[MT Summit IV, July 20-22, 1993, Kobe, Japan]

Automatic Speech Translation at ATR

Tsuyoshi Morimoto
[ATR Interpreting Telecommunications Laboratories]
Akira Kurematsu
[University of Electro-Communications]
Contact Person: Tsuyoshi Morimoto
ATR Interpreting Telecommunications Laboratories
Seika-cho, Souraku-gun, KYOTO, 619-02
07749-5-1301, 07749-5-1308
morimoto@itl.atr.co.jp

1. Introduction

Since Graham Bell first invented the telephone in 1876, it has become an indispensable means for communications. We can easily communicate with others domestically as well as internationally. However, another great barrier has not been overcome yet; communications between people speaking different languages.

An interpreting telephone system, or a speech translation system, will solve this problem which has been annoying human-being from the beginning of their history. The first effort was made by *NEC;* they demonstrated a system in Telecom'83 held in Geneva. In 1987, British Telecom Research Laboratories implemented an experimental system which was based on fixed phrase translation [Stentiford]. At Carnegie-Mellon University (CMU), a speech translation system was developed on doctor patient domain in 1988 [Saitoh]. These systems were small and simple, but showed the possibility of speech translation.

In Japan, on ATR Interpreting Telephony Research project, started in 1986 and terminated in March 1993, focused on basic research for speech translation and obtained fruitful results. Following that project, an Interpreting Telecommunication project was newly initiated.

This paper reports the technologies attained in the preceding project, and also describes the objectives of the new project.

2. Current Status of Basic Technologies for Speech Translation

In principle, three componential technologies are essential; speech recognition, language translation and speech synthesis. Furthermore, techniques concerning how to integrate speech recognition and language analysis are important. In this section, technologies attained so far are described.

2.1 Speech Recognition

Basically, two kinds of models are necessary for speech recognition; phonetic model and language model. For phonetic modeling, a Hidden Markov Model (HMM) approach was employed. Phone is apt to be acoustically affected by preceding and/or succeeding phones, so hundreds of allophone models are generated automatically from a huge speech database by use of the "successively-state-

splitting (SSS) algorithm [Takami]". For the language model, general context free grammar (CFG) was used. Comparing to other conventional language models such as bi-gram or tri-gram, it is superior in extendibility and maintainability. A new mechanism, a predictive LR parsing mechanism which is an extension of the generalized LR parsing algorithm, combines these two models dynamically and recognizes input continuous speech [Kita]. In this method, CFG rules are complied and converted to an LR table. The parser refers to the table and predicts the next possible phones, then verifies their existence in input speech by comparison with corresponding HMMs (Fig. 1). As will be shown in latter section, this method attains very high recognition rate.

As for non-specific user's speech, a speaker adaptation approach was adopted. By introducing the "vector field smoothing (VFS) algorithm [Ohkura]", only about ten words are sufficient to adapt to a new speaker's speech.

2.2 Integration of Speech Recognition and Language Analysis

The system accepts speech uttered phrase by phrase (Japanese bunsetsu) so that the speech uttered clearly. To treat such utterances, Japanese phrasal grammar rules are defined in the speech recognizer. In addition to them, sentential level (inter-phrasal) grammar rules are defined, which are used by the sentence recognition controller. It controls inter-phrase level parsing, and, coping with the phrase recognizer, recognizes input sentences as a whole rather than independent phrases. Then, all outputs from the recognizer are almost syntactically correct.

There still, however, remain several ambiguities in the outputs from the recognizer because only syntactic constraints are used in the process. To solve the problem, not only the best candidate (the best hypothesis) but several candidates (N-best hypotheses) are output from the recognizer. In the next step (a analyzer of Japanese) accepts such N-best hypotheses, and chooses the most plausible one that satisfies more accurate linguistic (syntactic and semantic) or even pragmatic constraints.

2.3 Spoken Language Translation

The style of spoken sentences is, especially in Japanese, quite different from that of written sentences. Spoken sentences includes various intentional expressions or ellipses. To treat such sentences, a new method called the "intention translation method [Kurematsu]" was developed (Fig. 2).

An input utterance is analyzed by the analyzer based on HPSG (and its Japanese version JPSG) grammar formalism and unification operation. In each lexical entry, syntactic, semantic and even pragmatic constraints are defined in a form of feature structures. In this paradigm, the inefficiency caused by the unification operation is the biggest issue, and various efforts have been made such as introducing medium-grained CFG rules [Nagata] or implementing a quasi-destructive graph unification algorithm [Tomabechi] to solve this issue. With these efforts, the processing time has been drastically decreased.

The next transfer component is composed of three phases; zero-anaphora resolution, illocutionary force type determination and conversion of source-language semantics to target-language semantics.

In spoken Japanese, words that are easily inferable from the context tend to be omitted. In particu-

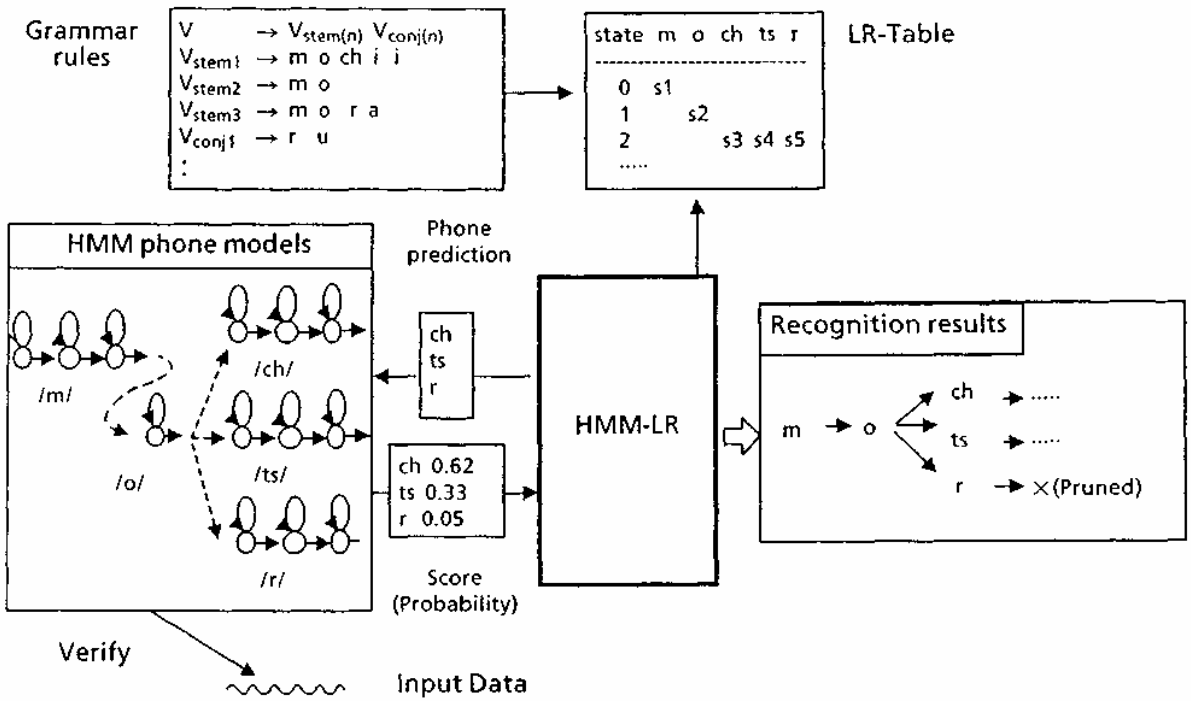

Figure 1. Basic Mechanism of Speech Recognition

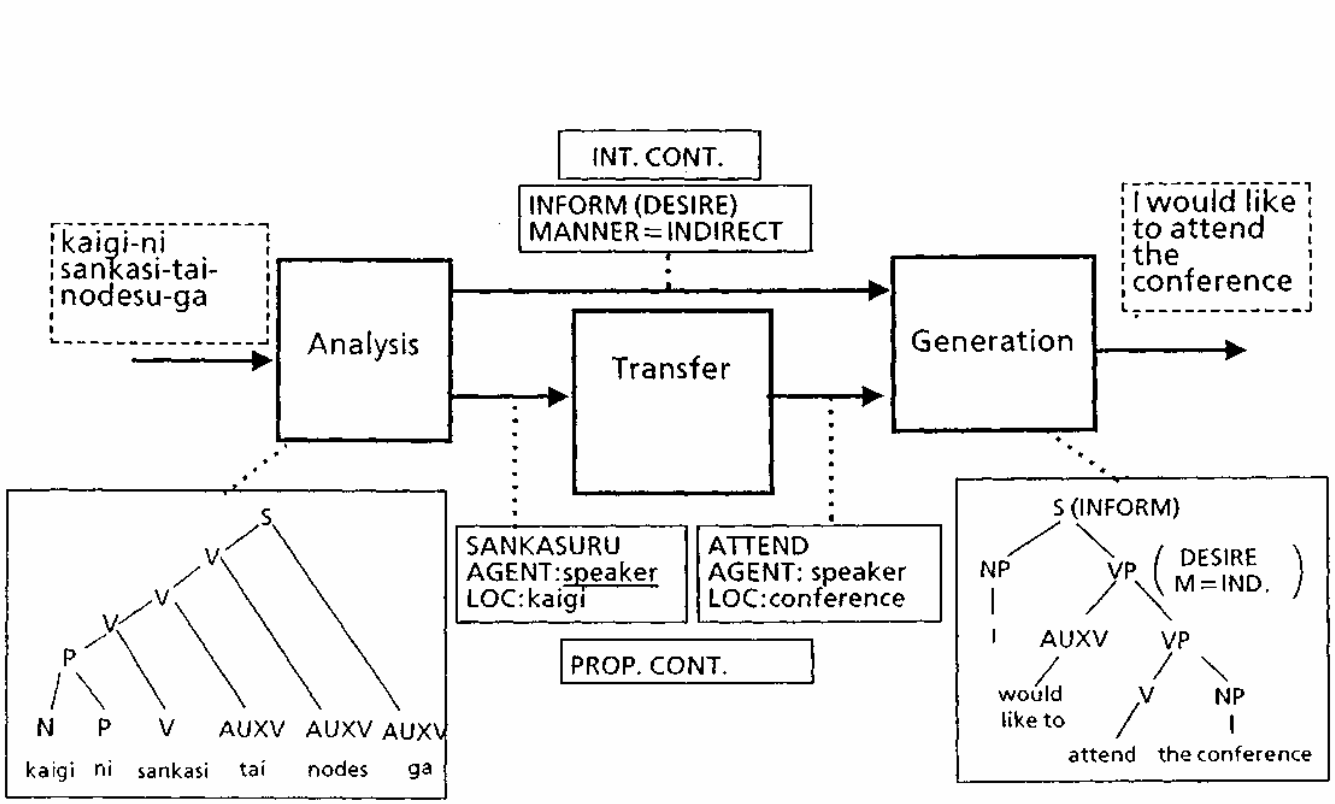

Figure 2. Intention Translation Method

lar, 'I' and 'you' are seldom uttered explicitly. In many cases, such zero-anaphora can be resolved by the use of pragmatic information such as honorifics appearing in the sentence.

The semantics of an input utterance can be divided into two parts; an intentional content part and a propositional content part. Roughly speaking, the former part indicates the speaker's intention or attitude, and the latter is a neutral proposition. From the former, an illocutionary force type of the utterance is determined. Typical illocutionary force types are shown in Table 1. In other words, an intentional content part is converted to language-independent concepts. A propositional content part described in Japanese concepts is converted to corresponding target-language concepts.

Table 1. Typical Illocutionary Force Type

Type	Explanation
PHATIC	Phatic expression such as those to open or close dialogue (Hello, Thank you)
INFORM	Inform a hearer of some facts
REQUEST	Request a hearer to carry out some action (Please tell me ···)
QUESTIONIF	Yes/No question
QUESTIONREF	WH question

The final component is generation. It accepts semantic feature structures in which both a illocutionary force type and a propositional content are described. The task of the generation system is to generate a syntax tree corresponding to input semantic feature structures. For this purpose, a set of sub-trees annotated with semantic feature structures (called "phrase definition" or PD) is defined in the system [Kikui]. Such PD is defined for each basic phrase structure as well as for each typical idiomatic expression of the target language. During generation, a set of PDs that can subsume the whole semantic feature structures of the input is selected, and combined by unification operation. Finally, a succession of lexical words appearing at the bottom of the generated syntax tree is output as a result. Some translation examples are shown in Table 2.

2.4 Speech Synthesis

In conventional speech synthesis, uniform speech units such as CVC (consonant-verbal-consonant) or VCV are prepared and the target speech is generated by connecting such units. In this approach, synthesized speech is not clear or natural enough because of distortion caused by the concatenation.

To improve the quality, a new method called Nyu-talk has been developed [Sagisaka]. In the method, various non-uniform units are extracted from huge speech database and stored in a synthesis speech file.

Table 2. Translation Example

No.	Input	Translated Output
Ex.1	会議の内容について教えて下さい。	Please tell me about the content of the conference.
Ex.2	今回の会議の話題は通訳電話です。	The topic of the conference this time is interpreting telephony.
Ex.3	私は英語が全然分からないのですが。	I don't understand English at all.
Ex.4	日本語への同時通訳を用意しております。	Simultaneous interpretation into Japanese is available.

For a sentence to be synthesized, the system dynamically selects the best combination (i.e. makes the least distortion) from these non-uniform units. Finally, prosody of the output speech is controlled according to the syntactic structure of the sentence.

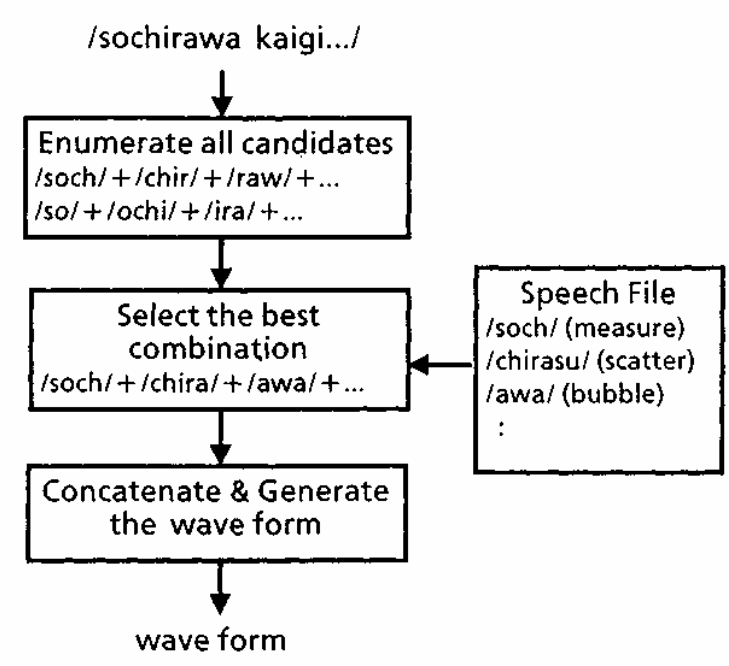

Figure 3. Speech Synthesis

3. ATR Speech Translation System

In 1989, ATR developed the first version of an experimental speech translation system from Japanese to English called SL-TRANS [Morimoto]. Then, many improvements both in the mechanisms and the efficiency were made, and the final version called ASURA was implemented. Most of the technologies mentioned above are integrated in it except for an English speech synthesizer; for this component, a commercial English speech synthesizer (DecTalk) is used.

The target domain is inquiry about "international conference". Experiments have been conducted over 12 dialogues, which cover various varieties of topics such as inquiry on "how to register", "how to cancel", "sight seeing tour", "hotel arrangement", etc.

Two versions which differ in vocabulary size have been developed. The first one (hereafter called "the standard version") covers all of the 12 dialogues and about 700 words are defined in the system. The second one, whose lexicon size is about 1,500, is more extended one (we call it "the extended version"), and covers not only the dialogues, but also more than 90 percent of standard expressions in Japanese spoken sentences.

The overall speech-to-speech translation accuracy of these two versions are shown in Table 3. On the standard version, more than 90 percent of the utterances are recognized and translated properly. On the extended version, the accuracy drops to about 63 percent. This is mainly due to an increase of ambiguity generated by the translation part. The processing time of the standard version and the extended version are about 25 seconds and 50 seconds, respectively (when two HP9000/750 are used). You might think that it is slightly too long. In the future, however, hardware innovation can still be expected and then near real-time processing might be achieved.

Table 3. Performance of ASURA

	Speech Recognition Rate	Translation Rate (System Total)
Standard Version	86.7% (1st) 92.5% (≧3rd)	90.3%
Extended Version	82.2% (1st) 90.7% (≧3rd)	63.3%

4. International Joint Experiment

ATR in Japan, CMU in the United States and Siemens Corporation/Karlsruhe University (KU) in Germany agreed to collaborate mutually in the area of speech translation, and started a consortium called C-STAR (Consortium for Speech Translation Advanced Research) a couple of years ago. The three parties decided to carry out an international joint experiment on an automatic interpreting telephone system, by interconnecting their speech translation systems. The parties shared equal responsibility; each site developed a speech recognition part and a speech synthesis part for its own language and a language translation part to the other two languages. In ASURA, all kinds of linguistic knowledge and the processing programs that use them are completely separated. Then, only transfer rules from Japanese to German and generation rules for German have newly been developed for Japanese-to-German translation. Other components such as Japanese analysis were used in common with those of Japanese-to-English translation. Consequently, a Japanese-German translation system was developed in very short time. The total system configuration for the experiment is shown in Fig. 4.

The experiment was conducted on January 28th, 1993. Several dialogues out of 12 were used in the experiment. In addition to the speech translation system, a Teleconference system were used so that the speaker could see what was going on at the different end. A large audience including the press and TV attended. As a whole, the experiment was successful and received a favorable evaluation.

5. For Further Enhancement and Extension

Interest in speech translation research has been growing; some works have been stimulated by the ATR interpreting telephony project. In the United States, several institutes such as CMU or Bell Labs. have been making efforts in speech translation research. In Germany, the VERBMOBIL project was recently launched, whose aim is to develop a portable face-to-face speech translation system. The same kind of big national project has also started in South Korea.

Most of their goals are very exciting and ambitious, i.e. "speech translation of spontaneous utterances".

At ATR, a new research organization (ATR Interpreting Telecommunications Research Laboratories) has recently been established supported by the Japan Key Technology Center and various private enterprises. It is the successor of the preceding project and will engage in basic research on advanced speech translation. In this section, a brief introduction of the new project will be given.

5.1 Objective

The objective of the project is to develop key technologies for translation of spontaneously or naturally spoken utterances. Such utterances include wide varieties of speech and language phenomena, which have not so much been investigated until now. In speech, phenomena such as strong coarticulation, phone variation depending on an individual person, collapsed or missing phone, etc., will appear quite often. On the other hand, prosody plays an important role for conveying extra-linguistic information like a speaker's intention. As language phenomena, fragmental and strongly context dependent utterances, inversions, repeating or re-phrasing, ungrammatical expressions, etc., will appear. The target area of the new project is shown in Fig. 5; the covered area by the preceding project is also indicated. In the following section, problems and approaches to be pursued are described.

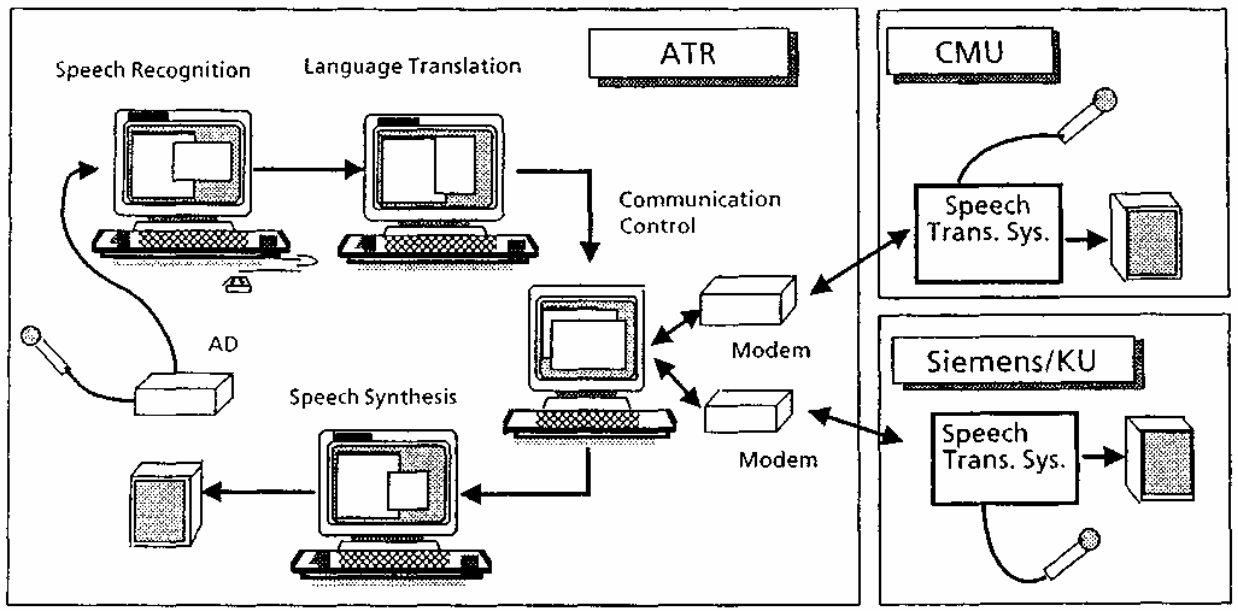

Figure 4. International Joint Experiment

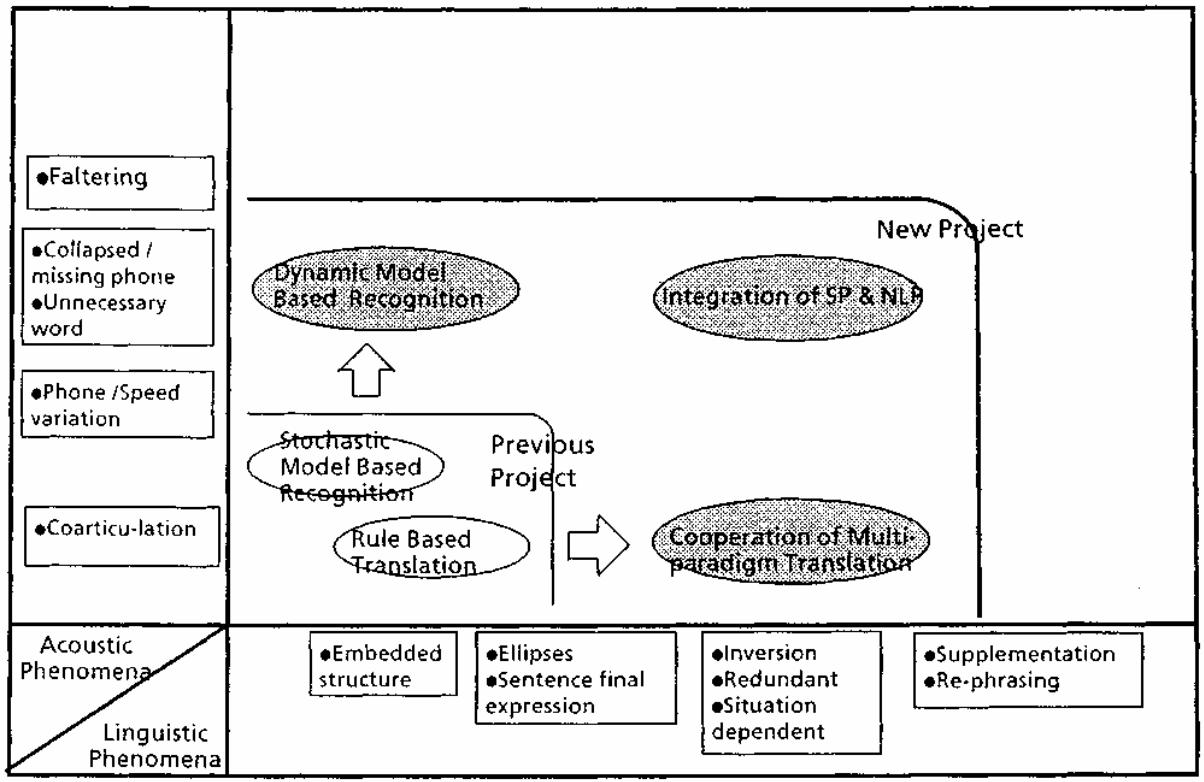

Figure 5. Target of the New Project

5.2.1 Recognition of Spontaneous Speech

Speech recognition must be robust enough against both acoustic and linguistic variations. In acoustic research, much effort will be paid to developing more precise and robust allophonic models to cover wide acoustic variations. In that way, effects from the linguistic environment might be carefully considered. The recognition of a non-specific user's speech is also important. Some dynamic speaker adaptation mechanism must be established so as to eliminate the undesirable necessity of uttering few words in advance. Adding to these themes, the problem of how to define and manage a language model should be investigated. Unnecessary or unimportant words such as "Uh" or "Oh" would be inserted frequently in spontaneous utterances. Difficulties may be the treatment of colloquial expressions like inversion and so on. Some management mechanism of a language model will be necessary, which will interact with a higher level language processing component and restructure the model dynamically according to progress of the dialogue.

5.2.2 Prosody Extraction and Control in Speech Synthesis

Prosody such as intonation, power and speed will play an important roll in spontaneous speech. It helps not only to resolve ambiguities in sentence meanings, but sometimes to give the extra-linguistic information such as the speaker's attitude, intention, or even emotion. Efforts will be made to establish an algorithm to extract prosody from speech and control it in speech synthesis.

5.2.3 Translation of Colloquial or Spontaneous Utterances

Most conventional translation is carried out by the use of several kinds of linguistic rules such as analysis, transfer or generation rules. It is based on the idea that all linguistic phenomena can be captured and written down as rules. However, we frequently observe various phenomena out of the cases. On the other hand, a new translation paradigm, called the "example based translation" approach, has recently attracted considerable attention. It translates an input by using a set of translation examples each of which is very similar to a portion of the input. Such examples are extracted from a large bilingual corpus. This approach seems to be very promising for translation of spontaneous utterances. However, if such examples are used without any linguistic knowledge or principle, the results would be disappointing. We believe that the best way is to somewhat integrate a rule-based approach and an example-based approach; should these two algorithms collaborate with each other, the most likely translation is generated. At the same time, the dialogue situation at that point should be taken into consideration to translate the very context dependent expressions properly.

5.2.4 Integrated Control of Speech and Language Processing

Especially in spontaneous speech translation, the integrated control of speech and language processing becomes very important. Appropriate information necessary for language models should be provided to speech recognition from the language processing side, and speech information such as prosody should be provided to language processing from the speech processing side as well. At the same time, the status of the dialogue should be recognized and maintained properly. Such situational information would be about the environment (such as the domain or the subject of the dialogue), the participants' statuses (such as their intentional or mental states) and the dialogue progression status (such as the topic or the focus). Such information would be referred to by both of

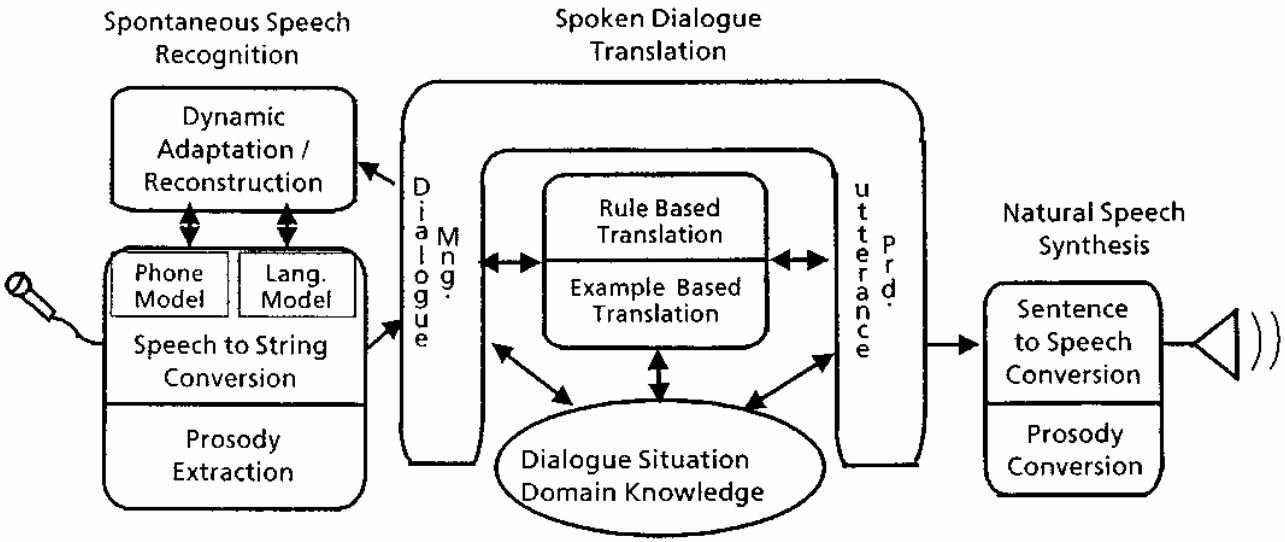

Figure 6. Future Speech Translation System

the speech processing and the language processing. The overall image of the future system would be like Fig. 6.

5.3 Time Schedule and Managemental Issues

The term of the project is seven years (from 1993.3 to 2000.3) and the total budget is expected to be 16 billion yen, which is nearly the same amount as the previous project's budget. The number of researchers will be about 50.

Considering the importance of international cooperation, the project ardently wants to have good collaborative relationships with outside active research organizations.

6. Conclusion

The main results achieved in the previous project are summarized below.

(1) Componential technologies necessary for developing a speech translation system have intensively been studied and a prototypical system has also been developed.

(2) An international joint experiment, connecting ATR's, CMU's and Siemens/KU's systems, has been conducted and was successful.

(3) Those efforts have shown the technical possibility of developing an "Interpreting Telephony System" in the near future.

The new project (following the previous project) was introduced. The mission of the project is to enhance and to extend the results attained in the previous project. We believe that these efforts will bring fruitful results, and let people in the world be able to speak freely without worrying about language differences at the beginning of the next century.

References

[Kita] K. Kita, T. Kawabata, H. Saito: "HMM Continuous Speech Recognition Using Predictive LR Parsing", ICASSP-89, 1989

[Kurematsu] A. Kurematsu, H. Iida, T. Morimoto K. Shikano: "Language Processing in connection with Speech Translation at ATR Interpreting Telephony Research Laboratories", Speech Communication, Vol. 10, No. 1, 1991

[Morimoto] T. Morimoto, M. Suzuki, T. Takezawa, G. Kikui, M. Nagata, M. Tomokiyo: "A Spoken Language Translation System: SL-TRANS2", COLING-92, 1992

[Nagata] M. Nagata: "An Empirical Study on Rule Granularity and Unification Interleaving Toward an Efficient Unification-Based Parsing System", COLONG-92, 1992

[Ohkura] K. Ohkura, M. Sugiyama, S. Sagayama: "Speaker Adaptation based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs", ICSLP-92, 1992

[Sagisaka] Y. Sagisaka, N. Kaiki, N. Iwahashi, K. Mimura: "ATR Nyu-Talk Speech Synthesis System", ICSLP-92, 1992

[Saitoh] H. Saitoh, M. Tomita: "Parsing Noisy Sentences", COLING-88,1988

[Stentiford] F. Stentiford, M. Steer; "A Speech Driven Language Translation System", European Conf. on Speech Technology, 1987

[Takami] J. Takami, S. Sagayama: "Successive State Splitting Algorithm for Efficient Allophone Modeling", ICCASP-92, 1992

[Tomabechi] H. Tomabechi: "Quasi-Destructive Graph Unification with Structure Sharing", COLING-92, 1992