

Speech-to-Speech Translation: Technology and Applications Study

C.J. Weinstein

10 May 2002

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS



Prepared for the Air Force, AFRL/IFEC, under
Contract F19628-00-C-0002.

Approved for public release; distribution is unlimited.

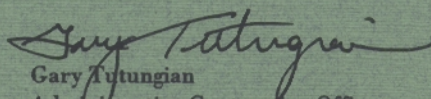
This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Air Force, AFRL/IFEC, Rome Research Site, under Contract F19628-00-C-0002 and does not represent a commercial offering by Lincoln Laboratory.

This report may be reproduced to satisfy needs of U.S. Government agencies.

The ESC Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER



Gary Tutungian
Administrative Contracting Officer
Plans and Programs Directorate
Contracted Support Management

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission has been granted by the Contracting Officer to destroy this document, when it is no longer required by the using agency, according to applicable security regulations.

Massachusetts Institute of Technology
Lincoln Laboratory

Speech-to-Speech Translation: Technology and Applications Study

*C.J. Weinstein
Group 62*

TECHNICAL REPORT 1080

10 May 2002

Approved for public release; distribution is unlimited.

ABSTRACT

This report describes a study effort on the state-of-the-art and lessons learned in automated, two-way, speech-to-speech translation and its potential application to military problems. The study includes and comments upon an extensive set of references on prior and current work in speech translation. The study includes recommendations on future military applications and on R&D needed to successfully achieve those applications. Key findings of the study include: (1) R&D speech translation systems have been demonstrated, but only in limited domains, and their performance is inadequate for operational use; (2) as far as we have been able to determine, there are currently no operational two-way speech translation systems; (3) intensive, sustained R&D will be needed to develop usable two-way speech translation systems. Major recommendations include: (1) a substantial R&D program in speech translation is needed, especially including full end-to-end system prototyping and evaluation; (2) close cooperation among researchers and users speaking multiple languages will be needed for the development of useful application systems; (3) to get military users involved and interacting in a mode which enables them to provide useful inputs and feedback on system requirements and performance, it will be necessary to provide them at the start with a fairly robust, open-domain system which works to the degree that some two-way speech translation is operational.

TABLE OF CONTENTS

Abstract		iii
1	INTRODUCTION AND OVERVIEW	1
2	SPEECH TRANSLATION MODELS AND COMPONENT TECHNOLOGIES	3
3	SPEECH TRANSLATION STATE-OF-THE-ART REVIEW	5
4	POTENTIAL MILITARY APPLICATIONS OF SPEECH-TO-SPEECH TRANSLATION	11
5	R&D DIRECTIONS FOR 2-WAY SPEECH-TO-SPEECH TRANSLATION	15
REFERENCES		17

LIST OF ILLUSTRATIONS

Figure No.		Page
1	Machine translation pyramid	3
2	Illustration of interlingua-based speech-to-speech translation system	4

1. INTRODUCTION AND OVERVIEW

This report describes a study effort in speech-to-speech translation technology and applications, which was conducted during FY2001 by MIT Lincoln Laboratory under AFRL/IFEC sponsorship. The purposes of the effort were to conduct a study on the state-of-the-art and lessons learned in automated speech-to-speech translation and its potential application to military problems; and to make recommendations on future military applications and on R&D needed to successfully achieve those applications. The study was summarized in a briefing to AFRL/IFEC personnel at Lincoln Laboratory on November 29, 2001. This is the final report of the study, and includes additional detail as well as an extensive set of references.

The organization of the report is as follows. This introductory section includes an outline of findings and recommendations. Section 2 gives an overview of speech translation models and component technologies, to set context for the rest of the report. Section 3 contains a selective state-of-the-art review, which is supplemented by a set of references on speech translation R&D. In Section 4, potential military and government applications are discussed, and comments are included on our experience in interacting with the military concerning such applications. Section 5 summarizes our recommendations on future R&D efforts needed to enable speech-to-speech translation for military applications.

Our key findings are outlined as follows:

- (1) As far as we have been able to determine, there are currently no operational two-way speech translation systems.
- (2) R&D speech translation systems have been demonstrated, but they work only in limited domains and their end-to-end performance is inadequate for operational use.
- (3) One-way fixed phrase translation has shown promise for some military application, but is not in regular use.
- (4) Intensive, sustained R&D will be needed to develop usable two-way speech translation systems.
- (5) It is very difficult to obtain specifications for limited domain military (or other) speech translation applications from potential users, in terms of required vocabulary, grammar, and other parameters needed for system development; the users have difficulty specifying in advance what they need.

Our major recommendations are outlined as follows:

- (1) A substantial R&D program in speech translation is needed, including component technology development but especially including full end-to-end system prototyping, live collection of two-way translation dialog speech, system evaluation supported by corpus collection, and iterative development.
- (2) Close cooperation among researchers and users speaking multiple languages will be needed for the development of working systems and useful applications.
- (3) To get military users involved and interacting in a mode which enables them to provide useful inputs and feedback on system requirements and performance, it will be necessary to provide them at the start with a fairly robust, open-domain system which works to the degree that some two-way speech translation is operational.

A brief prior review of current levels of capability and future prospects for speech translation as of 1999 is presented in [Lazzari 1999]. The conclusions here are generally consistent with that earlier study. For example, [Lazzari 1999] states that spoken language translation, even in limited domains, still presents considerable challenges, which are the object of ongoing research. Another point made in the earlier study is that system evaluation, including the collection of appropriate corpora, can be expected to be a bottleneck impeding future research progress.

2. SPEECH TRANSLATION MODELS AND COMPONENT TECHNOLOGIES

The purpose of this section is to give an overview of speech translation models, and of the required technologies and resources for a speech translation system. This will set context for our discussion of specific speech translation systems, including their performance and potential applications.

We begin with the machine translation pyramid [Hutchins 1992], shown below, which is a useful diagram to set background for our discussion of speech translation. This diagram illustrates the relationships among various approaches to machine translation of either text or speech. It shows source language analysis along the left side and target language generation along the right side. Direct translation approaches do little source language analysis. Direct Translation can give quick results, especially for one pair of languages and a limited domain. An extreme example of direct translation would be a phrase-book approach where translations for all possible input phrases are simply stored and looked up as needed. The interlingua approach includes the most detailed analysis, producing a language-independent meaning representation of each input sentence. The interlingua approach, though challenging to implement, has major advantages for multilingual applications.

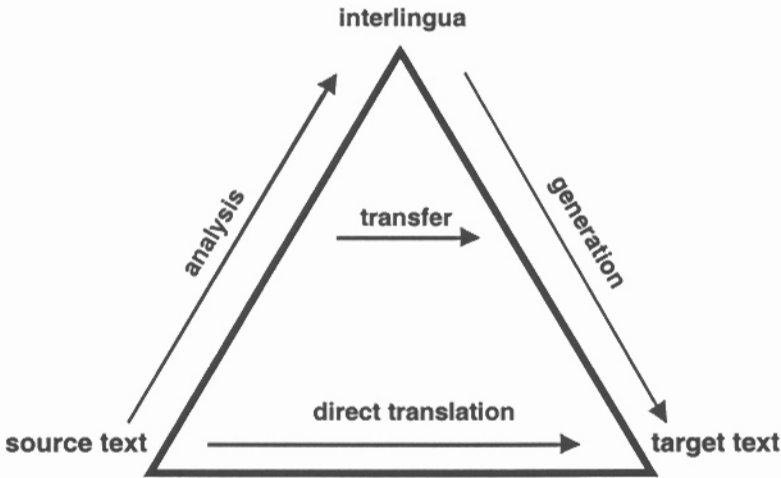


Figure 1. Machine translation pyramid.

To provide a framework for discussion of the processes, technologies, and resources needed for speech-to-speech translation, we show the following diagram of a multilingual, interlingua-based speech translation system. This diagram is based on a system we have developed at Lincoln Laboratory [Weinstein 1997] for English/Korean translation, but its elements are common to most interlingua-based speech translation systems.

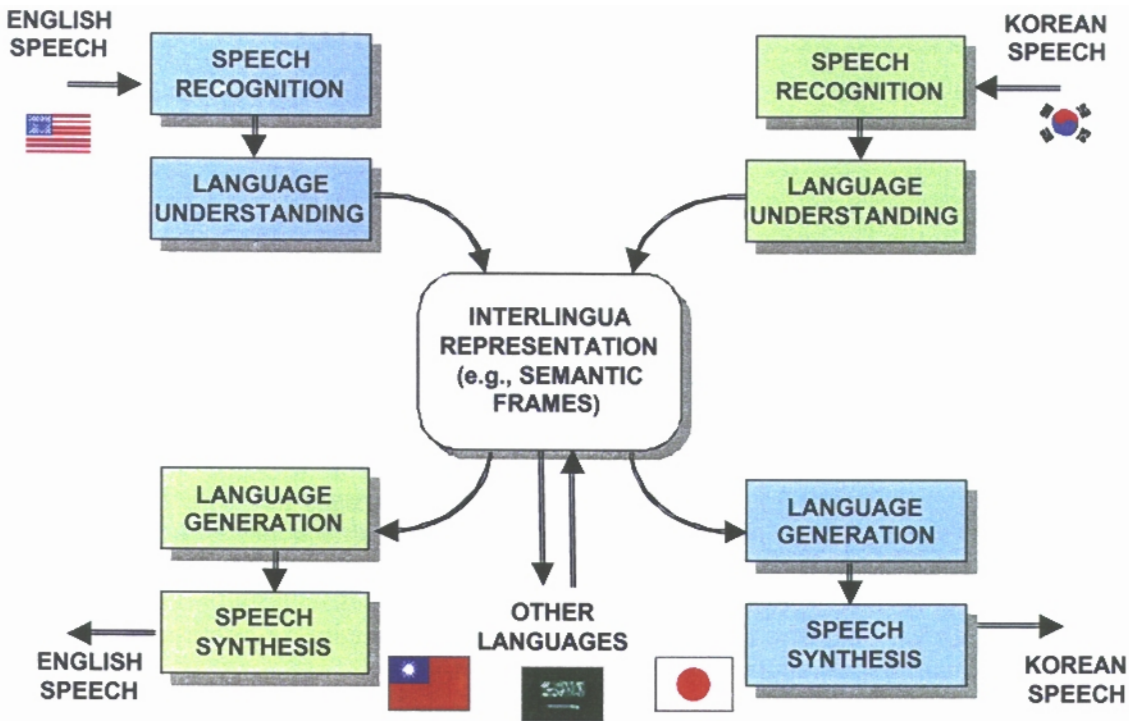


Figure 2. Illustration of interlingua-based speech-to-speech translation system.

The process flow for translating speech from one language to another includes: speech recognition to transform the input speech into a string of words; language understanding to create a language-independent interlingua-based meaning representation (which we call a semantic frame) of the words [Seneff 1992]; language generation to transform the semantic frame into words in the target language, and text-to-speech synthesis to produce the speech in the target language. If a transfer-based or direct translation system is used, the transfer would replace the understanding and generation. But speech recognition and text-to-speech synthesis would still be needed. Note that the elements of the processing chain must be implemented for every source language and for every target language for which speech translation is to be performed. The performance of the speech translation system will obviously depend strongly on the performance of the individual components in the processing chain, which in turn will vary substantially depending on: the languages used; the complexity of the task in terms of vocabulary and grammar; and other factors such as the acoustic environment and the users' knowledge of the system, the domain or topic, and each other.

The development of each module in the diagram requires substantial supporting resources, which are needed for each language. Speech recognition resource requirements include transcribed training speech, statistical language models for the task of interest, and phonetic dictionaries. Language understanding resource requirements include parsers, syntactic and semantic grammars, and training text with linguistic annotations. The actual translation process requires translating lexicons, either between each language and an interlingua representation, or bilingually between language pairs. Language generation requires generation grammars to produce well-formed sentences in the language of interest. Text-to-speech synthesis requirements include: text-to-sound rules; large corpora of training speech (for example-based synthesis); and phonological and stress rules. Achievable system performance will depend strongly on the availability and proper utilization of resources such as these for each language to be included in the translation system.

3. SPEECH TRANSLATION STATE-OF-THE-ART REVIEW

Our review of research and the state-of-the-art in speech translation begins with a discussion of the work of two research consortia: (1) the Consortium for Speech Translation Advanced Research (C-STAR), an international group [C-STAR 2001] which has been a focus for speech translation R&D since about 1991; and (2) the VERBMOBIL consortium [Wahlster 2000] of 19 academic and 4 industrial partners, who conducted a substantial R&D effort in speech translation during the 1990s, focusing on real-time translation to aid face-to-face communication in German, English, and Japanese. Then we describe specific research efforts at a few representative laboratories, with emphasis on those efforts, which develop full end-to-end speech-to-speech translation systems with substantial vocabularies and reasonable open grammars. In commenting on the research, we note the importance of system evaluation; as additional background, we make reference to long-term and ongoing translation evaluation work at the Defense Language Institute. Finally, we also discuss work on a “one-way” system, which translates a fixed set of spoken English phrases into the corresponding foreign language phrases. This report includes a thorough set of references to speech translation research, although the discussion in this section focuses on those references we consider to be most representative of the state-of-the-art. A number of the papers referenced are from special conference sessions on speech translation, which were held at ICASSP 1997, EUROSPEECH 1999, and ICSLP 2000. There is much written in the popular press about speech translation, and about machine translation in general. An interesting history of machine translation, including a timeline and a fanciful view of the future, is presented in the May 2000 issue of *WIRED Magazine* [WIRED 2000].

The C-STAR Consortium for Speech Translation Advanced Research [C-STAR 1996], [C-STAR 1999], [C-STAR 2001] was originally established in 1991. It is a voluntary research consortium aimed at international collaboration in speech translation technology. By 1999, there were 6 C-STAR “partners” and about 14 affiliate members. In total, 10 countries were represented. Each partner agreed to develop, at minimum, a “half-system” that could participate in multilingual C-STAR demonstrations and experiments. The “half-system” would translate from speech in the partner’s language to text in at least one other language, and would also convert text (in one’s own language) received from the translation system of another partner into speech in one’s own language. The affiliates attend C-STAR meeting, present their research, and share ideas, but do not need to develop C-STAR-compatible systems. A major 6-language C-STAR speech translation demonstration [C-STAR 1999] was carried out in July 1999 by the 6 partners, namely: ATR (Japan); CMU (USA); ETRI (Korea); Karlsruhe (Germany); IRST (Italy); CLIPS++ (France). The demonstration system included both a meeting task in which conversants were to set up meeting dates, and a travel task in which travel arrangements were to be made in conversation with other researchers taking on the role of travel agents. The core translation technologies employed included both an interlingua approach and an example-based approach, although by 1999 C-STAR’s emphasis was on an interlingua approach which would have major advantages as discussed earlier for multilingual applications. The July 1999 demonstration was a major event which for the first time showed that it was feasible to integrate a real-time, multilingual speech translation system. In addition to the core demonstration, advances in a number of related technologies (e.g., multimodal and multimedia interface) were shown. But although the demonstration illustrated what could be possible in speech translation, a careful reading of the individual papers and looking at the results in detail shows that much work is needed before systems achieve performance which would be usable operationally. In fact, the introduction to the C-STAR 1999 Proceedings predicted that it would take 5-10 years before commercialization of speech translation in limited domains would be achieved.

The VERBMOBIL Project is described in detail in [Wahlster 2000], which contains a large collection of papers by the participants at the end of the second phase of this 8-year project. An earlier overview of VERBMOBIL is [Kay 1994]. The focus of VERBMOBIL was on real-time translation to aid two-way face-to-face communication, rather than telecommunications. The VERBMOBIL scenario assumes a native speaker of German and a native speaker of Japanese, both of whom possess at least a passive knowledge of English. In case the active knowledge of English turns out to be insufficient, the partners are allowed to use their native language, and a goal of the VERBMOBIL system is to support the users by providing translation on demand from German or Japanese into English. So although the component technologies developed under VERBMOBIL are basically the same as for two-way speech translation, the goals and criteria are somewhat different. [Tessiere 2000] contains evaluation results in a VERBMOBIL context based on a simplistic criterion of translation accuracy in conversational turns: "the translation of a turn is approximately correct if it preserves the intention of the speaker and the main information of his/her utterance." Using this metric of translation accuracy, the authors discuss the impact of word error rate on translation accuracy. When word recognition error rates range from 20% to 50%, translation accuracy ranges from the 60% range to the 80% range, depending on a variety of factors (language direction, additional training for translation accuracy, etc.). There is also discussion of the ability of the translation system to enable users to accomplish various business-related tasks in the evaluation (meetings and travel arrangements). The overall percentage of successful task completions was just under 90%. Papers describing some of the technologies in VERBMOBIL include [Block 1995], [Bub 1997], [Engel 2000], [Finke 1997], [Niemann 1997], [Reithinger 1999], and [Ney 2000].

In the next few paragraphs, we discuss some of the research associated with C-STAR in more detail.

The Carnegie Mellon University Speech Translation Research group is one of the leading speech translation teams in C-STAR, and is both a technical and organizational leader in C-STAR. CMU's C-STAR system, referred to as JANUS, has used combinations of example-based and interlingua approaches, and currently focuses on an interlingua approach. JANUS-III results reported for on a travel domain in [Levin 2000] indicate that about 60% of English speech utterances yielded translations which were either "perfect" or "OK". This means that 40% of the translations were unsatisfactory, which would at a minimum require rephrasing or repeating the input sentence. This travel domain, in which was tested, was a big domain—total vocabulary was about 10,000 words. A sustained development, test, and evaluation cycle of effort [Waibel 1996], [Gates 1997], [Lavie 1997] on the simpler meeting scheduling domain (about 2,000 word vocabulary) led to somewhat better results. The [Gates 1997] paper gives a comprehensive discussion of evaluation methods and issues, and [Lavie 1997] reports achievement of 70% acceptable Spanish-to-English translation in the meeting domain. It is worth noting that most of the results reported for evaluation of speech translation systems, including the results just cited, are based on evaluations of the accuracy of off-line translation. Very little has been reported on translation effectiveness evaluation for real-time two-way speech translation. Such evaluations are difficult but are needed. Another effort at CMU, not directly focused on C-STAR, is the DIPLOMAT Project [Frederking 1997], which aims to enable rapid development of speech translation in multiple languages. CMU has been continuing to work in speech translation [Lavie 2001-1], including another international project called NESPOLE Negotiating through Spoken Language for E-commerce [Lavie 2001-2], which is focused on international commerce applications. At the time of this report, components of NESPOLE were under development and initial demonstrations had been performed [NESPOLE 2002].

ATR in Japan has conducted a sustained and strong effort in speech translation [ATR 2002], and has been a leading partner in C-STAR. ATR does emphasize development and test of full end-to-end translation systems [Shimizu 1997], [Iida 1998], [Sugaya 1999], [Yamamoto 2000], [Gruhn 2000], and has reported positive end-to-end results in travel and hotel reservations domains, as well as in more limited domains such as meeting arrangements. ATR has started some work in porting to other domains and to other language other than English/Japanese, but to date that porting work still requires a great deal of human labor and a substantial amount of difficult-to-collect bilingual data. An important issue with results, which are reported by the various research groups is that they are difficult to compare and interpret, because each group has its own scoring system. For example, ATR reported [Sugaya 1999] on an end-to-end evaluation for English/Japanese translation in a hotel reservation task that the “subjective score was 3.8 on scale of 5, indicating that users could acceptably carry out the task.” The fact that a two-way user test was carried out is very important, but comparative interpretation of the results with others is difficult. Clearly, there is a need for common corpora and common evaluation in speech translation. (For a description of a common evaluation that was carried out for text translation, see [White 1994].) But it is also clear that obtaining suitable, comprehensive corpora, and developing and agreeing upon an evaluation metrics, is a significant challenge. In the same study, ATR reported on task completion time and user adaptation, which are important aspects of speech translation. They found that over three sessions, users were able to reduce task completion time by about 25%. Part of this reduction occurred because users learned to reduce the length of their input utterances (i.e., average utterance length was reduced from 10 to 6 words over the three sessions). ATR’s fundamental translation approach requires a substantial amount of bilingual data to train the system. Such data is particularly difficult to collect for a two-way speech translation application. On the synthesis side, ATR has an excellent multilingual text-to-speech system based on concatenation of human speech samples [Black 1995], [Campbell 1996]. Thus the quality of the speech output in ATR’s translation systems is quite good.

The CLIPS++ group has developed a French system as part of C-STAR, and has focused on a system that transforms spoken French into an interlingua form (IF) consistent with other C-STAR systems [Blanchon 2000], [Boitet 2000]. They report that although they were able to achieve good demonstrations, many challenges remain in speech translation. In particular, they note the need for a more dialog-oriented, interactive, and tunable architecture, and indicate that the need to develop a task-oriented interlingua format is an important limitation on system flexibility.

IRST in Italy has been active both in C-STAR and in NESPOLE. Research in transforming Italian speech in the C-STAR travel domain into a collection of “dialog acts” represented in an interlingua format is described in [Angelini 1997]. Efforts in utilizing the dialog history to improve translation are also described. Additional research on the language component of this system is described in [Corraza 1999]. In [Lazzari 2000], the extension of this speech translation work into a multimodal, multimedia environment is described, with focus on future e-commerce applications in NESPOLE.

Korean-to-English and Korean-to-Japanese speech translation systems have been developed by ETRI, and have been demonstrated in the travel domain as part of C-STAR [Yang 1997]. Again here it is noted that the demonstrations were effective, and some promising results were obtained, but there were significant limitations on performance, especially for spontaneous speech. In addition to the speech recognition and understanding systems, ETRI has developed an excellent Korean text-to-speech system, based on concatenative synthesis [Jeon 1998]. The ETRI text-to-speech system has also been used successfully to produce Korean speech in the MIT

Lincoln Laboratory English/Korean speech translation system [Lee 2002]. More information on related ETRI research activities can be found at the ETRI web site [ETRI 2001].

MIT Lincoln Laboratory conducted some initial speech translation work [Tummala 1995] in 1994-5, which showed some ability of robust parsing to overcome speech recognition errors for 700-1000 word tasks. Since that time, Lincoln has participated in C-STAR as an affiliate, participating in C-STAR technical meetings [Weinstein 1996], [Lee 1999-1] but not in the actual C-STAR system development and demonstrations. Lincoln Laboratory's translation work has built upon and extended understanding [Seneff 1992] and generation [Glass 1994] technology developed at the MIT Laboratory for Computer Science, and has consistently focused on an interlingua-based approach to translation. From 1995-1999, Lincoln research focused on text translation [Weinstein 1997], [Lee 1999-1], [Lee 1999-2] (although some work on speech translation is reported in [Weinstein 1997]). Lincoln Laboratory's speech translation work was reactivated in 1999, focusing on two-way English/Korean speech translation in missile and medical domains. A number of successful real-time demonstrations were shown, but the translation performance for 1000-2000 word tasks was not adequate to support two-way dialog. For example, work [Lee 2002] on English/Korean speech translation in a new medical domain produced initial results where about 42% of sentences yielded translations sufficient to convey the speaker's intention. Research was conducted in robust parsing to cope with recognition errors and in utilizing discourse modeling to aid in recognition and understanding.

Here we note several additional research efforts in speech translation. An early English/Spanish system [Roe 1992] was developed in a collaboration between AT&T and Telefonica and Spain. It used a finite state grammar with 374 morphological entities in the domain of currency exchange. An interesting research project conducted by SRI in collaboration with Telia in Sweden has developed a spoken language translator [Rayner 1993], [Rayner 1997] in the air travel information system (ATIS) domain, and has combined rule-based and statistical techniques. The work builds on SRI's earlier ATIS work in human-machine interaction, and has included translation among English, French, and Swedish. A Spanish-to-English speech translation system, utilizing an example-based learning approach, and operating on about a 700-word vocabulary in a hotel front-desk communication domain, is described in [Vidal 1997]. Research at the Chinese Academy of Sciences has been carried out Chinese-to-English [Zong 2000-1] and Japanese-to-Chinese [Zong 2000-2] spoken language translation. The Chinese-to-English system uses a 12,000 word Chinese vocabulary and works in a hotel reservation domain. Results reported include 51% correct translation for sentences of average length 6.1 words. An English/Japanese system aimed at facilitating home shopping over the Internet is described in [Hiraoka 1997]. Speech translation aimed at portable systems is described in [Watanabe 2000] and [Matsui 2001]. The work at NEC [Watanabe 2000] deals with two-way English/Japanese translation for a travel domain. The Matsushita system [Matsui 2001] is implemented on a small PC, and includes two-way English/Japanese and two-way Japanese/Chinese translation. Recognizing that the translation capability is limited, the system is designed to aid the user in selecting input phrases, which are likely to be translated successfully. Recent research at AT&T Laboratories [Alshawi 2000] utilizes statistical language translation models called "dependency transduction models", while taking account of the hierarchical phrasal structure of language. The method has been applied to create English-Spanish and English-Japanese translation models for speech translation applications. Accuracy rates in the 70-75% range are reported, using a simple string edit-distance measure.

We have mentioned the importance of evaluation of translation several times. A useful source for translation evaluation resources is the Defense Language Institute, the organization within the Department of Defense responsible for foreign language training for the Armed

Services. Over the past decades, DLI has developed and implemented standardized evaluation metrics for human language translators in training. The metrics are designed to measure the difficulty of language samples as well as the performance of the human translators. In particular, a set of levels for calibrating skills of human translators has been developed, referred to as the Interagency Language Round Table (ILR) language skills for speaking and listening [DLPT 1991], [Child 1993], [SIL 2002], [DLI 2002]. Although the ILR levels (which rate speaking and listening skills on a scale from 0 to 5) have been used for decades as a reliable tool for training human translators, machine translation evaluation metrics have not incorporated them. Another potential resource for evaluation metrics is the National Institute of Standards and technology (NIST) Machine Translation Conference [NIST 2002]. This is a conference organized by NIST in 2002 for the automatic evaluation of text-to-text machine translation systems. The conference is closely associated with the TIDES program [TIDES 2002]. Although it does not address speech-to-speech machine translation, some of the automatic methods being developed here may be applicable for evaluating speech translation systems.

The DARPA One-Way Project, rather than aiming for providing advanced machine translation technology to the military, has taken the approach of providing spoken phrase translation systems (PTS) which use COTS speech recognition to recognize among a fixed set of phrases (typically about 500 English phrases), and a table lookup to play the output in the form of a recorded human translation in the foreign language. A good deal of information about the DARPA One-Way project is available at the Marine Acoustics website [Sarich 2001]A big advantage of the One-Way Project has been rapid deployment and accessibility for users. But clearly the capability is limited. A big problem is that users could not remember the set of phrases in the “book” and would get frustrated when recognition failed because a new phrase was spoken.

4. POTENTIAL MILITARY APPLICATIONS OF SPEECH-TO-SPEECH TRANSLATION

The potential for useful military application of machine translation, including speech-to-speech translation, as a force multiplier for military operations, is very great. The United States military operates worldwide in a variety of international environments that require language translation, and good human translators are a scarce commodity. The language barrier significantly reduces the speed and effectiveness of coalition operations. During hostilities, or during other military operations such as peacekeeping, any time saved by systems that can accurately and quickly translate between languages could reduce the possibility of miscommunication with allies or civilians and could provide an advantage over the adversary. However, this potential has generally not yet been realized because the technology is not sufficiently advanced or robust to support military applications. This is especially true of speech-to-speech translation, which is particularly challenging. In this section, we outline a number of potential military applications of speech-to-speech translation, with the caveat that advances will be needed to realize these applications. Then we discuss applications of the DARPA one-way phrase translator, whose partial success provides some useful information for developing speech translation applications. Finally, we comment on some issues and lessons from our experience in interacting with the military regarding speech translation and its applications.

The following is a list of potential military and government applications of speech-to-speech translation. Note that performance is the key to making any of these applications usable. Also note that because the military is involved in a wide variety of operations short of war, there is a great deal of overlap between military and civilian applications.

Military Operations Applications

- (1) Communication with allies—for example: support of face-to-face communication between commanders or soldiers; or multilingual radio communication in a coalition battlefield environment
- (2) Communication with local population—gathering information from civilian about the local environment or about opposing forces; providing information to avoid civilian casualties
- (3) Peacekeeping operations—including communication with multinational peacekeeping forces, with local population, and with opposing elements
- (4) Interrogation – of prisoners
- (5) Intelligence screening—of suspected adversaries or potential informants
- (6) Ship boardings and inspections—communication with ship crews to facilitate searches for weapons, military supplies, or contraband
- (7) Logistics and travel communications—communications to facilitate necessary movement of troops and supplies

Government and Law Enforcement Applications

- (1) Customs inspections and border patrols—translation to facilitate communication and to enhance the effectiveness and efficiency of these operations
- (2) Coast Guard inspections—communication with non-English speaking crews both for detection of possible criminal or terrorism activity, and for general safety and law enforcement
- (3) Police department and fire department operations—communication with non-English speaking people, mostly for law enforcement and safety but also with homeland defense applications

Humanitarian Assistance and Medical Applications

- (1) Disaster relief—communications support similar to peacekeeping applications
- (2) Medical diagnostics and treatment—translation for doctor/patient interaction
- (3) Local information and direction requests—translation to get information, travel directions, etc., from local population (this also would support military operations)

The DARPA One-Way Project [Sarich 2001] has taken the approach of deploying existing technology in order to provide rapid, though limited, translation aid for the military. A phrase translation approach is utilized, where the speaker is limited to a fixed set of English phrases. This speech translation system was based on an earlier system developed at the Naval Operational Medical Institute where the user would select an English phrase (in text) from a menu of choices, and the system would play out the translation in a selected foreign language. Existing, commercially available speech recognition technology is used to transform the spoken phrases to English text. The “one-way” approach avoids the need for foreign language speech recognition. The English speech recognition can be done with high accuracy for a fixed set of phrases, even if several thousand phrases are allowed, since each phrase is effectively treated by the recognizer as one long unit. Translation into the foreign language is done via a direct table lookup, and the foreign speech is produced by playing out a human recording of the designated phrase. One-Way translators have been deployed to Bosnia (1997), to the Arabian Gulf (1998), and to the Rim-of-the-Pacific (RIMPAC) exercises (2000). Foreign languages have included Serbo-Croatian, Arabic, Farsi, Korean, Japanese, and others. An initial applications of One-Way translation in Bosnia in 1998 included both force protection (“can you help me find landmines?”) and medical interviews. In the case of medical interviews, the doctor/interviewer could say things like “show me where it hurts,” or “does it hurt here.” The interviewee could respond with a yes or no gesture, or by other gestures such as pointing. A reasonably effective doctor/patient interview could be conducted with one-way, rather than two-way, translation. In the Arabian Gulf, the application was ship boardings (“Do you have any cargo bound for Iraq onboard?”) and inspections. In RIMPAC, the applications were refugee management and medical interviews. A big advantage of the One-Way Project has been rapid deployment and accessibility for users. In addition to the obvious limitation of supporting speech by only one of the parties in a conversation, another key limitation was the fact that English-speaking users would have difficulty recalling which phrases were allowed. The provision of a capability to add new phrases in the field has been proposed as a useful, though partial, solution to this problem. Although the capability is limited, we would suggest that the experience of the One-Way Project can be very useful in defining and developing more advanced applications which require an actual machine translation capability. In fact, most of the applications listed above can be initiated in a limited way by one-way phrase translation, and the results can be used to help define domains and requirements for two-way speech translation.

Lincoln Laboratory experience in interacting with the military regarding applications of speech translation dates back to a visit to Korea in October 1994, where we presented a design for a speech translation system and discussed potential application tasks with the military. The discussion, which was similar to other discussions with prospective military users, went something like this. The technologists (in this case, Lincoln) say that they can develop a speech translation system for a limited domain if the users can help define the domain (i.e., the vocabulary and grammar) and can help obtain training data. The users indicate strong interest in speech translation and offer some general ideas about domains, but are not able to provide any sample dialogs or training data. Basically, the problem is that the technologists would have to provide a working, user-friendly system with a reasonably open vocabulary and grammar before the users would be able to provide any useful feedback and data. And providing a working, user-

friendly system, which would translate in a domain of interest to the military is a task, which has been beyond the state-of-the-art. So we were in a chicken-and-egg situation. During our first visit to Korea in 1994, we redirected our effort from speech translation to text translation, in accordance both with user priorities and with where the most likely avenue for success pointed. We did have substantial success in text translation for US Forces Korea. But with regard to two-way speech translation for actual military applications, it appears to us that the dilemma, which we encountered in 1994 still applies. Open domain two-way speech translation has still not been achieved, and specific, limited-domain military applications for two-way translation haven't yet been defined or exercised. This has prevented the user feedback, which would be necessary for a successful development. Since military personnel generally have little time to assist in testing and evaluating new technology, we suggest that the development of 2-way speech translation technology is best done in a laboratory environment. This will require, among other things, working real-time systems and a set of personnel speaking various languages to train and evaluate the systems. More discussion of R&D directions for speech translation is presented in the next section.

5. R&D DIRECTIONS FOR 2-WAY SPEECH-TO-SPEECH TRANSLATION

Despite the significant past R&D efforts and the substantial progress that has been made in speech translation technology, it is clear that many challenges remain and that a substantial, sustained R&D efforts are needed to address these challenges. R&D is needed in development of component technologies, development and collection of speech and language resources, corpus collection, full end-to-end system development, and system evaluation for a substantial set of tasks and corpora. The R&D will continue to require close cooperation among researchers and users speaking multiple languages, and hence will require international collaboration. C-STAR, VERBMOBIL, and NESPOLE provide good models and experience as a basis for continuing multilingual, multinational collaborations.

Some specific challenges, which need to be addressed, are outlined below. It should be noted that most of these are not new challenges, but that despite prior progress more R&D is needed to produce effective speech translation systems. Challenges include:

- (1) Robust recognition of spontaneous speech—word error rates are still too high (typically >10%) for vocabularies of reasonable size (>2,000) words;
- (2) Robust parsing & understanding of spontaneous speech—this involves development of robust understanding system which incorporate both syntactic and semantic constraints, and can deal effectively with spontaneous inputs and with speech recognition errors;
- (3) Application of dialog constraints—more work is needed in utilizing dialog information to assist in obtaining correct translations;
- (4) The big challenge of dealing with many languages—this probably will eventually require an interlingua approach, and developing an interlingua approach which extends effectively not only across multiple languages but also across multiple task domains is a long-term challenge;
- (5) Evaluation—this will require substantial effort including development of effective, usable methodologies, definition of tasks and collection of corpora, and coordination among groups dealing with different languages; one direction which could aid in evaluation would be to attempt to calibrate machine translation performance to the human performance ILR levels, to provide a standard frame of reference for people working in defense-related activities; a possible scenario for evaluating a speech-to-speech machine translation system would be to utilize [Child 1993] the Oral Proficiency Interview to assess how much improvement in understanding is provided by the speech-to-speech machine translation system;
- (6) Bootstrapping—to make progress, especially with military (or non-military) user involvement, it will be necessary to develop initial systems which are robust and open enough to get users, or at least experimental subjects, talking over them so data can be collected and iterative development can proceed;
- (7) Combining efforts from multiple laboratories in a coherent way—this is a significant challenge, which must continue to be addressed for a multi-lingual effort.

Many of the organizations cited in the state-of-the-art review above are continuing to work on various aspects of speech-to-speech translation. To our knowledge, there is currently no ongoing DoD R&D program in speech translation. (DARPA is supporting a substantial effort in text translation from selected foreign languages to English under the Translingual Information

Detection, Extraction, and Summarization program [TIDES 2002].) However, DARPA is planning to embark, in the near future, on a new speech translation program. The Project is called Babylon, and a Workshop [Babylon 2001] to define program directions was held in Santa Monica on October 18-19, 2001. We should note that Babylon is not projected as a long-term research effort, but rather as a two-year effort aimed at speech translation in a Personal Digital Assistant (PDA). Its stated objective is to develop a very small (PDA-size), speech-to-speech translation system for use in the tactical environment. The proposed Babylon strategy is to achieve the small size and high accuracy by constraining the vocabulary (and corpus) to job specific tasks, where the tasks can be changed by the user or automatically.

Assuming that the Babylon Project goes forward as projected, it is certain to make significant contributions to advancing the state-of-the-art in speech-to-speech translation and in enabling military applications. But additional, long-term, multinational research efforts will continue to be needed to address the significant challenges of speech-to-speech translation.

REFERENCES

- [Alshawi 2000] Alshawi, H.; Bangalore, S.; Douglas, S., "Head-Transducer Models for Speech Translation and their Automatic Acquisition from Bilingual Data," *Machine Translation journal*, vol. 15 number 1, 2000, pp. 105-124.
- [Angelini 1997] Angelini B., Cettolo M., Corazza A., Falavigna D., Lazzari G., "Multilingual Person to Person Communication at IRST," *ICASSP 97*, pp. 91-94.
- [ATR 2002] ATR web site: <http://www.slt.atr.co.jp/>.
- [Babylon 2001] Meeting website: <http://www.dsic-web.net/ito/meetings/babylon2001oc>.
- [Bangalore 2000] S. Bangalore, G. Riccardi, "Finite-State Models for lexical reordering in Spoken Language Translation," *ICSLP 2000 Proceedings*, Beijing, China, pp. 422-425.
- [Black 1995] A. Black and N. Campbell, "Optimizing Selection of Units from Speech databases for Concatenative Synthesis," *Proc. Eurospeech 1995*, pp. 581-584.
- [Blanchon 2000] H. Blanchon, C. Boitet, "Speech Translation for French within the C-STAR-II Consortium and Future Perspectives," *ICSLP 2000 Proceedings*, Beijing, China, pp. 412-417.
- [Block 1997], H. Block, "The Language Components in VERBMOBIL," *ICASSP 97*, pp. 79-82.
- [Boitet 2000] C. Boitet, J. Guilbad, "Analysis into a Formal Task-Oriented pivot without Clear Abstract Semantics is Best Handled as "Usual" Translation," *ICSLP 2000 Proceedings*, Beijing, China, pp. 436-443.
- [Bub 1997] VERBMOBIL: T. Bub, W. Wahlster, A. Waibel, "The Combination of Deep and Shallow Processing for Spontaneous Speech Translation," *ICASSP 97*, pp. 71-74.
- [Campbell 1996] N. Campbell, "CHATR: a High Definition Speech Resequencing System," *Acoustical Society of Japan, Third Joint Meeting*, Dec. 1996.
- [Child et al. 1993] Child, James R, Ray T. Clifford, Pardee Lowe, Jr. Proficiency and Performance in Language Testing. *Applied Language Learning*. 1993. Vol. 4. Nos. 1 and 2. pp 19-54.
- [Corazza 1999] Corazza, A., ITC - IRST, Italy, "An Inter-Domain Portable Approach to Interchange Format Construction", *EUROSPREECH 1999 Conference Proceedings*, vol.6, Budapest, Hungary, p.2419-2422.
- [C-STAR 1996] C-STAR II Proceedings of the Workshop, ATR International Workshop on Speech Translation, September, 1996, Kyoto, Japan.
- [C-STAR 1999] C-STAR 1999 Workshop Proceedings, September 1999, Schwetzingen, Germany.
- [C-STAR 2001] C-STAR Web Site, <http://www.is.cs.cmu.edu/cstar/>.

[DLI 2002] Defense Language Institute Foreign Language Center website:
<http://pom-www.army.mil/pages/dliflc.htm>.

[DLPT IV 1991] Defense Language Proficiency Test (DLPT IV) Familiarization Guide. DLIFLC Pamphlet 350-14. 1 March 1991. Defense Language Institute. Foreign Language Center. Presidio of Monterey, CA 92944-5006.

[Engel 2000] R. Engel, "CHUNKY: an Example Based Machine Translation System for Spoken Dialogs," ICSLP 2000 Proceedings, Beijing, China, pp. 426-429.

[ETRI 2001] ETRI research web site: <http://comso.etri.re.kr/eng/>.

[Finke 1997] Finke M., Geutner P., Hild H., Kemp T., Ries K., Westphal M., "The Karlsruhe VERBMobil Speech Recognition Engine," ICASSP 97, pp. 83-86.

[Frederking 1997] Frederking, R., Rudnicky, A., and Hogan, C., "Interactive Speech Translation in the DIPLOMAT Project," presented at the Spoken Language Translation workshop at the 35th Meeting of the Association for Computational Linguistics, ACL-97. Madrid, Spain. 1997.

[Gates 97], Gates, D. A. Lavie, L. Levin, A. Waibel, M. Gavalda, M. Woszczyna and P. Zhan. "End-to-end Evaluation in JANUS: a Speech-to-speech Translation System". In Dialogue Processing in Spoken Language Systems: Revised Papers from ECAI-96 Workshop, E. Maier, M. Mast and S. LuperFoy (eds.), LNCS series, Springer Verlag, June 1997, <http://www.cs.cmu.edu/afs/cs/user/alavie/WWW/papers/ECAI-96-evaluation-book.ps>.

[Glass 1994] J. Glass, J. Polifroni and S. Seneff, "Multilingual Language Generation Across Multiple Domains," 1994 International Conference on Spoken Language Processing, Yokohama, Japan, 1994.

[Gruhn 2000] B. Gruhn, et al., "Cellular-Phone Based Speech-to-Speech Translation System ATR-MATRIX," ICSLP 2000 Proceedings, Beijing, China, pp. 448-451.

[Hiraoka 1997] Hiraoka S., Hoshimi M. (MRIT); Matsui K. (CRL); Junqua J. (STL), "An Experimental Bidirectional Japanese/English Interpreting System Video Phone System Using Internet," ICASSP 97, pp. 115-118.

[Hutchins 1992] W. J. Hutchins and H. L. Somers, *An Introduction to Machine Translation*, Academic Press Limited, 1992.

[Iida 1998] H. Iida, H., "Speech Communication and Speech Translation," proceedings of the Workshop on Multilingual Information Management: Current Levels and Future Abilities. Granada, Spain, 1998.

[Jeon 1998] J. Jeon, S. Cha, M. Chung, J. Park, K. Hwang, "Automatic Generation of Korean Pronunciation Variants by Multistage Applications of Phonological Rules," ICSLP 1998 Conference Proceedings.

[Kay 1994] Kay, M., J.M. Gawron, and P. Norvig, *Verbmobil: A Translation System for Face-to-Face Dialog*, Lecture Notes No. 33, Stanford University Center for the Study of Language and Information, ISBN (paperback) 0937073954, 1994.

[Lavie 1997] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavaldo, T. Zeppenfeld, P. Zhan, "JANUS-III: Speech-to-Speech Translation in Multiple Languages," ICASSP-97 Proceedings, Vol. I, pp.99-102, April 1997.

[Lavie 2001-1] A. Lavie, L. Levin, T. Schultz, C. Langley, B. Han, A. Tribble, D. Gates, D. Wallace, K. Peterson, "Domain Portability in Speech-to-Speech Translation," DARPA/NSF Human Language Technology (HLT2001) Conference Proceedings, March 2001.

[Lavie 2001-2], A. Lavie, C. Langley, A. Waibel, F. Pianesi, G. Lazzari, P. Coletti, "Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-Commerce Applications," HLT2001 Conference Proceedings, March 2001.

[Lazzari 1999] G. Lazzari, ed., "Speaker-Language Identification and Speech Translation." Chapter 7 in E. Hovy and N. Ide, eds., *Multilingual Information Management: Current Levels and Future Abilities*, A report Commissioned by the US National Science Foundation and also delivered to the European Commission's Language Engineering Office and the US Defense Advanced Research Projects Agency, April 1999. (This report is available as <http://www.cs.cmu.edu/~ref/mlim/index.html>. It has also been published, in *Linguistica Computazionale*, Volume XIV-XV, by the Insituti Editoriali e Poligrafici Internazionali, Pisa, Italy, ISSN 0392-6907.)

[Lazzari 2000] G. Lazzari, "Spoken Language Challenges and Opportunities," ICSLP 2000 Proceedings, Beijing, China, pp. 430-435.

[Lee 1999-1] Lee, Y. S., and Weinstein, C. J., *An Integrated Approach to English/Korean Translation and Translingual Information Access*. C-STAR (Consortium for Speech Translation Advanced Research) Workshop, Schwetzingen, Germany, September 23-24, 1999.

[Lee 1999-2] Lee, Y. S., Clifford J. Weinstein, C. J., and Hong, S. H., *Machine Assisted Language Translation for U.S./ROK Combined Forces*. Army RD&A Magazine, November-December 1999, pp. 38-41.

[Lee 2002] Y.S. Lee, D. Sinder, C. Weinstein, "Multi-lingual Speech translation: A Case Study from CCLINC English/ Korean 2-Way Speech Translation," Journal Article submitted for publication to *Machine Translation Journal*.

[Levin 2000] Levin, L., A. Lavie, M. Woszczyna, A. Waibel, "The JANUS-III Translation System, "Machine Translation," <http://www.cs.cmu.edu/afs/cs.cmu.edu/user/alavie/www/papers/mt-special.ps>.

[Matsui 2001] K. Matsui, Y. Wakita, T. Konuma, K. Mizutani, M. Endo, M. Murata, *An Experimental Multilingual Speech Translation System*," ACM Workshop on Perceptive User Interfaces, PUT 2001, Orlando, FL, November 2001.

[NESPOLE 2002] NESPOLE web site, "Negotiating through Spoken Language in E-commerce," <http://nespole.it.it/>

[Ney 2001] H. Ney, F. J. Och, S. Vogel, "The RWTH System for Statistical Translation of Spoken Dialogues," proceedings of HLT 2001 Workshop on Human Language Technology, March 2001, Morgan Kaufmann, pp. 302-307.

[Niemann 1997] H. Niemann, E. Noth, A. Kiesling, R. Kompe, A. Batlinger, "Prosodic Processing and Its Use in VERBMOBIL," ICASSP 97, pp. 75-78.

[NIST 2002]. Machine Translation Benchmark Tests at the National Institute of Standards and Technology. Project Website:
<http://www.nist.gov/speech/tests/mt/mt2001/index.htm>.

[Rayner 1993] Rayner, M. et al., "A speech to speech translation system built from standard components," proceedings of the 1993 ARPA Human Language Technology Workshop. Princeton, New Jersey.

[Rayner 1997] Rayner M., Carter D. (SRI International), "Hybrid Language Processing in the Spoken Language Translator," ICASSP 97, pp. 107-110.

[Reithinger 1999] Reithinger, N. DFKI GmbH, Germany, "Robust Information Extraction in a Speech Translation System", vol.6, p. 2427.

[Roe 1992] Roe, D.B., F.C. Pereira, R.W. Sproat, M.D. Riley, "Efficient grammar processing for a spoken language translation system." proceedings of ICASSP-92, vol. 1 (213—216). San Francisco.

[Sarich 2001] DARPA One-Way project: Marine Acoustics website
(<http://www.sarich.com/translator>).

[Seneff 1992]. Seneff, "TINA: A Natural Language System for Spoken Language Applications," Computational Linguistics, 18:1, pp. 61-88, 1992.

[Shimizu 1997] Shimizu T., Singer H., Sagisaka Y. (ATR-ITL), "Fast Word-Graph Generation for Spontaneous Conversational Speech translation," ICASSP 97, pp. 95-98.

[SIL 2002] The ILR (FSI) proficiency scale.
<http://www.sil.org/lingualinks/LANGUAGELEARNING/MangngYrLnggLrnnngPrgrm/TheILRFSIProficiencyScale.htm> Summer Institute of Linguistics.

[Sugaya, 1999] Sugaya, F., Takezawa, T., Yokoo, A., Yamamoto, S. ATR ITR Labs, Japan, "End-to-End Evaluation in ATR-MATRIX: Speech Translation System between English and Japanese", EUROSPEECH 1999 Conference Proceedings, vol. 6, Budapest, Hungary, pp. 2431-2434.

[Tessiere 2000], L. Tessiere and W. Hahn, "Functional Validation of a Machine Interpretation System: Verbmobil". In [Wahlster 2000], pp. 611-631.

[TIDES 2001] DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) Project website:
<http://www.darpa.mil/ito/research/tides/index.html>.

[Tummala 1995] Tummala, D., Seneff, S., Paul, D. B., Weinstein, C. J., and Yang, D., CCLINC: System Architecture and Concept Demonstration of Speech-to-Speech Translation for Limited-Domain Multilingual Applications, Proceedings of the 1995 ARPA Spoken Language Technology Workshop, Austin, TX, January 1995, pp. 227-232.

- [Vidal 1997], E. Vidal, "Finite-State Speech-to-Speech Translation," ICASSP 97, pp. 111-114.
- [Wahlster 2000] W. Wahlster, ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer-Verlag, ISBN 3540677836, September 2000.
- [Waibel 1996] Waibel, A., M. Finke, D. Gates, M. Gavalda, T. Kemp, A. Lavie, M. Maier, L. Mayfield, A. McNair, I. Rogina, K. Shima, T. Sloboda, M. Woszczyna, T. Zeppenfeld, and P. Zahn, "JANUS-II-Translation of Spontaneous Conversational Speech," Proceedings of ICASSP (pp. 409-412).
- [Watanabe 2000] T. Watanabe, A. Okumura, S. Sakai, "An Automatic interpretation System for Travel Conversation," ICSLP 2000 Proceedings, Beijing, China, pp. 444-447.
- [Weinstein 1996] Weinstein, C. J., Tummala, D., Lee, Y. S., and Seneff, S., Automatic English-to-Korean Text Translation of Telegraphic Messages in a Limited Domain, C-STAR II Proceedings of the Workshop, ATR International Workshop on Speech Translation, September 1996, Kyoto, Japan.
- [Weinstein 1997] Weinstein, C. J., Lee, Y. S., Seneff, S., Tummala, D. R., Carlson, B., Lynch, J. T., Hwang, J. T., and Kukulich, L. C., Automated English/Korean Translation for Enhanced Coalition Communications, Lincoln Laboratory Journal, Vol. 10, No. 1, pp. 35-60, 1997.
- [WIRED 2000], "Hello, World (past, present, and future of machine translation)," WIRED Magazine, May 2000, pp. 220-235,
http://www.wired.com/wired/archive/8.05/timeline_pr.html.
- [White 1994] J. White and T. O'Connell, "Evaluation in the ARPA Machine Translation Program: 1993 methodology," Proceedings Human Language Technology Workshop, Morgan-Kaufmann publishers, pp. 135-140, March 1994.
- [Yamamoto 2000] S. Yamamoto, "Toward Speech Communication Beyond Language Barrier – Research of Spoken Language Technologies at ATR", ICSLP 2000 Proceedings, Beijing, China, pp. 406-411.
- [Yang 1997] J. W. Yang and J. Park, "An Experiment on Korean-to-English and Korean-to-Japanese Spoken Language Translation," ICASSP-97 Proceedings, Vol. I, pp. 87-90, April 1997.
- [Zong 2000-1] C. Zong, T. Huang, B. Yu, "An Improved Template-Based Approach to Spoken Language translation," ICSLP 2000 Proceedings, Beijing, China, pp. 440-443.
- [Zong 2000-2] C. Zong, Y. Wakita, K. Matsui, Z. Chen, "Japanese-to-Chinese Spoken Language Translation Based on the Simple Expression," ICSLP 2000 Proceedings, Beijing, China, pp. 418-421.