RESEARCH ON AUTOMATIC TELEPHONE INTERPRETATION

Akira Kurematsu

ATR Interpreting Telephony Research Laboratories
Seika-cho, Soraku-gun, Kyoto 619-02, Japan

1. Introduction

An automatic telephone interpretation system is envisaged which transforms automatically and simultaneously the spoken dialogue from the speaker's language to the listener's. Fundamentally, three constituent technologies are necessary for such a system: speech recognition, machine translation, and speech synthesis. These individual subsystems will then be integrated to form an automatic telephone interpretation system.

Since this system is a brand-new concept, numerous studies and evaluations must be made regarding its feasibility. Among the matters to be considered are the degree of performance that can be expected in each of the constituent technologies, along with the ease of use, or "user friendliness" of the system. ATR started basic research for automatic telephone interpretation in 1986, and extensive research has been undertaken in exploring each component technology.

2. Implications of An Automatic Telephone Interpretation System

Analysis of telephone conversations through an interpreter has revealed many interesting points. First, user friendly, machine aided interpretation is essential, since speech recognition and machine translation of natural spoken language is sometimes difficult even for a human interpreter. Secondly, the initial stage of an Automatic Telephone Interpretation system is an interactive dialogue translation system. The translation will be consecutive rather than simultaneous. Speaker and hearer can actively participate in the dialogue.

Functions to be supplied in the system will be as follows: display of questionable words, keyboard editing functions, informing speakers of the resulting translation, and parallel transmission of the speakers dialogue. These functions will compensate for less than perfect performance.

3. Research Directions

3.1 Speech Recognition

Since conversational speech is normally continuous, with most words running together, recognition of phrases of continuous speech is necessary. Reliable phoneme recognition and segmentation has been studied leading to considerable improvements over conventional approaches.

One effective approach to the problem of speaker independence is the incorporation of a system for speaker adaptation. A small number of words is used to adapt to speaker characteristics.

Prosodic information such as pitch, stress, and duration, along with information on syllable boundaries, will be used in order to increase the precision and speed of algorithms for word and phrase recognition. However, careful analysis will be needed to extract effective information, since prosodic features in Japanese spoken dialogue are not particularly stable.

3.2 Integrating Speech Processing and Language Processing

Integrating speech and language processing is an important area to be tackled by real time high speed software technology. This requires word prediction based on language model. A continuous speech recognition system combining HMM phoneme recognition with the generalized LR parsing algorithm has been successfully implemented. The linguistic constraints of syntactic, semantic and pragmatic information are utilized to narrow the number of word candidates. Information from the speech recognizer to the language processor is in the form of a phrase lattice. In language processing systems, a function that can use syntactic and semantic knowledge to select the most appropriate candidate is necessary.

3.3 Machine Translation

Spoken language differs from ordinary written language in both vocabulary and grammar. Dialogue interpretation requires intention extraction. Input utterances are analyzed according to unification-based lexico-syntactic, syntactico-semantic principles. Syntactico-semantic analysis permits an integrated description of information from various sources, and lexico-syntactic analysis provides modularity. The proposed translation method can be characterized by two translating processes: one which extracts intentions in utterances such as requests, promises, greetings etc. and another which transfers propositional parts of utterances. A method of analyzing dialogues is being developed using a discourse structure based on topic information and the discourse function of sentences. Both topic information and discourse functions are represented by feature structures as well as other syntactico-semantic information. Intra-sentential and inter-sentential structures can be analyzed using the same unification-based phrase structure framework.

It is essential that the telephone interpretation system be able to comprehend meaning in context in order to disambiguate expressions and compensate ellipses.

3.4 Speech Synthesis

A high quality speech synthesis system by rule must be realized for translation output. A speech synthesis system using flexible speech synthesis units of various lengths is under development. Another matter for research is prosody, which is closely correlated with meaning and naturalness. Further research into the incorporation of conceptual information in speech synthesis will be undertaken .

Individualization of synthetic speech will be accomplished by use of voice conversion from one speaker's to another's. High quality speech personalization is expected in cross-linguistic speech synthesis.

4. Conclusion

An extensive effort must be made to raise the level of technology in each of the areas of speech recognition, machine translation, and speech synthesis to realize an Automatic Interpretation of Telephone Conversations. Massive databases of speech and language information will be essential to further promote related research. Moreover, advanced high-speed processing technology will be required to implement a real-time automatic telephone interpretation system. Because of the vast complexity of natural language, however, this goal can be reached by specifying the application domain in the constrained area. Also, it will be necessary to consider the expandability of the system in terms of domain size, different domains, and multi-language application.