HUMAN MACHINE INTERACTION BY VOICE AND GESTURE

Nikil Jayant

Multimedia Communications Research Laboratory
Bell Laboratories, Lucent Technologies

ABSTRACT

Voice and gesture represent fundamental and universal modalities in interhuman communication. With recent advances in automatic methods of speech recognition and synthesis, human-machine interaction by voice is rapidly becoming a technological and commercial reality. Although less mature and deployed, gesture recognition by machine is becoming reliable enough to be considered as a serious supplement to the voice interface between humans and machines.

1. VOICE AND GESTURE

Among the several forms of human-machine interaction that are currently being researched, voice- and gesture-based methods stand out because of their fundamental and universal roles in interhuman communication. Machines that understand speech and respond by voice are becoming slowly but surely a part of multimedia communications technology; and in the context of small vocabularies, gesture recognition by machine is reaching levels of reliability and robustness that are reminiscent of the initial milestones in voice recognition about a decade ago.

Section 2 of this paper presents a high-level summary of the current state of human-machine interaction by voice, and demonstrates typical current capabilities with the example of a *voice-activated* web browser.

Section 3 of the paper describes a specific example of gesture recognition, characterized by robust and reliable understanding of visual messages out of a small vocabulary, as conveyed through the human hand and its fingers. The visual recognition technology is demonstrated by the example of *gesture-activated* scene-browser.

2. HUMAN-MACHINE INTERACTION BY VOICE

Following about two decades of serious work by the research community, automatic speech recognition (ASR) and text-to-speech synthesis (TTS) are beginning to cross thresholds of performance in a way that is producing several compelling applications of automatic speech recognition and understanding (ASR), voice-response based on text, without real-time human intervention (TTS), and two-way dialogue between human and machine (combination of ASR and TTS)

2.1. Automatic Speech Recognition (ASR)

Figure 1 depicts the capabilities in ASR along the dimensions of speaking vocabulary, fluency and environment.

An example of a simple ASR application is the situation of a 2- to 5-word vocabulary (such as *yes - no* or *collect, person-to-person, . .*), with words or phrases articulated carefully and in isolation, and in an acoustic environment characterized by moderate to high levels of signal-to-noise ratio. Call-routing in the telephone network is a typical and quite successful application of this class of ASR technology. At this time, there is no fundamental reason why this ASR application cannot be ubiquitous and speaker-independent, except in the most unfriendly of acoustic environments and speaking disciplines. A soft-failure mode, with lapse-back to a manual telephone operator, makes the application particularly acceptable.

At the other end of the ASR-acceptability spectrum is the scenario of a high-strung, fast-speaking, super-vocabulary human pacing around on a noisy convention floor with a dictation machine, expecting its text-output to be almost totally error-free. Even if we ignore the computing and memory limitations of the machine that may be at the disposal of this human, we do not currently have the knowledge to perform the above ASR task, or even many simpler versions of it, with technical decency or commercial viability.

As readers of this summary have no doubt observed by now, there are several challenging and yet realistic applications of ASR that lie between the extreme scenarios in the last two paragraphs. Typical of these intermediate applications are those for automated transaction processing. In these applications, vocabularies

Copyright 1997 IEEE

can be moderate (several hundreds of words). However, the speaking style presents at least three types of serious challenge. First, the articulation is not in the form of isolated words. Second, the level of speaking discipline is typically low, and is characterized by the inclusion of non-speech sounds and disfluencies. Finally, the user tends to present concepts to the ASR system, rather than carefully thought out words or phrases. It has taken the very best combined efforts in signal processing, speech science, and modeling of context and grammar, to provide experimental speech-understanding systems that are beginning to be useful assistants for transaction-processing.

Another class of challenging applications are those in which humans use special speaking discipline to make sure that dictation tasks necessary for an efficient work style proceed with acceptable levels of performance.

Finally, tending toward the simpler small-vocabulary situation, but nevertheless very significant from the commercial viewpoint, are applications of ASR such as hands-free name-dialing in an automobile environment and hands-free control of personal information devices such as telephone answering machines.

Conditioning of the acoustic environment is crucial to all applications of ASR. Directional microphone systems and intelligent microphone arrays serve to reject extraneous noise and reverberation, thereby maximizing the speech quality at the input of the ASR system, and consequently its accuracy.

2.2. Text-to-Speech Synthesis

The elimination of human-generated speech from the voice-response process provides the valuable flexibility of speaking essentially unrestricted text as demanded by the transaction at hand. (Only the fairly trivial of situations permit generating a large number of possible responses by splicing together sub-responses of human speech; even in these cases, the overall quality of the response is compromised in general by the unnaturalness inherent in the splicing process.)

As in the case of ASR, most of the progress in TTS technology has occurred in the last one to two decades. Currently, the best-designed systems offer an excellent level of pronunciation accuracy at the local levels of phonemes or even words. Prosody and intonation models, usually created by expert systems trained from a large corpus of human speech, are constantly improving but are still not robust or universal enough to produce high overall quality at the long-phrase, sentence and paragraph levels.

On the other hand, our ability to customize the TTS voice to user preferences has provided a parallel dimension of performance which, together with accurate pronunciation (of common words, as well as key elements such as proper names and abbreviations), is raising the acceptability of some TTS systems to hitherto unprecedented levels. Finally, in several laboratory experiments, the addition of a text-driven agent face has provided higher subjective acceptance of TTS voice.

An often-cited "killer application" of TTS is in the application of a navigational agent such as a GPS-aided system for talking an automobile driver through a maze of highways or city-streets in a hands-free, eyes-free, map-free fashion. Less demanding applications include reverse-directory assistance and stylized machine prompts in transaction processing.

The language models and pronunciation dictionaries in TTS are also needed in realizing efficient ASR systems. Several TTS systems have multilingual capability and their language models are directly applicable to multilingual ASR.

The combined working of ASR and TTS is an obvious requirement for many classes of two-way human-machine dialog, and for automatic language translation. There is great interest in developing pragmatic, domain-specific examples of such systems.

2.3. The voice-activated web browser

In this (videotaped) demonstration, the user employs speaker-independent ASR to navigate through a book of web pages. Voice commands serve to perform operations such as page-flipping, page-advance and returns to home-page in a keyboard-free style. Voice commands also serve to branch out into one of several new pages based on corresponding textual labels or cues that explicitly appear on a current page. The user-acceptability of ASR in voice-activated web browsing tends to be heightened by the fact that although the ASR grammar is page-specific, implying moderate levels of vocabulary and complexity, the user of the book of pages gets the perception of a large-vocabulary ASR system. Feedback to the user, in the form of TTS, further enhances the overall appeal of the voice interface. The greatest payoff of the voice-activated browser is in the context of mobile, portable computing with a 'webphone' without a conventional keyboard.

3. MACHINE RECOGNITION OF HAND GESTURE

We now describe a specific example of human-machine interaction by gesture. The example is limited in that it relies only on gesture recognition, and does not incorporate gesture-synthesis by machine. It is also spe-

Copyright 1997 IEEE

cific in that we consider the particular, albeit important subclass of hand gestures.

Perhaps the simplest form of hand gesture recognition is one where the vocabulary consists of numbers *one* through *five*, as indicated by the number of clearly straightened and isolated fingers. Indications of fractional counts (such as *one-half* as in a half-straightened finger) are also conceptually straightforward, as are extensions using the two hands of the user.

Other uses of hand-gesturing that are common in inter-human communications are those that signify (the eight) directions in (3-D) space and other intuitive physical attributes such as speed and intensity.

Challenges in high-accuracy gesture recognition include the provision of a fairly high-quality hand image, as well as algorithms that are robust in the contexts of arbitrary image placement in the field of view and at least partial occlusion of finger positions. Further, if the system is to be user-independent, normalization for hand- (and finger-) size and shape should be part of the algorithm.

The system to be demonstrated achieves these capabilities in the context of a vocabulary on the order of ten, as in Figure 2. The salient technical features of the algorithm include robust locations of dominant (finger) contours and vertices, and matches of shapes based both on deterministic graphs and probabilistic models.

In the videotape demonstration of the system, the user employs the high-accuracy gesture recognizer to navigate successfully and smoothly through a panoramic 3-D scene.

4. CONCLUSION

Human-machine interaction by voice is a well-recognized reality. Machine recognition of human gestures is about to become a reality as well. As we enter the next century, as the qualities of our acoustic and visual transducers become increasingly better, and as humans seek greater variety and complexity in the network-connected multimedia information station of the future, one can only imagine the many synergies that will come about when humans and machines converse both by voice and gesture.

5. ACKNOWLEDGEMENT

I thank several colleagues at Bell Laboratories for inputs leading to this paper; in particular, thanks to Mike Brown and Jakub Segen for the videotape demonstrations of speech and gesture recognition.

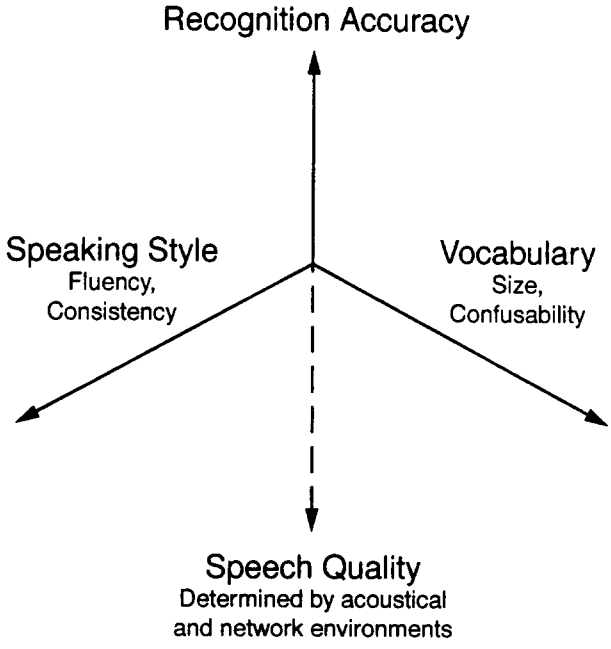

Figure 1

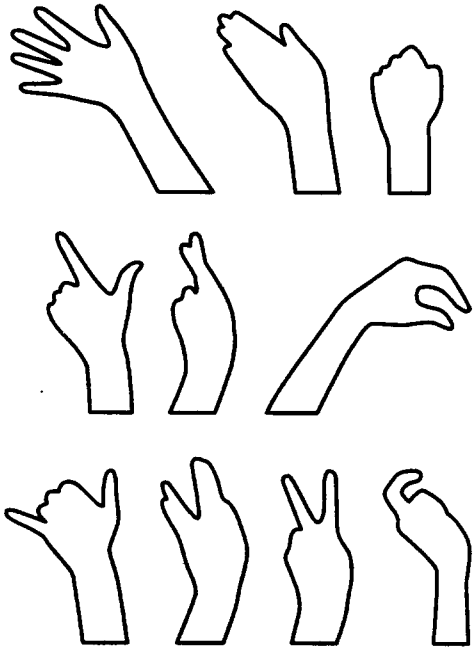

Figure 2

Copyright 1997 IEEE

Copyright 1997 IEEE