or under — the user’s normal clothing, in the style of MIT’s MIThril [10] vest. For this we plan to use eHolster’s Ballistic Nylon support system, with cabling running along the harness and the electronics contained within pouches. Also, instead of the laptop we plan to use a smaller computer, probably the Advantech PCM-5822 miniature motherboard with a 200 MHz Pentium-compatible processor and various useful connectors. Power supply will be a problem: we need something lightweight, compact, stable, and long-lasting. We plan to craft our own, using two of Sony’s lithium-ion batteries (which are, incidentally but conveniently, identical to those used by the MicroOptical displays) with a custom-built regulator board to adjust the voltage.

**Easy to See** The Glasstron was of course poor as a computer display, being heavy, bulky, and with a narrow field of view and poor-quality picture. For the experiments reported below, we used a MicroOptical CO-3 eyeglass-mounted display, with which we were quite happy. The display takes up less than a ninth of the field of view and can be adjusted to any position, although most right-handed people will opt for the lower left corner.

The content of the display was also an issue. Since we wanted the system to be usable outdoors in visually cluttered locations, we decided to keep menus simple (currently no menu has more than 5 items at a time), to show all text with large (50 pt), high-contrast letters, and to provide very obvious visual feedback through color and highlighting.

**Easy to Hear** Since the device needs to produce only a finite number of utterances, we chose to use pre-recorded samples, as being more intelligible than text-to-speech output. For the experiments reported below, the samples were recorded by a young male native speaker of Japanese with a deep voice, compatible with the voice one would expect from the user. Utterances were recorded in a slightly flat voice, on the assumption that warmth, “color” and animation would come from the user’s smiles and gesture.

For audio output, we used a speaker powered by two AA batteries, adjustably mounted at chest level on the shoulder strap of the bag. This provided sufficient volume, even when used next to a

busy highway, but was a little too loud for inside use. Ideally the system should adapt its output volume automatically based on the ambient noise level.

The device’s utterances were designed to constrain the likely responses of the native and to encourage him to provide non-verbal cues (“point”, “show”, “draw”, etc.).

**Reliable** Since the device only needs to handle a finite number of utterances, “translation” was merely the lookup of the Japanese utterance corresponding to the selected English menu item (done within Yascript, below).

This choice meant that we did not have to do compositional translation, that is, we avoided the error-prone process of having a system build up the translation of a full sentence by doing syntactic and semantic analysis of the input and translating each word and then assembling the results.

**Easy to Operate** Regarding the question of how the user specifies what she wants the machine to say, we initially considered using a keyboard. However the difficulty of text entry on today’s wearable keyboards led us to reject this option.

We also considered using speech input, that is, having the user speak in her native language and having the machine recognize the utterance and output the target language equivalents. With an unrestricted vocabulary the speech recognition rates would obviously be unusably low, but we also found that even with a small vocabulary (of a dozen items, corresponding to the items currently shown a menu), that mobile speech recognition was too error prone, even indoors. We were also somewhat concerned that natives would be confused as to whether the user was directing her utterances to the system or to them.

So we decided to restrict the input modality to selections among choices on menus (see Figure 6). A number of different input devices were tested, including a normal USB mouse (difficult to move the pointer accurately) and a FinRing wireless mouse with tilt sensor (difficult to use). We finally settled on Victor’s Handy Mouse, a USB mouse in the form of a joystick-type pointer specifically designed for one-handed use.

Each decision the user has to make is presented as a menu: the user can scroll up and down with the mouse (only one degree of freedom needed) and press the *select* button to choose. At any stage

the user can press the *cancel* button to backtrack (when pre-cueing) or repeat (if in an interaction), or select the *abort* option to stop immediately and return to the initial menu. This was simple enough for the user to do quickly, reliably, and without looking at her hand.

To further ease the selection task, we organized the inventory of phrases hierarchally. First the options are organized by task (buy train ticket, buy noodles, etc.), and then temporally by screen within each task. For example, the buy-noodles task leads to the select-dish screen, after which follows the select-drink screen, and so on.

Deciding on the appropriate phrases and organizing them is a fair amount of work. To build a full device like this it will probably be more efficient to license the contents of some existing phrase-book.

**Fast** We also decided to take advantage of the fact that exchanges like these typically can be seen as the execution of some plan. Thus we moved as many decisions as possible into the planning phase. For example, if you are going to ask someone on the street for directions, first you have to decide where you are trying to go, and then you have to decide whether you want him to try to show you on a map. Thus in each task, there were one or two initial parameter-setting screens. In the ramen-ordering task<sup>1</sup>, the parameters are what you want to eat and what you want to drink.

Having the user make these decisions off-line, before entering into an interaction, has two advantages. First it reduces the cognitive load on the user during the conversation, so she can concentrate on listening and on looking friendly. Second, it lets him produce follow-up responses quickly enough so the native doesn't get frustrated.

## 6 System Internals

Our system — dubbed the Yak as a weak pun on *honyakki* (Japanese for translation device) and the English verb *to yak* — is a fairly straightforward implementation of the design sketched out above.

The software of the devices is a user interface/display engine entitled Yakkey and a corresponding scripting language called YakScript. Both script

<sup>1</sup>this may seem absurdly specific, but in Japan ramen shops have ordering and paying conventions, as well as vocabularies, which are different from those in sushi restaurants, family restaurants, bars, and so on

```
#
# Ask for directions to $location
#
finder.init=
  Play finder_$location;
  Pause $SELECT if they understood,<P>
    $CANCEL to repeat question;
  FALSE : Goto finder;
finder.title=Ask directions
finder.action=
Repeat slowly (Blank Asking...;
              Play finder_repeat)
Point direction (Blank Asking...;
                Play finder_point)
Show on map (Blank Please show the map.;
            Play finder_showmap)
Draw a map (Blank Please offer...\
           Play finder_drawmap)
Thanks & goodbye (Blank Thanking...;
                 Play finder_thanks;
                 Load index.script)
```

Figure 5: An example of YakScript



Figure 6: A user's view of Yakkey (simulated)

interpreter and display engine were written in Java, allowing rapid development and portability to both Windows and Linux environments. This threefold division into program logic (individual scripts for each particular environment), program control (the YakScript interpreter) and user display and input (the Yakkey interface) also allowed the system to evolve smoothly as new requirements were discovered, e. g. the need for the user to backtrack in the script or more advanced conditionals in the script.

A YakScript script consists of a sequence of menu *frames*, which in turn contain pairs of *options* and *actions*. Options are bits of HTML text used to build a display. When an option is selected by the user, its corresponding actions are performed. As an example, the code in Figure 5 generates the display shown in Figure 6. The language supports branches to other frames, variable substitution in both options and actions, conditional execution and return values, but little else; should the need arise, external programs can always be invoked with the *Exec* command.

YakScript as a programming language is rather conventional, although its internal design has some more unusual features. The YakScript interpreter was written in such a way that each action (such as *Set*, *Goto*, *Play* is its own Java class. All actions inherit from the abstract *Action* class and can inherit each other as well, e. g. *Sub* is a subclass of *Goto* that simply throws all current state onto the stack before invoking the parent.

In designing YakScript we left hooks so that it could be extended to handle speech input, in the spirit of SpeechWear [11], VoiceXML [12] and the nascent Multimodal Dialog Markup Language (MMDL) [13].

## 7 Preliminary Experiments

At time of writing the device has been operational for just over a week. So far we have tested the device outside the lab only informally.

Our “user” for these experiments was the first author, who is, moreover, fluent in Japanese. Therefore we are unable to assess yet how usable the device will be for people dealing with a language of which they have no knowledge. The experiments were nevertheless informative.

The first task we set the user was to ask a passer-by for the location of a *ramen* (noodle soup) shop. We did the experiments just outside the university, on the sidewalk of Hongo Avenue, in Tokyo.

The natives the user approached were for task 1 were: two individuals who were leaving the university

entrance, one couple sitting outside, and one person entering a subway station. Of these, 2 people, when accosted by the device, looked surprised, changed their path, and walked away. 2 people seemed momentarily surprised, but then entered into the conversation. The surprise may have been due to the appearance of the head-mounted display, or to being greeted by a machine (the speaker sound quality was fair, but clearly not a normal human voice, and in addition the speaker is mounted at a fair distance from the user’s mouth).

Video clips of two of these interactions are available at the <http://www.sanpo.u-tokyo.ac.jp/~jani/yak/video/>.

In both of the conversations which occurred the user was able to successfully obtain directions, in one case with the help of a map. Holding the map at the same time as the mouse was a little awkward but not impossible.

We found that it was indeed possible for the user of the device, while operating it, to maintain eye contact and gesture with one hand (point directions or draw a map). However the interactions were not as natural as normal face-to-face conversations. One problem is that giving instructions to the machine requires focusing on the display for a moment, resulting in, from the native’s viewpoint, a traveler with vacant, expressionless face.

Interestingly, the user felt unnatural not speaking himself, as if he had to suppress the reflex to speak with his own voice.

In the first run, after an initial greeting (“Hello, excuse me”), the user launched an explanation (“I do not speak Japanese, so I will talk through this interpreting device. Is this OK?”). However we felt that this was not necessary and later did without it. Certainly the function of the device was fairly obvious (or perhaps there is a reflex response to respond when spoken to, even by a mechanical voice).

When faced with Yak, the natives spoke fluently and without hesitation in their own language. Thus, the system seemed to make it easy for people to deal with the traveler. Perhaps they were assuming that the translation device was also capable of translating the other way too, or perhaps this too was a reflex response (to the fluent utterances produced by the machine). With real users this may be a problem — fluent native utterances may come with gestures which are too swift and subtle to detect. If this is true, we may need to have the machine produce utterances which are slightly halting and ungrammatical. (Of course, the user can always launch an explicit request “please speak slower”, but that is disruptive.)

In human-human conversation, when two people are speaking the same language, turn-taking is swift, but in cross-language conversations the flow is not quite so smooth. This was seen for the Yak too. In principle it should be possible for the user to launch things at the exact instant when the other finishes speaking. But in practice there were some awkward pauses while operating the device, although this only approached the level of annoyance on one occasion when the mouse got stuck.

The second task was to go in a ramen shop and have lunch. Unfortunately, we could not videotape this conversation, so we had to rely on notes from a confederate on the scene. In this case the waitress showed little surprise or interest in the device, and just took the order without comment. At one point an utterance by the device overlapped with the response by the waitress — in human-human conversations this generally does not happen, since people naturally signal non-verbally when they are about to speak, and over-talk is resolved quietly when one party backs off to let the other continue. However, the Yak currently has no provision for aborting an utterance in progress.

The Yak was easy to carry around and stays out of the way while not in use, but the weight of the eyeglass-mounted display did cause some strain on the ears.

## 8 Future Work

Upon the completion of the construction and testing of the Yak-2, we plan to test the following hypotheses:

- 1 People will be receptive to communicating with a person whose utterances are produced synthetically.
- 2 In many dialog situations, it will be possible to communicate even if the utterances of only one party are translated.
- 3 Users can produce gestures and facial expressions that are synchronized with and usefully supplement synthetic utterances.

The hypotheses will be tested by asking users equipped with either the Yak-2 or a reference system to complete predefined tasks, such as ordering a vegetarian meal or purchasing a train ticket to a given destination. Reference systems will include a printed phrasebook, an electronic phrasebook, a remote over-the-phone human interpreter, and a physically present

human interpreter. Based on quantitative data collected by the device itself and qualitative feedback provided by the user, it should be possible to evaluate the hypotheses and assess the strengths and weaknesses of the system, and heads-up communication aids more generally.

## References

- [1] Robert Frederking, Alexander Rudnicky, and Christopher Hogan, “Interactive Speech Translation in the Diplomat Project,” *Machine Translation*, to appear.
- [2] Robert Frederking, Alexander Rudnicky, and Christopher Hogan, “Interactive Speech Translation in the Diplomat Project,” in *Spoken Language Translation Workshop. 1997*, ACL.
- [3] Nigel Ward and Dan Jurafsky, “Machine translation,” in *Speech and Language Processing*, Daniel Jurafsky and James H. Martin, Eds., pp. 720–751. Prentice-Hall, 2000.
- [4] Christian Fuegen, Martin Westphal, Mike Schneider, Tanja Schultz, and Alex Waibel, “LingWear: A Mobile Tourist Information System,” in *HLT 2001 preliminary proceedings*, 2001, pp. 373–377.
- [5] Yasunari Obuchi, Atsuko Koizumi, Yoshinori Kitahara, Jun’ichi Matsuda, and Toshihisa Tsukada, “Portable speech interpreter which has voice input and sophisticated correction functions,” in *Eurospeech99*. 1999, pp. 2023–2026, ESCA.
- [6] Thomas Bub, Wolfgang Wahlster, and Alex Waibel, “Verbmobil: The combination of deep and shallow processing for spontaneous speech translation,” in *ICASSP-97*, 1997, pp. 71–74.
- [7] B. Pellom, W. Ward, J. Hansen, K. Hacioglu, J. Zhang, X. Yu, and S. Pradhan, “University of Colorado Dialog Systems for Travel and Navigation,” in *HLT 2001*, 2001, pp. 345–350.
- [8] Shin Ogata, Satoshi Nakamura, and Shigeo Morishima, “Multi-modal translation system: Model based lip synchronization with automatically translated synthetic voice,” in *Interaction 2001*. 2001, pp. 203–210, Information Processing Society of Japan.
- [9] Jie Yang, Jiang Gao, Ying Zhang, and Alex Waibel, “Towards automatic sign translation,” in *HLT 2001*, 2001, pp. 269–274.
- [10] MIT Media Lab, “Mithril, the next generation research platform for context aware wearable computing,” <http://wearables.www.media.mit.edu/projects/wearables/mithril/>, 2001.
- [11] Alexander I. Rudnicky, Stephen D. Reed, and Eric H. Thayer, “Speechwear: A mobile speech system,” in *ICSLP 96*, 1996, pp. 538–541.

- [12] L. Boyer, P. Danielsen, and J. Ferrans, “Voice eX-tensible Markup Language (VoiceXML) 1.0,” Tech. Rep., W3C : World Wide Web Consortium, 2000.
- [13] T. V. Raman, B. Lucas, P. Kapanen, et al., “Multimodal requirements for voice markup languages,” Tech. Rep., W3C : World Wide Web Consortium, 2000, working draft.