Chapter 2

SPEECH RECOGNITION TECHNOLOGY IN MULTIMODAL/UBIQUITOUS COMPUTING ENVIRONMENTS

Sadaoki Furui
Tokyo Institute of Technology, Department of Computer Science, Tokyo, Japan
furui@cs.titech.ac.jp

Abstract In the ubiquitous (pervasive) computing era, it is expected that everybody will access information services anytime anywhere, and these services are expected to augment various human intelligent activities. Speech recognition technology can play an important role in this era by providing: (a) conversational systems for accessing information services and (b) systems for transcribing, understanding and summarising ubiquitous speech documents such as meetings, lectures, presentations and voicemails. In the former systems, robust conversation using wireless handheld/hands-free devices in the real mobile computing environment will be crucial, as will multimodal speech recognition technology. To create the latter systems, the ability to understand and summarise speech documents is one of the key requirements. This chapter presents technological perspectives and introduces several research activities being conducted from these standpoints.

Keywords: Speech understanding; Speech summarisation; Human-computer interaction; Robustness; Adaptation; Spontaneous speech.

1. Ubiquitous/Wearable Computing Environment

The continuing progress in hardware and software development technologies have lead to the augmentation of computer performance at such a rapid pace that it improves several hundred times in every 10-year period. Resultingly, computers are getting smaller, more powerful and cheaper. Regardless of whether and to what degree they are noticed by users, computers will proliferate into every facet of our daily lives. People will actually walk through their day-to-day lives wearing several computers at a time. Thus, making com-
puters mobile and portable, exemplified by the present PDA (personal digital assistant) technology, is considered to be the transition phase to wearable computing (Pentland, 1998). Making computers more functional and smaller will generate not only quantitative changes but also qualitative changes in the way we use computers. In the near future, various computers including portable equipment existing everywhere will work together in autonomous collaboration (Weiser, 1991). Indeed, the new characteristics of computing will greatly change the focus and approach of human-computer interaction.

The transmission channel capacity of portable terminals will be easily expanded to the level of several Mbps, taking advantage of the technological progress available through, for example, MMAC (Multimedia Mobile Access Communication) systems. The exchange of dynamic information will be possible in addition to that of simple characters and voice information. In turn, this will give rise to sophisticated collaboration and coordination of human-machine systems based on autonomous protocol and information exchange between computers distributed everywhere.

2. State-of-the-Art Speech Recognition Technology

Speech is the primary, and the most convenient means of communication between people (Juang and Furui, 2000). Therefore, speech recognition systems are expected to play important roles in the ubiquitous/wearable computing environment (Furui, 2000b). The field of automatic speech recognition has witnessed a number of significant advances in the past 5-10 years, spurred on by advances in signal processing, algorithms, computational architectures, and hardware. These advances include the widespread adoption of a statistical pattern recognition paradigm, a data-driven approach which makes use of a rich set of speech utterances from a large population of speakers, the use of stochastic-based acoustic and language modelling, and the use of dynamic programming-based search methods (Juang and Furui, 2000; Rabiner and Juang, 1993; Furui, 2001; Furui, 2000a). Major applications of speech recognition technology can be classified into (a) conversational systems for accessing information services and (b) systems for transcribing, understanding and summarising ubiquitous speech documents such as meetings, lectures, presentations and voicemails.

Figure 1 shows a mechanism of state-of-the-art speech recognisers (Ney, 1997). Common features of these systems are the use of cepstral parameters and their regression coefficients as speech features, triphone HMMs as acoustic models, vocabularies of several thousand or several tens of thousand entries, and stochastic language models such as bigrams and trigrams. A word sequence, $w_1, \ldots, w_k$, which maximises the a posteriori probability is selected.