TR-I-0001

Automatic Telephone Interpretation: A Basic Study
自動翻訳電話の基礎研究

Akira Kurematsu
榑松 明

1987.5

Abstract

The Automatic Telephone Interpretation system is envisaged as a facility for enabling a person speaking in one language to communicate readily by telephone with someone speaking in a different language. It will do so by automatically transforming the dialog content from one speaker's language to that of the listner's, as it is being spoken. Three constituent technologies are necessary for construction of such a system: speech recognition, machine translation, and speech synthesis. In this report, an outline of research tasks and research approach for these fundamental technologies is presented.

ATR Interpreting Telephony Research Laboratories
ATR 自動翻訳電話研究所

1. Introduction

International communication is increasing in many countries. Along with the increase in opportunities for communicating with people of different languages, the problem arises as to what sort of language can be used to convey one's meaning most effectively.

Given the difficulty and importance of these problems, great hope is being placed in future technology to find a solution to them. The dream is that, by making use of the latest information processing technology, ways will be found to overcome language barriers and facilitate communication among different peoples of the world.

The automatic telephone interpretation system is envisaged as a facility for enabling a person speaking in one language to communicate readily by telephone with someone speaking in a different language. It will do so by automatically transforming the contents of a dialog from the speaker's language to that of the listener's, as it is being spoken.

Creation of such a system will first require developing the various constituent technologies for performing the separate tasks involved. These technologies will be used in building a speech recognition system for spoken word comprehension, a machine translation system for language conversion, and a speech synthesis system by which a computer will produce the voice of the "interpreter." These individual subsystems will then be combined to form the automatic telephone interpretation system.

Since this system is a brand-new concept, numerous studies and evaluations must still be made regarding its feasibility. Among the matters to be considered are the degree of performance capability that can be expected in each of the constituent technologies, along with the ease of use, or "user friendliness" of the system. According to a feasibility study on automatic interpretive telephony sponsored by the Japanese Ministry of Posts and Telecommunication, realizing such a system will require fifteen years of research and development.

The ATR Automatic Interpreting Telephony Research Laboratories were established in April 1986. Presently, a seven year basic research project necessary for realization of an automatic telephone interpretation system has been started.

2. Envision of the automatic telephone interpretation system

To depict the type of system the research project has in view, the functions of the system can be illustrated using this Japanese/English interpretation example.

Suppose a Japanese speaker picks up the phone and says,
Moshi-moshi, Satoh-san onegai-shimasu.

This utterance will enter the system, through the telephone mouthpiece, as a speech signal. The Japanese language speech recognition unit will analyze these series of sounds into individual Japanese words. Once the spoken words have been recognized, they will then be analyzed by the Japanese-to-English translation unit and transformed to an equivalent English utterance:

"Hello, Mr. Satoh, please."

This English will then be converted to a voice utterance, by the English speech synthesis unit, and will be output at the listener's end.

The listener may respond to this utterance with,

"Mr. Satoh is not at his desk."

The process will then be repeated, this time from English to Japanese. The Japanese equivalent of the response will thus be produced for the Japanese speaker, again as synthetic speech.

Exactly how the automatic telephone interpretation system develops will depend on the degree of technological progress that is made over the coming years. It can be anticipated, however, that neither speech recognition nor machine translation technologies will reach the point of 100 percent recognition or translation accuracy, at least within the foreseeable future. Accordingly, a number of functions will have to be provided to make the system more usable.

Basic functions that ought to be made part of the system are along the following lines.

(1) Automatic translation on/off function

This will allow the facility to be used only when translation is needed.

(2) Repeat function

When the spoken utterance cannot be recognized or translated, the system will notify the speaker and ask for it to be repeated more clearly or more simply.

(3) Selection of translation level

Conversations take place on many different levels of complexity. This function will allow the speaker to select in advance the required level of translation capability. For example, just enough to have the intended party come to the phone, or the level of everyday conversation, or at a level that includes specialized terminology. The purpose of this function is so that the system may be used most efficiently for a given purpose.

(4) Display of translation result

To enable confirmation as to whether or not the translation has been correctly made, this function will display the translation result in written form.

(5) Editing functions

The resulting translation can be stored temporarily, allowing the speaker to choose from among different suggested translations, add new words to the system's vocabulary, or correct errors before sending. These functions will be particularly useful in the case of voice mail.

(6) Addition of visual information

Through the use of video telephone or other such multi-media services, communication between speakers can be enhanced by means of facial expression, gestures and the like.

(7) Resulting translation function

This function will keep the speakers informed as to how their conversation is being translated and to ensure that the dialog proceeds smoothly. In addition, by transmitting the speaker's actual voice along with the synthetic speech of the "interpreter," it will be possible to get across more of the nuance and feeling of the original utterances.

(8) Translation confirmation accuracy function

As another way of letting users know whether or not their ideas are being translated accurately, this function will re-translate the result back into the speaker's language.

3. Research Task

As noted earlier, realization of an automatic telephone interpretation system will basically require establishment of the three foundational technologies of speech recognition, machine translation, and speech synthesis. In each of these areas, technological development has already progressed to a certain level. In each case, however, this is far from the level of advancement needed to realize an viable system. Thus, extensive fundamental research is still necessary for each of these fields. Here an outline of the main research tasks that must be taken up in each of the constituent technologies is presented.

3.1 Speech Recognition

Of the constituent technologies that will make up the automatic telephone interpretation system, the area with the most problems to be solved is no doubt that of speech recognition. At the present time, progress toward practical realization is being made for speaker-dependent speech recognition systems, where the speaker records his or her voice in the system in advance. If the utterances are pronounced separately, in discrete words, such systems are capable of recognizing anywhere from 500 to 5,000 words and the recognition rate is around 98-99 percent,.whereas for natural, continuous speech the number is a little more than 100 words.

Speaker-independent recognition, on the other hand, which can apply to any speaker, presents even more difficulties. A system now in use for handling telephone inquiries about bank account balances can recognize a few dozen words spoken discretely. At the experimental level, recognition of a few hundred words can be made with about 95 percent accuracy. These systems, however, require an effort by the speaker to conform to the machine's limitations; and recognition rates are much lower for some speakers than for others.

The goal for the study of speech recognition is to establish a technology for the recognition and understanding of speaker independent continuous speech which can recognize the content of a telephone conversation irrespective of the speaker using a large vocabulary.

(1) Recognition of continuous speech

The speech signals uttered by human beings have certain peculiar characteristics. For one thing, utterances are constrained by the speaker's language. This means, for example, that words which are not lexical items of that language will not appear. Moreover, extremely difficult sounds are prevented from occurring by limitations on movement of the tongue and other speech organs. Since conversational speech normally is continuous, with most words running together without breaks, the first problem of speech recognition is to discern the segmentations between words and phrases. This is especially true in Japanese, where stem boundaries, verb endings and adverb endings are not explicit and human are not aware of the definitions of postpositional particles or auxiliary verbs in conversational utterance. Continuous speech in phrase should be recognized in Japanese.

From a linguistic viewpoint, speech can be broken down into minimal units known as "phonemes." The Japanese language (as spoken in the Tokyo area) consists of just over 20 phonemes. For recognition of continuous speech, first the phonemes which are the basic units of that speech must be discerned, then on this basis words and phrases can be recognized. One problem here is that, depending on the speed at which an utterance is spoken, and other circumstances, several variations may appear in the speech signals of identical phonemes.

(2) Large vocabulary speech recognition

The automatic telephone interpretation system will eventually be required to recognize large vocabularies, ranging from a few thousand to several tens of thousands of words. The study, on initiation, will be conducted with a medium-size vocabulary defined as approximately 1000 words continuously spoken. The ultimate aim of the study is a technology that will recognize continuous speech containing a large vocabulary of approximately 3000 words with several hundred specific task words. Languages will be Japanese and English.

As the vocabulary grows larger, the problems of speech recognition are compounded enormously. Besides the problem of variations in speech signals of identical phonemes, a larger vocabulary increases the number of similarsounding words that must be distinguished (like "fender" and "vendor"). The processing involved becomes accordingly much more complex.

The approach taken for this task is first to group the input words roughly on the basis of such factors as types of phonemes and accent pattern. Next, the candidates are narrowed down by means of connection probability using the words preceding and following, and grammatical constraints. More minute differences are investigated to come up with a final result.

In the case of continuous speech, however, in which the segmentations among words or between words and affixes are not clear-cut, and especially in Japanese, this process still may leave a number of possible candidates. Here use must be made of knowledge regarding the topic of the conversation and the context, in order to come to a final understanding of the speech content.

(3) Speaker-independent speech recognition

One of the most problematic aspects of speech recognition is that speech characteristics differ from one speaker to another. If the interpretation system is to be made available for general use, by both men and women, young and old, with all sorts of speech habits, speaker-independent speech recognition technology will have to be developed.

The most effective approach to this problem will be the incorporation of a system for speaker adaptation. The role of this system is to recognize the peculiar characteristics in the speech of a given speaker, so that the system may make the necessary adaptations. Asking each speaker to utter in advance the entire vocabulary of the conversation to be spoken would of course be impractical, since it would impose too great a burden on the user. Accordingly, research is needed into algorithms that will enable speaker adaptation based on a small amount of speech data, along with information already contained in the system as to various possible speech variations.

(4) Understanding of conversational speech

Conversational speech has a structure which differs greatly from single utterances or recitation. Therefore, it will be necessary to study the characteristics of utterance and grammatical structure peculiar to conversation. Moreover, the treatment of meaningless words that are injected during conversation, such throat-clearing, interjections and other sounds will also be necessary.

A technology which uses information about languages and background knowledge to understand speech content is necessary. Knowledge processing or language processing will also be approached. Conversational speech recognition on some domain will be considered. Studies of syntactic, semantics and topical analysis, and creation of a model for speech understanding will be developed. Word spotting will be used to recognize speech in high performance. As for the treatment of meaningless words, several kinds of meaningless words will be deleted.

(5) Influence of interference

Telephone speech is mixed with surrounding noises and telephone circuitry interference. A practical system will have to include functions for eliminating such sounds as office noise from the recognition processing. Moreover, there are a variety of telephone transmitter and circuit noises that commonly mix with voice inputs. Establishment of a technology for speech recognition which eliminates these influences is vital.

To overcome such interference, a technology to detect speech intervals which can distinguish between interference and speech, and that can eliminate interference from sound is to be studied. An interference control technology which can distinguish speech and interference and can normalize these sounds according to the special qualities of each noise characteristic and transmission type will be constructed. Bandwidth limitations on communication channels (3.4 kHz or less) are also a factor here, but by the time the interpretation system reaches the implementation stage, progress will no doubt have been made in the incorporation of intelligent functions in terminal equipment. Also a highly directional microphone should be incorporated. Along with these advances and the digital techniques of terminal equipments should come an easing of bandwidth limitations.

3.2 Machine Translation

The goal of machine translation research in this project is to establish a machine translation technology for Japanese/English conversation.

Systems presently available for translating written texts from one language to another are applicable only to certain limited fields. Moreover, they require human intervention both for pre-editing the text to be translated, and for post-editing the resulting translation. In other words, these are not yet complete translation systems, but in most cases are little more than translation support systems, by which a computer assists human beings in the translation process. Further development of semantic processing techniques, and of the ability to consider surrounding context, is required in order to raise the quality of machine translation to an acceptable level. Qualitatively improving the language processing in machine translation to make it applicable to the automatic telephone interpreting system will involve solving the following types of problems.

(1) Machine translation of conversational sentences

The main object of translation by the automatic telephone interpreting system will be conversational sentences. Spoken language differs from ordinary written language in both vocabulary and grammar. It contains many characteristic linguistic phenomena that are not found in written texts. Analysis of a number of conversations in Japanese reveals the following such peculiarities.

(1) Frequent appearance of imperative and interrogative statements.
(2) Frequent omission of the subject and part of the predicate.
(3) Frequent use of polite, respectful language.
(4) Use of a variety of different expressions and synonyms.
(5) Frequent occurrence of proper nouns (undefined words).
(6) Use of periphrastic expressions.
(7) Variations in word order.
(8) Frequent use of direct expression.
(9) Frequent ambiguities.

In addition, the exchange of words in a conversation takes place on the basis of context and a certain amount of knowledge shared by the participants. This dependence is especially frequent in conversational sentences. Thus analytical methods that have been successful to a degree in written language thus often prove inadequate when applied to conversation.

Given the peculiar forms of expression used in conversational sentences, a translation procedure and syntactical rules particularly applicable to conversation need to be established. When the sentences are input as speech, analysis in the language processing section is made more difficult by the problem of homonyms; and ambiguity of sentences is also greater than for written sentences. Overcoming the various ambiguities and interpreting meaning will require a procedure for making use of a knowledge base, based on semantic analysis, contextual analysis, and representation of knowledge. To this end, a way must be found to construct flexible knowledge networks consisting of knowledge related to syntax and meaning, as well as other relevant information. Making use of these knowledge networks will require some sort of inference structure, for which the syntactic analysis and other processing results act as triggers.

In order to simplify the problems of translating conversational speech and correspondence, a framework has to be set that will focus the speech intention, or "task," along with the speech circumstances. An essential part of a telephone conversation or correspondence is the common exchange that takes place at the beginning, for example, to confirm the other party. Next, ideas are exchanged regarding the topic of the communication. If a way could be found to limit this topic within a certain known framework, the usual difficulty of understanding a conversation can be largely avoided. Creating an effective procedure for processing sentences in a dialog requires that a discourse structure model be devised , by which it will be possible to determine the relationship the purpose of the dialog has to recognition of focus, theme, subject matter and other such aspects.

(2) Machine translation of written correspondence

Considering the growing use of electronic mail and other text telecommunications, the necessity of a means for translating such correspondence texts is evident. Further, the foreseeable trends in development of new telecommunications media would seem to indicate that the automatic telephone interpretation system should not be limited to speech alone. The most efficient approach would be to develop a system that is applicable also to mixed media containing text and image information. In such a case, the system would be required to translate correspondence in the form of written texts as well.

The style of such correspondence texts differs from that of texts in other fields such as technical documents or articles: They contain expressions that are not ordinarily found in written language. For this reason, research will need to be carried out with regard to sentence patterns and contextual matters characteristic of correspondence, so that a translation method may be developed for application to document telecommunications.

This research will have to clarify the special nature of correspondence texts, and devise rules for them by discerning the types of sentence patterns actually used and then classifying them.

(3) Techniques for constructing lexical databases and knowledge bases

A lexical database containing the terminology occurring in conversational sentences and in correspondence texts, as well as the idiomatic expressions that appear frequently in conversation, is a requirement for realizing a high-level machine translation system. Also essential is a large-scale lexical database that includes grammatical rules. These databases are necessary for the syntactic analysis used to process the grammatical structures of a language, and for semantic analysis. The information needed for each type of analysis should be included under one entry, in a unified manner. Since it is vital that a translation correctly grasp the relation of a sentence to what precedes and follows, contextual information has to be extracted and made use of in the translation. A high-level translation is made possible by combined use of these approaches.

The structure should also be flexible enough to permit ready use throughout the world. This requirement is based on the future outlook for development of worldwide systems that translate telecommunications among many different languages

It was noted earlier that conversational sentences and written correspondence feature many words with more than one meaning, and sentences that can be interpreted in more than one way. This is called ambiguity. Thus the accuracy of translation can be improved only by achieving a deeper understanding of the meaning of words. Assuming the domain or task of a conversation or correspondence is specified, then possessing the specialized knowledge of that domain is obviously a help in understanding the content. Technology must therefore be established for grasping and describing that knowledge conceptually, along with the mutual relationship among concepts. Then the concepts expressed by words, and the explanations of those concepts, will need to be organized systematically into a knowledge base, which can aid greatly in the process of understandingmeaning.

3.3 Speech Synthesis

The speech output produced at the end of the automatic telephone interpretation process is based on techniques for converting texts freely to synthetic speech signals. Systems have already been realized which are able to convert text data from texts to speech, in English, French, Spanish, Japanese and other languages. The quality of the synthesized speech is still low, however, lacking in both clarity and naturalness. Below are some of the main problems that remain to be solved.

(1) Highly natural synthetic speech

Techniques for achieving high quality speech synthesis marked by clarity and naturalness should make use of compound speech units of various lengths. A high level rule-based speech synthesis should be developed, which is applicable to Japanese, English and other languages. Rules will have to be established regarding the deformations and connections between speech synthesis units when they are combined to form the words and phrases that make up continuous utterances.

Another matter for research is tone, which relates closely to meaning and naturalness. When the resulting translation is converted to speech, that speech should have a "tone of voice" in keeping with the meaning of the utterance. This involves finding ways to control such factors as accent, stress, and speed on the basis of meaning, and will require further research into the incorporation of conceptual information in speech synthesis.

(2) Individuality of voice quality

The synthesized speech should resemble the voice of the original speaker to the extent possible. As a minimum, this means offering a choice of male and female voices. It should also be possible to synthesize different speech qualities for younger and older people, for example. Here the required techniques include extraction of the factors involved in speech quality information, and methods for controlling speech quality.

4. Research Approach

Research approach of the project at ATR Automatic Interpretive Telephony Laboratories will be described.

4.1 Speech Recognition

The basic plans for research on speech recognition in the telephone interpretation system will now be discussed.

(1) Phoneme recognition

The two main approaches to speech recognition are methods employing a features base and those using pattern matching. Research will begin with an investigation of features base methods for speech recognition. One example is an approach based on spectrogram readings. These will be compared with pattern matching techniques. Consideration will also be given to combining the two methods.

Progress has already been made in establishing techniques for speech recognition by means of pattern matching.

The use of digital signal processor, vector quantization, and other techniques has made realtime processing with pattern matching possible. Because of the differences in spectra, articulation, and other factors for each speaker, however, it has proved difficult to develop application to larger vocabularies or continuous speech recognition in the case of unspecified speakers.

In most of the approaches using a features base, including those based on spectrogram readings, the basic recognition unit is the phoneme. From the standpoint of system configuration techniques, use of phoneme units is well suited to development toward large-vocabulary and continuous speech recognition. It also makes it easier to take advantage of knowledge in the area of speech linguistics.

(2) Continuous speech recognition

Research of continuous speech recognition will proceed with phoneme recognition considered as the basic recognition unit.

All three of these units, phonemes, syllables, and words, are conceivable as basic recognition units. The advantage of using phonemes is that this facilitates handling a dictionary containing a large vocabulary. It is also well suited to a spectral reading approach. However, CV(Consonant Vowel) and VCV syllables will be used for phoneme chains having strong co-articulation. Taking phonemes as the basic unit, then, large amounts of speech data will be investigated in an effort to clarify allophones, phonotactic rules, and phonological rules. Furthermore, word units will be used in the case of high-frequency words.

(3) Word spotting

Algorithms, based on word spotting with phonemes as the basic unit, will be devised for recognition of continuous speech.

In continuous speech recognition, an algorithm employing two-stage a DP(Dynamic Programming) is effective. It recognizes the words in order from left to right, as long as there are no unknown words and the vocabulary is limited to around 100 words. The amount of computation increases drastically, however, as the vocabulary is expanded. A possible replacement for this is an algorithm based on word spotting . Research will be carried out both on word spotting by means of spectral matching and on word spotting using a features base. Further, considering the detected words as "actors," this approach will be developed into a technique for finding the words that immediately proceed and follow the actor.

(4) Word prediction

Dependency relations based on case grammar will be rewritten as an algorithm enabling word prediction. Use of statistical data will also be investigated.

The word spotting module will receive feedback in the form of linguistic information. This information will consist mainly of dependency relations based on case grammar. An algorithm which was developed for natural language processing, based on verbs, will be rewritten as an algorithm which can serve to predict the next word going from left to right. Use of the nearest approximate meaning will also be investigated, with use being made of statistical data such as that of the occurrence probability of trigrams and digrams .

So that this algorithm may be implemented most effectively, formulization will be attempted using dynamic scheduling methods. Besides the word string that appears most accurate, all the word names that were tested by word spotting and their matching scores can be sent to the machine translation module. The word spotting module will also be constructed to enable top-down initiation by the machine translation module.

(5) Prosodic and linguistic processing

An investigation will be made as to the correlation between prosodic and linguistic information, and efforts will be made to enhance the word spotting and other algorithms.

Use will be made of prosodic information such as pitch frequency, stress, and duration, along with information on syllable boundaries, in order to increase the precision and speed of algorithms such as those for phoneme recognition and word spotting. Analysis will also be made of the correlation between prosodic and linguistic information.

(6) Speech signal analysis

The LPC(Linear Predictive Coding) method will be adopted as the main approach to speech analysis. Fundamental studies will also be made, however, of other analytical methods for the purpose of extracting dynamic features.

The LPC analysis has demonstrated a higher level of performance than bandpass filter analysis in recognition of spoken words and in other areas. It will accordingly be the major method used for speech analysis. In the spectrogram reading approach as well, a comparison will be made between FFT spectrograms and LPC spectrograms, so that spectrograms produced by LPC analysis may be employed.

Application of an algorithm from the image processing field will also be made. This will be one method for getting a top-down grasp of the working of formants and other phenomena on the spectrogram. Methods for analyzing the working of formants, and the short-term working of bursts will be studied, centering around processing on the waveform level.

(7) Speaker adaptation

Speaker adaptation algorithms will be researched, taking up the use of vector quantization. Application of this approach to spectrogram formalization will also be attempted.

Without speaker adaptation, it would be quite difficult to achieve adequate speech recognition for unspecified speakers in large-vocabulary or continuous speech situations. Some sort of algorithm must thus be devised for adapting to different speakers. As a means of spectral pattern learning, one promising approach is the use of vector quantization for speaker adaptation. It is also possible to reduce to a certain extent the co-articulation differences among speakers by introducing a multi-frame vector quantization method. Another possibility is the use of a multi-template word dictionary to absorb differences in pronunciation, co-articulation, and the like.

As for speech recognition employing a features base, statistical parameter learning is one approach which is believed to be effective. Initially, an attempt will be made to develop this method of learning with a teacher, where a list of word utterances to be learned must be spoken in advance. Then it will be extended to learning without such a list or a teacher.

(8) Speech database

A speech database is essential if the above researches are to be carried out efficiently. It is necessary to establish a technology for constructing a database to study and evaluate speech recognition and understanding. This database can then be used in the development of wide-scale speech recognition and understanding, and speech synthesis research.

Speech data will be gathered, in the form of important words, continuous short sentences, and phoneme balance lists uttered by announcers and other trained speakers. Then labels, including the name of the phoneme and other important information, will be attached based on the spectrograms. The range of speakers used for compiling the database will gradually be expanded to include a variety of non-professional speakers.

Conversational speech data on some domain topics will be accumulated. Technology to manage a wide scale speech database will also be established. Also a system will be created for management of this speech database with phoneme labels.

A speech database management system will be built to enable the database to be employed most efficiently for research. Consideration will also be given to adding pitch and other prosodic information to the database.

4.2 Machine Translation

The basic research areas in language processing which will be carried out in order to construct the telephone interpretation system will cover the following categories:

- Dialog models
- Mechanism for understanding dialog
- Written language translation techniques
- Lexical database construction methodologies.

Research work in the dialog models will be aimed at a broad and thorough elucidation of the various conditions governing the process whereby a dialog is effected. Rules will be formulated where possible for the different aspects pertaining to speech phenomena.

The aim of this research work will be to probe the computer processing mechanism needed for understanding spoken dialog. This is essential since such a capability is an indispensable prerequisite to machine translation of spoken language.

A major focus of this work will be on certain forms of correspondence text, which in terms of linguistic structure, are regarded as falling between written and spoken languages. The research of written language translation techniques will involve research into techniques for speeding up and enhancing the current level of written language translation techniques so that they can be applied to translation of spoken language.

Research work in the lexical database construction methodologies will focus on different methodologies for constructing a database containing linguistic and knowledge data. These methodologies will be needed at the research stage and during the construction of the translation system.

The specific themes of studies that are to be taken up in each category are briefly explained below.

(1) Dialog structure

Studies will be carried out focusing on dialogs that take place via the telephone. Research will be done on linguistic models that enable dialogs to be effected. This will include examining the conditions and circumstances making a dialog possible as well as the tacitly understood presuppositions and rules allowing a dialog to develop.

To date, few attempts have been made to formulate a model of the dialog structure, and a model has yet to be proposed that would be a worthwhile object of evaluation.

The research areas that will receive special emphasis include: techniques for recognizing structural segments by using the pauses that occur in reading a written text as clues; analysis and systematization of the role played by connectives and connective-like prepositional phrases at the beginning of statements; analysis and systematization of certain styles for indicating the subject; establishment of conversation units; recognition of the focus, theme, and subject, and clarification of their connection with the purpose of the dialog; and the relationship between the segmental structure of a dialog and ellipsis/anaphora.

(2) Rules governing speech dialog

Efforts will be made to formulate rules for the linguistic presuppositions, or implicit conditions, that facilitate the development of a dialog. Rules will be formulated to define the framework for the application and inheritance of these presuppositions. Work will also be directed toward verifying and extending conversational postulates for dialogs in general.

Certain conversational postulates have been proposed, but they have yet to be fully verified. Further work must be done to advance verification of these postulates as they apply to certain specific fields.

Therefore, studies focusing on these specific fields alone will be carried out to systematize and formulate rules for the implicit conditions underlying conversation. Investigations will also be undertaken regarding the rules governing the usage of pronouns, demonstrative pronouns, and ellipses.

(3) Schemes for understanding meaning

Investigations will be made into schemes for understanding the meaning of each conversation unit, taking into consideration the conditions surrounding the development of a dialog. The main areas of research will be: how to accurately represent the state of comprehension on the speakers' part, which changes according to the surrounding circumstances, how to predict what will be said next, and how to manage changes in the topics and statements of the speakers.

Present systems might be able to produce an appropriate sentence on the basis of a correct understanding of each individual conversation unit making up an utterance. Still, they are unable to judge whether the import of the generated sentence is suitable to the flow of the dialog. Further, they are incapable of processing anaphora, ellipses, and the inheritance of linguistic premises.

It is essential that the telephone interpretation system be able to comprehend the meaning in context. Toward this end, studies will be done concerning the introduction of semantic constraints for the semantic markers that occur in the case structure as the units being processed expand from nouns to phrases to sentences. Investigations will also be carried out regarding the introduction of a flexible script or frames for managing the development of a dialog.

(4) Knowledge representation schemes

Studies will be carried out to find a suitable scheme for representing linguistic knowledge as well as common sense outside language and specialized knowledge peculiar to specific domains. The results obtained will be used to establish a flexible knowledge representation scheme suitable for natural language processing. In addition, a knowledge representation language will be developed to describe the representation scheme.

Up to now, the two main schemes used for representing knowledge or the state of comprehension have been predicate logic and semantic networks. With predicate logic, it is possible to specify the operations by which it is executed. This is difficult to do with semantic networks since they do not have a structural configuration. Nonetheless, semantic networks allow tremendous representation flexibility and they offer greater representation capacity.

Research will be carried out in an effort to find a representation method that will allow a semantic network to be given a structural configuration. Each node of the network will be designed so that its contents can be varied to match changes in the surrounding circumstances. To make this possible, studies will be done on node representation that will facilitate knowledge inheritance and dissemination. ,Further, a method will be developed for establishing nodes as objects.

(5) Dialog processing system

Studies will be carried out to find a processing mechanism that can understand conversation units as a dialog unfolds. Research work will focus primarily on the areas of: identifying anaphora, confirming the contents of ellipses, clarifying the process by which linguistic presuppositions are inherited and transmitted, and predicting changes in circumstances using associational knowledge.

A typical example of anaphora is the question of establishing the relationship between a pronoun and its antecedent. Understanding how dialogs develop is an essential prerequisite for improving the accuracy level of the anaphoric inferring problems. A method for implementing the mechanism for understanding how dialogs unfold will also be established.

Studies will be made of each conversation unit case structure to clarify the contents of ellipses. A method of using frames corresponding to a highly flexible script will be investigated as a way of dealing with the intervals between conversation units.

(6) Inference mechanism designed for understanding dialog

The process of understanding a dialog involves an ongoing construction of the state of comprehension. To do this, it is necessary to generate and modify the surrounding circumstances through the use of both knowledge contained within the language and knowledge existing outside of it.

Studies will be done on methods of applying syllogisms and other rules of inference, high-speed unification mechanisms, and ways of using associational knowledge.

(7) Mechanism for high-speed analysis of correspondence text

Various studies will be carried out to develop an analytical mechanism that can be augmented to handle the task of understanding dialogs.

Investigations will be carried out concerning the introduction of templates for the units making up a sentence. Templates will be developed for the sentences, clauses, and phrases peculiar to correspondence text - within certain limited topic areas. Such templates will be needed in order to analyze how a conversation is intentionally developed within certain limited domains. As to the conditions for making syntactic selections, syntactic elements will be established that allow easy treatment of conversation units. This will be accomplished through expanded interpretations of the system comprising the various parts of speech.

Conditions facilitating syntactic and semantic selections will be added to each element to enable high-speed analysis to be performed. Effective use will be made of the results obtained from studies on semantic markers. A mechanism for judging when these conditions should be applied will also be researched. This mechanism will determine the suitability of the selection conditions by using combination rules based on the operators involved.

(8) System for correcting erroneous input

A system will be developed that will automatically correct errors occurring in input words, phrases, clauses or sentences, by making use of syntactic and semantic knowledge. Studies will also be carried out on an error correction system that employs contextual information. The aim of this work will be to enhance the error correction system or to select the most appropriate candidate so that it can comprehend spoken dialog.

It is essential to establish such an automatic error correction system because the strings of symbols input into the language processing system are not always linguistically correct. In order to integrate the language processing and speech recognition systems, it will be necessary to have a system that can use syntactic and semantic knowledge to correct errors in the input sentences. The knowledge that is to be used will range from qualitative linguistic rules to statistical data on the probability of word combinations.

(9) System for representing linguistic structures

A research support environment will be developed which will allow the language analysis and understanding processes to be indicated in a suitable form. Studies will be carried out to establish system configuration techniques that will facilitate the integration of dissimilar processes.

A feedback loop will be established to facilitate smooth circulation of the experimental results. Such a loop will be needed to promote efficient development of the overall system, including the implementation of tests and experiments for verifying all the component technologies.

Every effort will be made to enhance the efficiency of the development work for the various technologies essential to the system. In addition, interfaces will be provided that will allow these technologies to be linked together easily. In this way, the results obtained in each area of research will be integrated into the construction of the system.

(10) Constructing a knowledge base

The knowledge base required for understanding dialog must be able to handle a wide range of dissimilar knowledge, both linguistic and non-linguistic. Methods will be studied for describing non-linguistic knowledge, knowledge involved in making associations, and that used for predicting the development of events. Techniques will then be determined for configuring a knowledge network, for designing nodes made up of frame expressions, and the like.

Knowledge network nodes will constitute stereotypes for realizing a multiple viewpoint, with a hierarchical structure adopted for each node. Methods will also be investigated for achieving a compact knowledge base structure that enables knowledge inheritance to be described. Further, in the area of associational knowledge, a network structure will be considered in which the degree of association varies according to the particular situation. A knowledge base will be compiled within a specific domain, and will be tested in the system for correspondence text translation and the prototype telephone interpretation system.

(11) Research on constructing dictionaries of meanings

Methods will be investigated for constructing dictionaries for use in language processing. A lexical database will then be compiled that will include the necessary morphemic, syntactic, and semantic analysis information for each word or other syntactic unit.

In the translation dictionary, the semantic units must be expanded beyond words to phrases, clauses, and sentences. This will require a new notional structure for expressing the conceptual meaning of expressions and sentences. Research will accordingly be carried out regarding the detailing of primitive actions such as the conceptual classification of declinable words, and structured rules for the generation of these on lattice expressions. The lexical database will also contain bilingual information, mainly on equivalent expressions in Japanese and English.

(12) Research on dictionary structure

Research will be made into dictionary configurations and access mechanisms. The aim is to make it possible to bring the various types of linguistic and non-linguistic knowledge into effective relation with each other, as well as enabling rapid access to the required knowledge and information.

The research will focus on methods for implementing the results of investigations on compiling knowledge bases and dictionaries of meaning. This will involve specifically such matters as file management for high-speed access, techniques for creating hashing tables which are optimized for the situation, and multiple access methods as well as ways of managing them.

4.3. Speech Synthesis

The research effort in the speech synthesis theme will be directed to the following items.

(1) Suitable speech units

Studies will be carried out to find the optimum method for selecting the complex speech units to be used in speech synthesis. These speech units have been proposed largely from the standpoint of linguistic units. There is no assurance, however, that they would produce the best speech quality when judged in terms of acoustic and perceptual distortion. Moreover the techniques for deriving these speech units from actual speech have not yet been fully studied.

In this research, the question of selecting and generating the most suitable speech units will be addressed statistically, using different yardsticks to evaluate various physical characteristics such as spectral distortion. Instead of seeking a one-to-one correspondence between linguistic and speech units, which has been the conventional approach, the aim of this work will be to establish a speech synthesis method using complex speech units of varying length. In addition, throughout the course of this research, efforts will be made to automate speech unit file generation, a task that has traditionally been performed by technicians, and to expand the scope of application to include multi-lingual speech synthesis.

(2) Speech unit transformation and combination rules

Research will be carried out to establish the rules governing the transformation and combination of speech units. Such rules are necessary in generating continuous speech, such as that represented by words and sentences, on the basis of the speech units.

There has been little quantitative analysis of the phenomena involved in co-articulation, and adequate rules have yet to be established which could be incorporated into a speech synthesis system. Since there are a large number of factors contributing to the phenomena which are collectively referred to as co-articulation, it is difficult to treat them independently. Consequently, elucidation of these phenomena will require models that are suitably designed for analysis, as well as an adequate speech database.

The research work will begin by examining the methods for transforming and combining speech units. At the same time, efforts will be directed toward the construction of a large-scale database for speech synthesis. A major area of concern throughout the course of this research will be to clarify acoustic parameter changes based on the use of speech generation models. Those results will then be used to formulate transformation and combination rules. The speech generation models that are obtained through this research will then be applied to analyses of co-articulation phenomena. That work will then lead to the formulation of rules for controlling the speech spectrum.

(3) Speech synthesis employing intention of utterance

The interrelationships among the elements involved in the creation of prosody in speech will be analyzed and a model will be proposed for controlling all of them on an integrated basis. What is needed is to develop a control procedure that more faithfully reproduces the control mechanism employed by humans. Speech synthesis using intention of utterance will be studied by using modal and contextual information.

Analyses will be made of the interrelationships among the three elements involved in prosody, and an integrated control model will be proposed. In addition, a mechanism will be developed whereby the system parameters of the model can be optimized automatically. Work will be directed toward achieving an automatic acquisition process for the system parameters of the model. The scope of that effort will then be expanded to include development of an automatic acquisition algorithm for speech generation.

(4) Control over speech characteristics of individual speakers

The factors governing the speech characteristics of individual speakers will be analyzed from the standpoint of formulating rules for use in speech synthesis. Rules will then be designed to impart individualized characteristics to synthesized speech. The analytical findings will be further examined in auditory perception tests and those results will be used in establishing an appropriate procedure for controlling the speech characteristics of individual speakers.

5 Anticipated Benefit of This Research

Before an automatic telephone interpretation system becomes a practical reality, considerable basic research will be needed to overcome each formidable problem. This research effort will have results that go beyond the field of telephone interpretation. Developments in each of the element technologies that are part of the project will undoubtedly be applicable to a wide range of other fields as well.

Speech recognition techniques, for example, can be applied to a whole range of speech-activated functions, such as dialing by speech, information retrieval by speech, and automatic dictation. The machine translating function has uses also in text communications, lexical database creation, and automatic text proofreading, to name a few. Speech synthesis research will have obvious benefits for development of high quality audio response units, audio guidance units, and automatic text reading equipment for the blind.

Initially the automatic telephone interpretation system will have a limited usable vocabulary and will be restricted in its applications. To enable the most effective utilization, the user will have to register in advance the information relating to the particular field of use. Suppose, for example, that a given user were to employ the system only for making international hotel reservations. This would still solve an important language problem for someone who was not bilingual but was required to make such reservations frequently.

When a full-fledged telephone interpretation system finally becomes a reality, it will mean that the language barrier, which remains as a persistent obstacle to international telecommunications, will have largely been overcome. People will be able to communicate in their own language with others throughout the world, easing much of the difficulty of foreign contact. This should go a long way toward improving international exchange and mutual understanding among nations. In that sense it can be seen as a contribution toward world peace. Certainly for Japan, this project is looked upon as a way of getting beyond the feeling of isolation and "island mentality" imposed more than anything else by the language problem.

At the same time, development of an automatic telephone interpretation system can be expected to result in new approaches being devised for use of the telephone in international telecommunications. These would allow full use to be made of developments in related technologies.

The first of these is the development of multi-media translation communications, combining speech, text, and image. The question of how translation, speech recognition, and other processing is to be divided between the switching station and terminal equipment will have to await further consideration of the system configuration. It would seem, however, that limiting the terminal to sending and receiving of audio signals only, as with conventional telephone service, would not be taking full advantage of the system's potential. We are already seeing advances in terminal equipment technology, as typified by personal computers that incorporate telephone functions. The quality of service offered by the interpreting system would no doubt be improved by making appropriate use of multi-media terminal equipment capable of handling text and image media in addition to speech.

The second aspect is cultivating a diverse range of telephone interpretation techniques. Mailbox services, for example, are a way of dealing with the time differences among countries in international telecommunications. If a translation function were to be added to voice mail service, the resulting translation could also be checked in advance; moreover, the voice of the caller could be transmitted along with the synthesized speech output.

Communication between people in different cultures, of course, involves more than mere language differences. Many problems arise also from cultural differences, such as the differences in customs and in ways of thinking. Ideally, the telephone interpretation system should memorize cultural knowledge for various countries, and be able to refer to this so as to add a more human element to the communications.

7. Conclusion

Research topics toward realization of a system for automatic interpretation of telephone conversations have been presented in the three areas of speech recognition, machine translation, and speech synthesis. An attempt will be made to raise the level of technology in each of these areas qualitatively, based on enormous volumes of linguistic data. Inasmuch as this research is dealing with natural language, it will necessitate developing techniques for constructing massive databases of speech and language information, as well as system configuration techniques applicable to a telecommunications network that is equipped with an interpretation system. Moreover, because of the diversity of parallel processing required, techniques for creating software systems will need to be investigated. From the standpoint of users, research in the communications science field is also required for solving the human problems involved in dialog, and for clarifying language phenomena from the view point of human communications.

Research on automatic telephone interpretation has the primary aim of overcoming the language barrier to international telecommunications. Because of the vast complexity of natural language, however, this goal can only be reached by starting on a small scale and gradually building up knowledge and techniques. It is important, in other words, to proceed with a clear view of what is and is not technically feasible.

One more important requirement of this ambitious project is the mutual cooperation of research institutions in Japan and abroad.