3 Spoken Language Communication Technology

3-1 Overview of Spoken Language Communication Technologies

KASHIOKA Hideki

The goal of Spoken Language Communication Laboratory, Universal Communication Research Institute, NICT, is to realize multi language communication technologies with spoken language regardless of who, where, when, how and in which language users speak. Toward this goal, we will intensively develop ICT for a human-machine interface, such as multilingual speech recognition, multilingual speech synthesis, and spoken dialogue technology. In this paper, we indicate these technologies overview.

Keywords
Speech recognition, Speech synthesis, Dialogue processing

1 Introduction

With the development of information and communication technologies, it is desired to achieve communication between various people in various environments and situations, for example, communication between people in distant locations or who use different languages. To realize such communications, multilingual spoken language communications need to be studied for people's smooth interactions in any languages, at any time and in any place with any expressions. The research and development of spoken language communication technology is an important issue since spoken language communication is one of the most natural human communications. In our daily life, along with the rapid spread of smart phones, voice services to access various kinds of information are now available and used by many people.

In this paper we give an overview of speech recognition, speech synthesis, and dialog system technologies, which are the major technologies that constitute spoken language communication technology.

2 Spoken language communication technology

Spoken language communication technology is the technology of not only recording and utilizing the information of various speech-based daily communications but also overcoming communication barriers. The major technologies for spoken language communication are speech recognition technology for transforming speech dialog to text, speech synthesis technology for providing the speech output of text information, and dialog system technology for supporting speech-based interactions.

2.1 Speech recognition technology

We often receive a large amount of speech information not only in conversations with

others but also from various announcements, TV, radio, and videos on the Internet. Speech recognition technology transforms such speech into text.

For the transform, the technology learns from many corpora the models of extracting speech features and using them for the transform of speech. It then learns the models of verbalizing the obtained character strings as words, phrases and sentences, and checks the input speech against the learned models. The models of extracting speech features and using them for the transform of speech are called acoustic models and the models of verbalizing the obtained character strings as words, phrases and sentences are called language models. Spoken Language Communication Laboratory has developed a high-speed highly accurate speech recognition system [1] using Weighted Finite State Transducer (WFST) as a search model, as well as acoustic and language models, to browse the results of checking speech against these models. The speech recognition system is used in a spoken dialog system and a speech translation system. We have a Japanese dictionary of 650,000 words and can process a six-word speech within 1 RTF (Real Time Factor).

Speech that we receive in various environments may contain non-speech sounds. To perform speech recognition, we therefore need not only the above-mentioned speech-to-text transform technology but also technology to recognize the non-speech sounds as noise and a voice activity detection technology to extract the speech section that includes the target speech. In addition, microphones that receive speeches are not always ideal for speech recognition. A familiar example is applications on mobile terminals such as smart phones. To process various kinds of noises and speech, such applications use noise reduction processing and build a speech recognition model (in particular an acoustic model) of higher noise resistance from noise-contained sound data before making speech recognition.

A familiar application of speech recognition is the speech translation system and spoken dialog system used in devices like smart phones. Applications in call center systems are also expected. The utilization of speech recognition for captioning news and other TV programs or internet videos has been requested for a variety of reasons (e.g. for supporting people with disabilities or for recording motion pictures). To use speech recognition practically in these applications, we need to work on the issues of noise processing, long sentence processing, and precision improvement etc. Multilingual support is also an important issue.

2.2 Speech synthesis technology

Speech information is often necessary in various situations. In particular speech synthesis is actually used in public transportation and disaster-prevention announcements. When information that people want to receive in a spoken form is given as text, speech synthesis technology is used to create speech information from that text.

To synthesize speech from text, the speech synthesis technology is comprised of a text analysis unit for parsing the text with syntactic and semantic information, and a synthesis engine for determining how intonation and rhythm are used to produce sounds from the words and phrases in the text. The text analysis unit extracts the word and phrase information and the synthesis engine uses the acoustic model for speech synthesis. At present, Spoken Language Communication Laboratory uses an HMM speech synthesis method to implement speech synthesis supporting not only Japanese but also English, Chinese, Korean, Indonesian, Vietnamese, and Malay [2]. Since speech synthesis acoustic models cover different languages and have different voice qualities and speech styles, the speech style and voice quality can be changed by switching the models. With a focus on this fact, a voice selector was developed. It produces speech in a voice similar to the original speaker's by selecting a model that has similar voice characteristic to the speaker's in speech translation and other processes. The naturalness of synthesized

speech is also enhanced by improving the filter used for the development of the model.

The speech synthesis system is necessary to build a speech translation system and a speech dialog system. Since it is used in conversations with people, synthesized speech with enriched naturalness is required. Speech synthesis acoustic models are built from a speech corpus of the same person. However, research suggests that the model's naturalness is higher when the corpus is made from conversations than when it is made from reading text. Since building the speech synthesis acoustic models costs a lot, the automatic building of models is another important issue.

2.3 Dialog system technologies

To maintain a spoken dialog and make appropriate questions and answers, it is necessary to perceive the situation and environment of the dialog and understand the dialog itself. In speech communications, not only the dialog's content but also the speech information may provide information related to the situation and environment. Information can also be obtained by considering a series of speeches. Dialog system technology is a technology for managing dialogs comprehensively, understanding them, and predicting and producing the next speech in various occasions.

Dialog system technologies can be classified into speech understanding technology for the understanding of speech, context processing technology for producing a response according to the underlying dialog act as obtained in the speech understanding process, by considering the association with surrounding information services, and speech producing technology for producing a response sentence from the response content. To understand dialog, it is necessary to understand the dialog content and dialog act. The dialog content can be understood by describing the content as a logical proposition based on the understanding of proper nouns, various expressions, homonyms, and different expressions of the same target. The dialog act, such as a request, question, or information offering, can be acknowledged using speech expressions, intonations and other speech information. Spoken Language Communication Laboratory built a tourist information speech dialog corpus and developed a Dialog Management System [3] with WFST to achieve rapid and accurate speech language understanding.

The dialog system technology is used directly for building a spoken dialog system. Unlike QA-systems, users can keep a dialogue with this system and obtain appropriate information. It can also be applied to not only the dialog system but also a mechanism that understands and predicts the context. This is an important technology to realize appropriate context-based processing in various speech communication technologies.

3 Conclusions

We have given an overview of the speech recognition, speech synthesis, and dialog system technologies, which all constitute spoken language communication technology. The Spoken Language Communication Laboratory, Universal Communication Research Institute, NICT, has been promoting research and development with a focus on the problems and future of these elemental technologies. It not only performs research and development into the individual elemental technologies but also combines them with those developed by other laboratories to realize an integrated system that can be used in actual society. Specifically, the Laboratory has developed the speech translation system VoiceTra by integrating these speech recognition, speech synthesis, and multilingual translation technologies, and the spoken dialog system AssisTra by integrating speech recognition, speech synthesis, and dialog system technologies. These systems are released as applications available on smart phone devices for demonstration and field experiment purposes. In the future we will promote the research and development of technology for building a speech archive and of multilingual communication technology for

providing smooth communications against various communication barriers by utilizing speech information contained in speech data and video data as well as text information.

References

1 Dixon Paul Richard, Chiori Hori, and Hideki Kashioka, "A COMPARISON OF DYNAMIC WFST DECODING APPROACHES," In Proc. ICASSP, 2012.

2 Yoshinori Shiga, "EFFECT OF ANTI-ALIASING FILTERING ON THE QUALITY OF SPEECH FROM AN HMM-BASED SYNTHESIZER," In Proc. ICASSP, 2012.

3 C. Hori, K. Ohtake, T. Misu, H. Kashioka, and S. Nakamura, "Statistical Dialog Management Applied to WFST-based Dialog Systems," In Proc. ICASSP, pp. 4793–4796, 2009.

(Accepted June 14, 2012)

***KASHIOKA Hideki**, Ph.D.*

Director, Spoken Language Communication Laboratory, Universal Communication Research Institute

Spoken Language Processing, Speech Translation, Spoken Dialogue