2007－SLP－67 (9)
2007／7／20

NICTにおけるユニバーサルコミュニケーションのための音声言語研究

中村 哲[1,2]

[1]（独）情報通信研究機構 知識創成コミュニケーション研究センター 音声言語グループ
[2]（株）国際電気通信基礎技術研究所 音声言語コミュニケーション研究所
京都府相楽郡精華町光台2-2-2 satoshi.nakamura@{nict.go.jp,atr.jp}

内容梗概 ユニバーサルコミュニケーション技術とは、あらゆる利用者がネットワークに接続されたコンピュータなどの機器と自然にコミュニケーションするための技術である。筆者らのグループでは、だれが、いつどこで、どのように、何語で話しても、イントネーションや表情、ジェスチャーなどの非言語情報を考慮しながらことばを核とした自然なコミュニケーションができる技術の確立を目標に、2006年4月よりNICT（情報通信研究機構）の第2期中期計画の研究プロジェクトとして活動を開始した。さらに、具体的には、この目標に向けて、音声対話システム技術の研究開発、さらに、これまでATRで行ってきた音声翻訳技術の研究開発をユニバーサルコミュニケーション技術の重要な要素技術として位置づけて研究を進めている。本稿では、本プロジェクトのねらい、アプローチ、体制などについて紹介する。

Spoken Language Technologies for Universal Communication

Satoshi Nakamura[1,2]

[1] Spoken Language Communication Group, Knowledge Creating Communication Center, National Institute of Information and Communications Technology,
[2] ATR Spoken Language Communication Research Labs.
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan

Abstract. A universal communication technology is the one that enables human beings communicate with machines breaking any kinds of digital divide. Research goal of the spoken language communication research group is to achieve spoken language-based natural communication regardless of who or where speakers are, when or how they use them, or in which language they are communicating, taking advantage of paralinguistic information such as intonation, facial expressions, and gestures. Towards the goal, we had launched research on multi-modal spoken dialog systems in 2006. We also continue research works on a speech-to-speech translation research that has been done at ATR that enables natural oral communication among different language speaking people as one of the most important universal communication technologies.

1. Introduction

The advent of Internet technology has brought human beings into a new environment that provides various kinds of information through computers. However, the interface between humans and machines is still cumbersome. The goal of the spoken language communication research group is to achieve natural communication technologies regardless of who or where speakers are, when or how they use them, or in which language they are communicating, taking advantage of paralinguistic information such as intonation, facial expressions, and gestures. We will intensively develop ICT technologies toward achieving this goal of obtaining a human-machine interface, such as multilingual automatic speech processing technologies, using paralinguistic information processing technologies for intonation, facial expressions, and gestures,

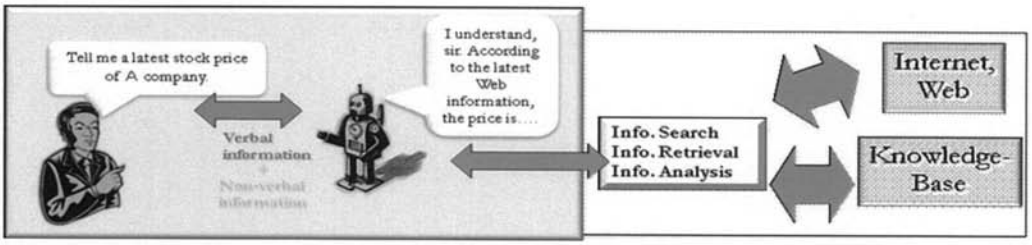

Figure 1 Spoken Dialog System

spoken language processing technologies for various colloquial expressions, multimodal synchronization technologies for speech and other modalities, and automatic corpus-collection technologies for multimodal speech and languages. Furthermore, we intend to study and demonstrate their feasibility through developing a prototype of a proactive spoken-dialog system with a multimodal dialog interface and conducting field trials.

Figure 1 shows a conceptual picture of the spoken dialog system for universal communication. We aim to develop natural spoken dialog systems that can provide any kind of information with users through natural conversation regardless of who or where speakers are, when or how they speak, or in which language they speak. Another target technology is speech-to-speech translation that enables natural oral communication between different language speaking people. In this paper, section 2 describes on the spoken dialog system and section 3 describes on the speech-to-speech translation. Section 4 describes summary and goals of the research and future works.

2. Spoken Dialog System

Recently the internet became an enormously rich information source. Any kinds of information can be found in the internet. However, it is often difficult for users to know how to retrieve the expected information. This is because novice user doesn't know what he wants to know, what actual queries are, and how to use the search engines. We aim to develop technologies for the users to clarify what he wants to know and retrieve the requested information by a series of spoken dialogs. The spoken dialog systems' technology also realizes an automatic call center system that responds to the customers query by a machine. The automatic call center system drastically saves a labor cost.

Spoken dialog systems have been studied and

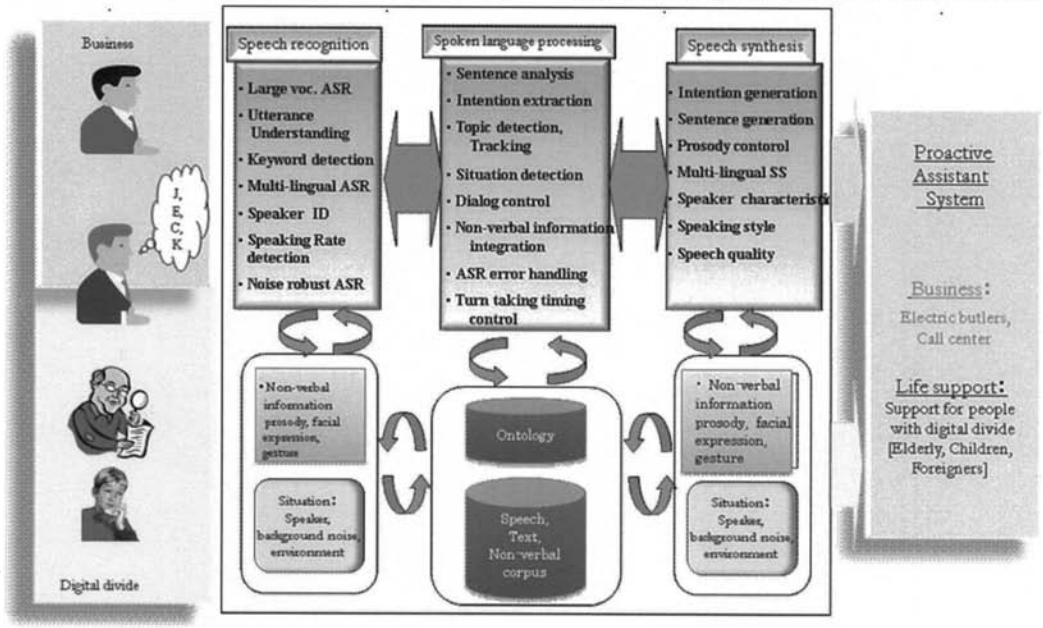

Figure 2 Block diagram and research topics of dialog systems

developed in many research projects and research laboratories. In early 90's DARPA supported a project named ATIS that is to develop a dialog system for air-line reservation [1]. In mid 90's DARPA funded a project named DARPA communicator based on the MIT Galaxy system [2].

The technology development targeting to an automatic call center has been conducted and launched the real service for directory assistance, help desk, trouble shooting, and reservations.

However, these systems are so domain specific that it is not applicable to wider and unstructured domains like information search on the WEB. Also these systems behave very unnatural since these systems only use text level linguistic information.

Current research topics of dialogue systems are mainly two directions. One is how to integrate non-verbal information like intonation, pause, head pose, face, and gesture to the verbal linguistic-based spoken dialog system. The conventional text-based dialog system could not understand intention clearly and serve natural and appropriate turn taking, thereby interface is far from natural and friendly one. In order to study this issue, we are currently collecting a dialog corpus on a Kyoto sightseeing task including the verbal and non-verbal information. System overview and research topics are shown in figure 2.

The second topic is robustness of the system against variety of disturbances such as: speaking style differences by age, gender, local accents, emotions, and different way of expression and query, different situations, domains, and tasks of dialog, different languages, and different acoustic environments. Especially, since different way of expression is quite serious for a spoken dialog system, we are planning to build multi-lingual ontology to overcome this problem. Figure 3 shows NICT-SLC group organization. SLC group is composed of four projects, Multilingual speech technology project for speech recognition and synthesis for multiple languages, Non-verbal information technology project for integration of non-verbal information to linguistic information, Natural language technology project for processing spoken language, and Spoken language technology project for developing dialog systems. In 2006 we have engaged multi-lingual spoken dialog system study. We have developed a baseline dialog system on Kyoto tourist information assistance in a client-server fashion. Figure 4 shows the client terminal based on Sony VAIO pc. We have installed a new functionality compared to conventional spoken dialog systems, which is speaking speed control. The system will respond with longer sentences with detailed information, when the user inputs his query utterances in slow speaking rate.

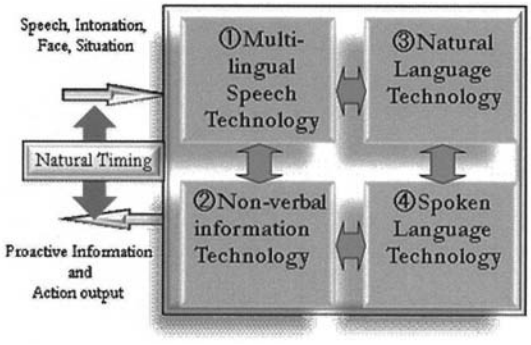

Figure 3 Organization of NICT SLC

3. Speech-to-speech Translation

Many research projects have addressed speech-to-speech translation(S2ST) technology, such as ATR[3], VERBMOBIL[4], C-STAR[5], NESPOLE![6], and BABYLON[7]. S2ST between Western languages and a non-Western language, such as English-from/to-Japanese, or English-from/to-Chinese, requires technologies to overcome the drastic differences in linguistic expressions. For example, a translation from Japanese to English requires, (1) a word separation process for Japanese because Japanese has no explicit spacing information, and (2) transforming the source sentence into a target sentence with a drastically different style because their word order and their coverage of words are completely different, among other factors.

The other factor for S2ST is that the technology must be portable for every domain because S2ST systems are often used for applications in a specific situation, such as supporting a tourist's

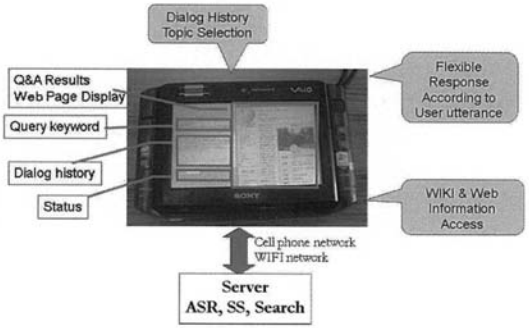

Figure 4 Prototype Dialog System

conversations in non-native languages. Therefore, the S2ST technique must include (semi-) automatic functions for adapting to specific situations/domains and specific language pairs in speech recognition, machine translation and speech synthesis.

Multilingual speech-to-speech translation devices are vital for breaking the language barrier, which is one of the most serious problems inherent in globalization. In S2ST, machine translation is the core technology for generating natural translation from original input. Most of the currently available machine translation systems are handcrafted rule-based translation systems designed for written text, mainly because it is difficult to gather data that exhaustively cover diverse language phenomena. In rule-based systems, efforts have been made to improve rules that abstract the language phenomena by using human insight. In taking this type of approach, however, it is difficult to port a particular system to other domains, or to upgrade the system to accommodate new expressions. Portability is one of the most important factors for S2ST, because S2ST systems are often designed for a specific domain and situation for various language pairs depending on their users. Therefore, customization for their domains and situations and for their language pair is obligatory work for S2ST.

With the increased availability of substantial bilingual corpora by the 1980s, corpus-based machine translation (MT) technologies such as example-based MT and stochastic MT were proposed to cope with the limitations of the rule-based systems that had formerly been the dominant paradigm. Since 1986, we at ATR have been conducting research on applying corpus-based

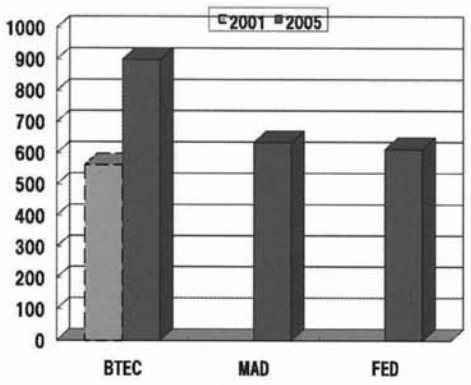

Figure 6 Speech Translation Performance (TOEIC)

methods to speech translation and developed many technologies. Our research experience shows that corpus-based approaches are suitable for speech translation technology. This is because corpus-based methods: (1) can be applied to different domains; (2) are easy to adapt to multiple languages; and (3) can handle ungrammatical sentences, which are common in spoken language. Now we at NICT collaboratively keep working on speech-to-speech translation with ATR. Figure 5 shows a block diagram of our speech-to-speech translation system. The system composed of automatic speech recognition, machine translation, and speech synthesis. All of the modules are corpus-based and statistical model-based systems.

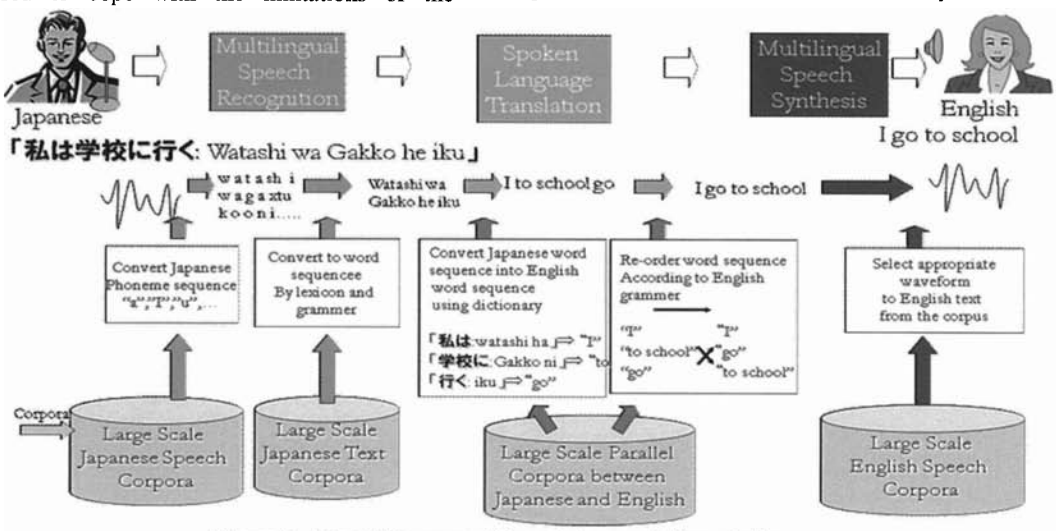

Figure 5 Block Diagram of Speech-to-speech Translation

Figure 6 shows a performance improvement for travel sentences compared to the performance in 2001. The performance is measured by the equivalent TOEIC test proposed by [8]. In the figure, BTEC, MAD, and FED indicate basic travel expressions, dialogs through speech translation system, and field data collected in Kansai International Airport, respectively. The figure shows that the performance of the current speech-to-speech translation system even outperform 600 TOIEC score.

We are going to extend the speech-to-speech translation research in the following directions,

- Wider domain including business,
- More language pairs,
- Simultaneous speech-to-speech translation.

4. Conclusions

This paper introduces research goals of NICT-SLC research group. NICT-SLC aims to achieve spoken language-based natural communication regardless of who or where speakers are, when or how they use them, or in which language they are communicating. The paper also introduces current research activity of spoken dialog system and speech-to-speech translation.

References

[1]Pallet, D. S., et al (1992). *DARPA Feburary 1992 ATIS benchmark test resuls.* Proceedings, Human Language Technology Conference.

[2]DARPACommunicator.http://communicator.sourceforge.net/sites/MITRE/distributions/GalaxyCommunicator/docs/manual/index.html.

[3]Nakamura,S et al, (2006). *The ATR Multilingual Speech-to-speech Translation System.* IEEE Transaction on Audio, Speech, and Language Processing.

[4]Wahlster W. (2000). Foundation of speech-to-speech translations. Springer Verlag.

[5]C-Star. http://www.c-star.org.

[6]NESPOLE! http://nespole.itc.it.

[7]BABYLON.http://www.darpa.mil/ipto/research/babylon/ap-proach.thm.

[8]Sugaya F, et al. (2000). Evaluation of the ATR-MATRIX speech translation system with a pair comparison method between the system and humans. Proceedings, ICSLP.