公開

TR-I-0184

Overview of ATR Basic Researches into Telephone Interpretation

ATRにおける自動翻訳電話の概要

Akira Kurematsu, Kiyohiro Shikano*,
Takeshi Kawabata**, Hitoshi Iida, Tsuyoshi Morimoto

榑松明、鹿野清宏*、川端豪**、飯田仁、森元逞

1990.11

内容梗概

本稿はATRにおける自動翻訳電話に関する研究の概要をまとめる。まず、自動翻訳電話の研究についての研究項目と研究概況を述べる。特に、音声と言語の統合処理に重点をおいて以下の項目を述べる。連続音声認識、HMM-LR方式、意図伝達を行う話し言葉翻訳、プラン認識やメモリー主導や並列処理に基づく高度翻訳方式、音声言語翻訳システム(SL-TRANS)について述べる。本資料は、1989年12月にATRで開催された自動翻訳電話基礎研究国際シンポジウム(ASTI)でATRから発表された報告をまとめたものである。

*NTTヒューマン・インタフェース研究所、**NTT基礎研究所

* NTT Human Interface Laboratories, ** NTT Basic Research Laboratories

Contents

1. *An Overview of ATR Basic Research into Telephone Interpretation*
   Akira Kurematsu 1

2. *ATR Approaches to Continuous Speech Recognition*
   Kiyohiro Shikano 15

3. *HMM Continuous Speech Recognition Using Predictive LR Parsing*
   Takeshi Kawabata 28

4. *Intention Translation Method: A Spoken Dialogue Translation System Using a Lexicon-Driven Grammar*
   Hitoshi Iida 40

5. *Advanced Dialogue Translation Techniques: Plan-Based, Memory-Based and Parallel Approaches*
   Hitoshi Iida 53

6. *Spoken Language Processing in SL-TRANS*
   Tsuyoshi Morimoto, Kentaro Ogura, Kenji Kita, Kiyoshi Kogure and Koji Kakigahara 62

Keynote Speech

An Overview of ATR Basic Research into Telephone Interpretation

Akira Kurematsu

ATR Interpreting Telephony Research Laboratories

1. Introduction

The objective of my talk is to give a perspective of research into speech and language processing for telephone interpretation , and to discuss the problems in this area. I shall begin my talk with a general description of requirements for spoken language interpretation. Then I will outline recent ATR research. I will conclude by discussing those areas which require further efforts.

An automatic telephone interpretation system will transform a spoken dialogue from the speaker's language to the listener's automatically and simultaneously. It will undoubtedly be used to overcome language barriers and facilitate communication among the peoples of the world.

Creation of such a system will first require developing the various constituent technologies: speech recognition, machine translation, and speech synthesis. These individual subsystems will then be integrated to form an automatic telephone interpretation system. An automatic telephone interpretation system provides two-way spoken language interpretation between persons speaking different languages. The relationship between speech processing and language processing is quite important.

2. Requirements for Spoken Language Interpretation

Requirements for a spoken language interpretation system are as followings.

(1) High performance recognition and translation must be achieved. There will be no "pre- or post-editing", since both the input and output signals will be speech. This means that the output of a spoken language interpretation system must be highly intelligible so that people speaking two different languages can communicate with each other without much difficulty.

(2) The users are actual participants in the dialogue. The hearer will understand the intention of the speaker upon hearing the interpreted speech. From the point of view of the interaction between human and machine, this system allows

human-to-human communication through machine processing. The active participation of interacting humans will permit a richer capability than a conventional machine translation system.

(3) The kind of sentences the spoken language interpretation system must handle is very different from those processed by a machine translation system for written texts. Although in spoken dialogues sentences are short and sentence structure not particularly complex, spoken dialogues include elliptic and anaphoric expressions. They also tend to include many syntactically ill-formed expressions. In addition to handling the often ill-grammatical nature, a spoken language interpretation system must tackle the errors or ambiguities made by its speech recognition unit which can never be perfect. Tolerating the inevitable recognition errors or ambiguities, a language parsing process followed by speech recognition will be required to efficiently use the parsing algorithms.

(4) One basic requirement is that a spoken language interpretation system must be operated in real time. The system simply cannot take more than several seconds to translate a sentence, as the hearer is right there waiting for the output from the system, and a consecutive reply is required to respond to the previous utterance. What this means is that high speed processing of both speech and language is required. Efficient algorithms are crucial if an exponential increase in computation time is to be avoided. It will be also possible to translate spoken sentences in a batch mode as in a speech mail box. However, such usage will be quite limited in normal telephone interpretation.

(5) Although the ultimate goal of the telephone interpretation system will be universal dialogue in an unlimited domain, the present goal, which is quite feasible, is a system which is limited to specific, task-oriented areas. Domain knowledge is expected to disambiguate the spoken language. Efficient processing by use of prediction in the limited domain is expected.

(6) It will be used by mono-lingual users, that is, the speaker will not know the target language and the hearer will not understand the source language. This situation also imposes stringent requirements on the accuracy of interpretation to prevent misunderstanding. As the automatic telephone interpretation system is a completely new one, the overall system design will be determined by considering the level of performance which can be expected from each of the constituent technologies as well as human nature. The ease of use, or "user friendliness" of the system will not be neglected and the possibility of using a multi-media terminal which deals with image and text in addition to speech is likely to compensate for the awkwardness of the telephone interpretation system.

3. Overview of Current Research

Currently, ATR Interpreting Telephony Research Laboratories are engaged in research in the areas of continuous large-vocabulary Japanese speech recognition, integrated processing of speech and language, machine translation of spoken language from Japanese to English, and high-quality speech synthesis. Here, the key component is the integrated processing of speech and language. It is intended to bridge the gap between speech recognition and language analysis. Based on the idea of the language source model, it carries out top-down prediction to the speech recognizor. The bottom-up results contains multiple candidates which must be narrowed down by use of linguistic constraints and various knowledge information. Additional areas are also being investigated with the collaboration of other research institutes.

(1) Speech Recognition

For recognition of large vocabulary continuous speech, first, phonemes are being recognized to the extent possible, then continuous words and phrases are recognized. Several improvements in HMM phoneme models have been introduced and applied to the recognition of continuous speech. Phoneme segmentation using spectrogram reading knowledge and phoneme recognition by neural network are being researched to enhance the capability of speech recognition.

As an effective approach to the problem of speaker independence, speaker adaptation has been taken. As a means of spectral pattern learning for speaker adaptation, one promising approach is codebook mapping. This algorithm realizes general speaker adaptation which does not depend on speech recognition systems. Twenty to thirty words will be enough to adapt to speaker characteristics.

(2) Integrating Speech and Language Processing

Integration of speech and language processing is extensively researched. For speech recognition, there are problems in the following areas:

- Phrase recognition rate of top five candidates will be around 95%.
- Phrase recognition uncertainty will produce many candidates.
- Increased perplexity is a severe condition for real-time data processing.

Perplexity means the average number of words from which the recognizor has to make a choice at any single point. It now seems that a spoken language system will have to be constrained on the language so that the system is tractable.

Linguistically-based constraints of syntactic, semantic and pragmatic information will have to be utilized to narrow the number of word candidates.

HMM Phoneme models are integrated with the generalized LR parser. In our approach, the next phoneme is predicted in the speech input and verified by estimating the overall probabilities. This integration algorithm is effective for a large vocabulary with high perplexity and is processed quite efficiently. For ease of recognition, we are adopting a method which recognizes continuous speech separated into phrases.

In language processing systems, a function which can use syntactic and semantic knowledge to select the most appropriate candidate will be necessary. A method which reduces the number of candidates from speech recognition is being studied in the analysis of spoken Japanese utterances. In this method, the Japanese dependency relationship is used. Using this information, probable candidates are selected from the speech recognition output.

A knowledge processing approach in the specific domain has been investigated. It has been studied to reduce the number of candidates from speech recognition by applying the knowledge base. The relationships express the associative relationship between words.

(3) Machine Translation

It will be desirable that the telephone interpretation system be able to comprehend meaning in context. Our research has mainly been directed to solving the following problems:

- ●How to make up for omitted words, how to disambiguate expressions or how to be coincident with pronouns,
- ●How to accurately represent the speaker's comprehension, which changes according to circumstances,
- ●How to predict what will be said next,
- ●How to manage changes in the speaker's topics and statements.

A dialogue model as well as a broad explanation of the dialogue process is effective for machine translation. Major linguistic phenomena peculiar to Japanese spoken dialogues have been investigated from a linguistic viewpoint in order to construct a discourse-dialogue model that can be implemented on a computer. Research topics into zero pronouns, honorifics, negation, and intention among other areas is now underway. The results obtained have been integrated, step-by-step, as a grammar for Japanese dialogue analysis.

Spoken Japanese, sentences contain certain inherent ambiguities, especially in the distribution of zero pronouns and construction of predicate phrases. In order to disambiguate them, pragmatic constraints on the uses of expressions must be extracted and the most plausible analysis candidate selected by using these constraints. The current approach to analyzing Japanese dialogues is based on a lexico-syntactic grammar framework in terms of typed feature structures and an analysis order controllable parser.

In devising a system for machine translation of telephone dialogues, one of the problems to be solved is how to adequately translate the underlying meaning of the source utterance, or the speaker's intention, into the target language. In dialogue, smoothness of communication depends on understanding the speaker's underlying meaning. Considerable research has been focused on a plan recognition model for understanding and translating dialogue. A computational model for context processing using constraints on dialogue participants' mental states is also being studied.

The proposed dialogue translation method is essentially based on the semantic transfer approach, and can be characterized by its two translating processes: one which extracts intentions in utterances, such as request, promise, greetings, etc., and another which transfers propositional parts of utterances. A feature structure is adopted as an integrated description of information for the whole process of analysis, transfer and generation. A method to efficiently handle the feature structures is also being studied.

In order to establish robust translation, an example-based paradigm has been studied. This consists of finding examples similar to the current object and adapting the examples to solve new problems. This method is now being applied to disambiguation of the speech recognizor output and the proper translation of connected noun groups.

An experimental spoken language translation system has been developed at ATR determining the major problems inherent in the integration of speech and language processing. It has turned out that an efficient algorithm for spoken language analysis is necessary.

(4) Speech Synthesis

Speech output produced at the end of the machine translation process will be based on speech synthesis by rule, where linguistic information including morphological information will be utilized as well as the text. When the resulting translation is converted to speech, that speech should have a "tone of voice" in keeping with the meaning of the utterance. The main factors in speech clarity and naturalness are the proper selection of the speech synthesis unit and control of the rules of prosody. Speech synthesis using compound speech synthesis units of various lengths are being explored.

Individualization of synthetic speech has been accomplished by use of voice conversion from one speaker's to another's. The required techniques include extraction of the factors involved in speech quality information, and methods for controlling speech quality.

4 Further Research

An extensive effort must be made to raise the level of technology of speech recognition, machine translation, and speech synthesis if automatic interpretation of telephone conversations is to be realized. Further research is now being directed to the following points.

The capability of recognizing large vocabulary continuous speech should be further enhanced. Our goal is 3,000 words. In order to obtain better phoneme recognition performance, a scheme to integrate current approaches of applying relevant knowledge about speech will be formulated.

In speaker-independent recognition, a method applicable to large-vocabulary continuous speech recognition will be explored by using a large scale speech database.

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

In the integration of speech and language processing, a scheme to predict the level of words or utterance will be studied. The introduction of the statistical characteristics of grammar or heuristics of sentence structure will be explored. Statistical constraints on the input which takes the form of estimates of the probability of a particular sequence of words will be used to reduce the perplexity. Further, utilization of higher level information such as dialogue structure will be investigated. The prototype of a spoken language translation system between Japanese and English will be demonstrated in this phase of the present research project.

Knowledge about language itself and domain-specific extralinguistic knowledge will be formulated as the common base for various aspects of spoken language interpretation.

In machine translation, the enrichment of grammar and lexical dictionaries to an advanced level will be carried out to cope with large vocabulary translation. To enhance deep understanding, translations based on context processing will be explored. The challenge will be directed to the general methodologies which can be expanded to large vocabularies and various task domains. Considering the requirement of real-time processing, a high speed computational scheme will be researched to shorten the considerable existing gap between theoretical computational linguistics and software implementation.

In speech synthesis, speech synthesis by rule will be enhanced to obtain more natural speech quality in conversational sentences. The linguistic information in language generation will be reflected to the control of rule in speech synthesis. Voice individualization over different languages will be developed.

Massive databases of speech and language corpora will be essential and necessary to further promote related research. Because of the vast complexity of natural language, however, the goal should be reached by gradually improving levels and techniques. Also, it will be necessary to consider the expandability of the system in terms of domain size, different domains application, multi-language application, and multi-speaker use.

Another important matter to consider in this ambitious project is international cooperation. Natural language, each country's mother tongue, will have to be deeply studied in the research organizations of countries around the world.

An Overview of ATR
Basic Research into Telephone Interpretation

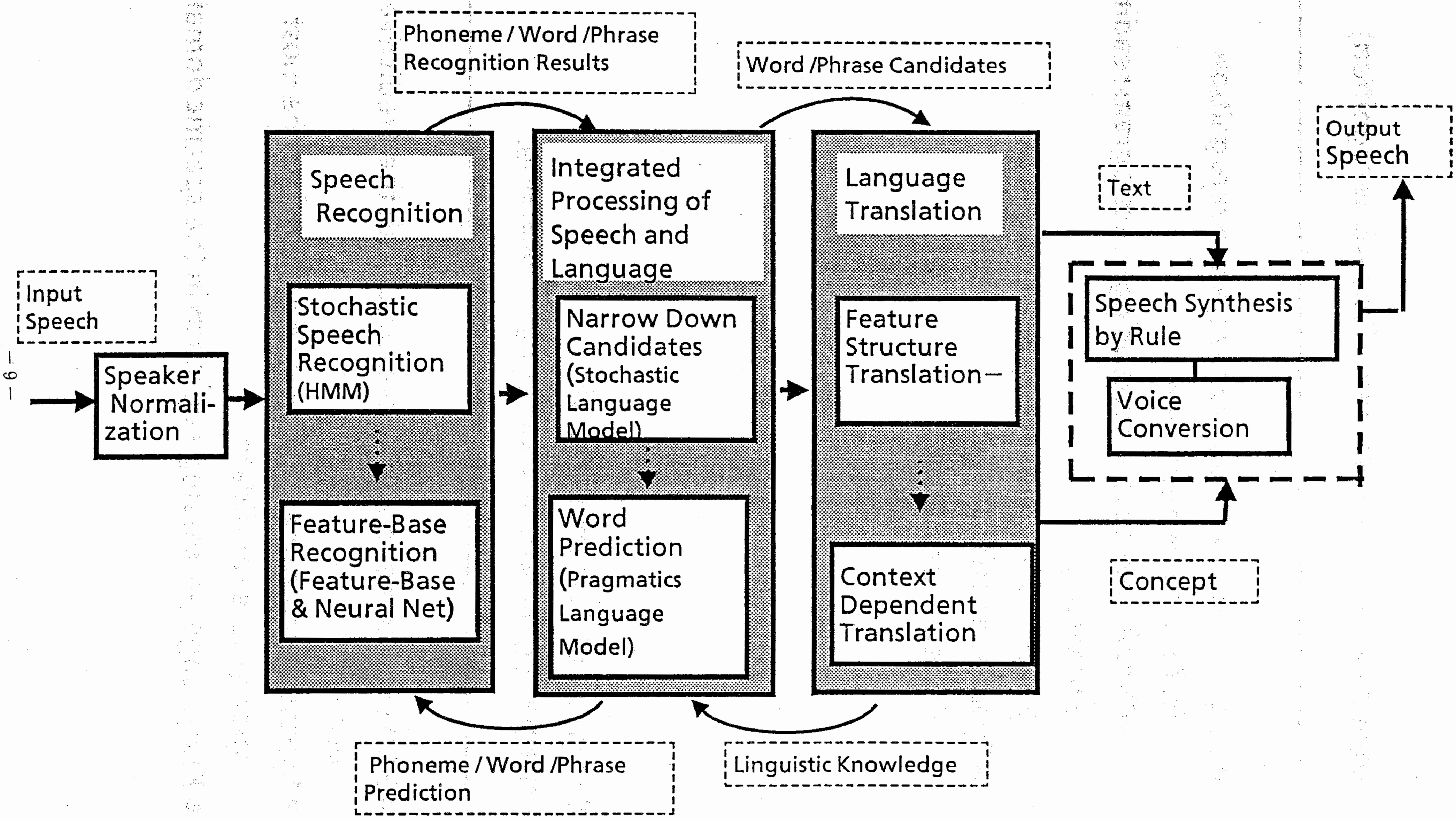
Akira Kurematsu

ATR Interpreting Telephony Research Laboratories

Requirements for Spoken Language Interpretation

- High performance recognition and translation
  - No "pre- or post-editing"
- Actual participants in the dialogue
- Ill-formed expressions
  - Errors or ambiguities
- Real time processing
  - Efficient algorithm
- Specific, task-oriented domains
- User friendliness

Fig. 1 Proposed Telephone Interpretation System


Phoneme / Word /Phrase Recognition Results
Word /Phrase Candidates
Output Speech
Text
Input Speech
Speaker Normalization
Speech Recognition
Stochastic Speech Recognition (HMM)
Feature-Base Recognition (Feature-Base & Neural Net)
Integrated Processing of Speech and Language
Narrow Down Candidates (Stochastic Language Model)
Word Prediction (Pragmatics Language Model)
Language Translation
Feature Structure Translation
Context Dependent Translation
Speech Synthesis by Rule
Voice Conversion
Concept
Phoneme / Word /Phrase Prediction
Linguistic Knowledge

Overview of Current Research (Speech Recognition)

- Large vocabulary continuous speech recognition
  - HMM phoneme models
  - Phoneme segmentation using spectrogram reading knowledge
  - Phoneme recognition by neural network
- Speaker adaptation
  - Codebook mapping

Overview of Current Research
(Integrating Speech and Language Processing)

- HMM Phoneme models integrated with the predictive parser
- Syntactic and semantic knowledge to select the most appropriate candidate
- Knowledge processing approach in the specific domain

Overview of Current Research (Machine Translation (1))

- Discourse-dialogue model
  - Zero pronouns, honorifics, negation, and intention
- Analysis of Spoken Japanese, sentence
  - Lexico-syntactic grammar framework in terms of typed feature structures
- Translate the underlying meaning of the source utterance
  - plan recognition model

Overview of Current Research (Machine Translation (2)

- Dialogue translation method
  - Intentions , such as request, promise, greetings
  - Propositional parts of utterances
- Robust translation
  - Example-based paradigm
- Experimental spoken language translation system
  - Integration of speech and language processing

Overview of Current Research (Speech Synthesis)

- Speech synthesis by rule
  - Speech synthesis using compound speech synthesis units
- Individualization of synthetic speech
  - Voice conversion

Further Research (1) (Speech Recognition)

- Recognition of large vocabulary continuous speech
  - Scheme to integrate current approaches of applying relevant knowledge about speech
- Speaker-independent recognition
  - method applicable to large-vocabulary continuous speech recognition
- Prosodic information

Further Research (2) (Integration of Speech and Language)

- Scheme to predict the level of words or utterance
  - Statistical characteristics of grammar or heuristics of sentence structure
- Prototype of a spoken language translation system between Japanese and English
- Knowledge about language itself and domain-specific extralinguistic knowledge

Further Research (3) (Machine Translation)

- Enrichment of grammar and lexical dictionaries
- Translations based on context processing
- High speed computational scheme

Further Research (4) (Speech Synthesis)

- More natural speech quality in conversational sentences
- Voice individualization over different languages

Further Research (5)

- Massive databases of speech and language corpora
- Expandability of the system
  - Domain size, different domains, multi-language, and multi-speaker use
- International cooperation

ATR Approaches to Continuous Speech Recognition

Kiyohiro Shikano
(ATR Interpreting Telephony Research Laboratories)

1. Introduction

In this article, ATR approaches to continuous speech recognition are overviewed. Our efforts aimed at speaker-dependent phoneme recognition and speaker-independent phoneme segmentation have resulted in dramatically improved phoneme recognition and continuous speech recognition performance. A continuous speech recognition system based on Hidden Markov Modeling attains a phrase recognition rate of 88% for a 1,000 word task.

We are taking four approaches to speech recognition: (1) *A Feature-Based* approach, especially for phoneme segmentation, (2) *A Hidden Markov Model* approach, (3) *A Neural Network* approach, and (4) *A Learning Vector Quantization* approach.

For speaker-independent speech recognitibn, a speaker adaptation approach has been undertaken using the concept of *Fuzzy Vector Quantization and Spectrum Mapping*. We also successfully applied this algorithm to an HMM-based continuous speech recognition system with a phrase recognition rate of 79%. In the area of language models, a language source modeling approach using the *generalized LR* parser and an *N-Gram* approach using Neural Networks have been investigated.

2. Large Vocabulary Continuous Speech Recognition

Reliable phoneme recognition and segmentation algorithms have been investigated leading to considerable improvements over conventional approaches. We have been pursuing three approaches: a *Feature-Based* approach, a *Hidden Markov Model* approach, and a *Neural Network* approach. These improvements resulted in the successful implementation of a continuous speech recognition system such as *HMM-LR* by combining HMM phoneme models with the generalized LR parsing algorithm. Hardware implementation has also begun.

2.1. Feature-Based Approaches

In particular, phoneme boundaries can be identified correctly in Japanese utterances using expert knowledge, which can deal with various kinds of coarticulation phenomena. Now we are developing an expert system(Hatazaki-89-05)

(Komori-89-09) for phoneme segmentation using the expert tool, *ART*. A Symbolics lisp machine for running *ART* and a micro Vax station/Vax 8800 for feature extraction of input speech are connected and used to implement a phoneme segmentation expert system. The phoneme segmentation knowledge for consonants is described by ART. The phoneme segmentation expert is integrated with a *TDNN* neural network for consonant discrimination(Komori-90-05). The system attains a consonant segmentation rate of 94.5% and a consonant recognition rate of 88.8%. Moreover, spotting of vowels, semi-vowels and the syllabic nasal has been studied using a vowel spotting TDNN neural network.

The final aim of this feature-based approach is the implementation of a phonemic typewriter, which is able to recognize continuous speech without language knowledge such as words or syntax.

2.2. Hidden Markov Models

Hidden Markov Models (HMM) were studied to determine how to best use models in the training stage on the forward-backward algorithm. The number of states, tied probabilities of transition and output, probability smoothing techniques, initial probability settings and duration control techniques were studied based on the task of phoneme recognition using our large vocabulary speech database(Kuwabara-89-05). After that the HMM phoneme models were improved by introducing *Separate VQ* (multiple codebooks) and *Fuzzy VQ* techniques. These improvements resulted in a 9% phoneme recognition rate improvement, from 86.5% to 95.7%.

The Hidden Markov models based on phoneme units have been applied successfully to word spotting in continuous speech(Kawabata-89-05). Moreover, the HMM phoneme models are combined with a generalized LR parser (a language source model) to efficiently recognize a Japanese phrase input, where the LR parser can predict phoneme candidates to the HMM phoneme models. This system is called *HMM-LR*(Kita-89-05). HMM-LR can recognize Japanese phrase inputs with a phrase recognition accuracy of 88%(Hanazawa-90-04). The performance correct for the top-two and top-five phrase candidates are 96% and 99%, respectively. Robustness studies for continuous speech recognition and speaker adaptation have been initiated to improve HMM-LR performance. HMM-LR hardware implementation has also been initiated in order to show the feasibility of a prototype interpreting telephony system. The HMM phoneme model approach is also applied to English word recognition.

2.3. Neural Network Approaches

Using Neural Networks, phoneme recognition and continuous speech recognition have been investigated. A time-delay neural network (*TDNN*) now achieves high phoneme recognition rates for the task of speaker-dependent discrimination not only among the voiced consonants, /b/,/d/, and /g/(Waibel-89-03) but also for full consonants(Waibel-89-05). All consonants were extracted from the phonetically labelled large vocabulary database , i.e., from 5,240 common Japanese words spoken by three speakers. The TDNN attains a 98.6% phoneme recognition rate for the /b/,/d/, and /g/ task, and 96.7% for the full consonant task. Phoneme spotting(Sawai-89-05) by TDNN aimed at determining a continuous speech recognition approach by neural networks has been carried out and shows a phoneme spotting rate of 98%(Miyatake-90-04). A preliminary experiment to recognize continuous utterances using TDNN phoneme spotting results has been tried by summing up the phoneme spotting outputs by means of a DTW algorithm.

Efforts to speed up the back-propagation algorithm resulted in a 1,000-fold speedup of the TDNN(Haffner-89-09). This speedup and the use of an Alliant mini-supercomputer make it possible to challenge larger scale neural networks. TDNN robustness and generalization studies for different speaking styles have been initiated.

Other neural network approaches such as a deterministic Boltzmann machine(Dang-89-10) and a neural prediction model are also being studied. A constructive neural network (CNN)(Kawabata-89-10) to recognize words in bottom-up fashion using a TDNN phoneme spotting network is also studied.

2.4 Learning Vector Quantization Approaches

Another neural network approach to continuous speech recognition using Kohonen's Learning Vector Quantization algorithm (*LVQ*) has been investigated at ATR Auditory and Visual Perception Research Laboratories.

In order to deal with dynamic features in speech, a 7-frame(70 msec) input is used to extract phoneme features. Therefore, LVQ can be regarded as a kind of matrix vector quantization with the ability of phoneme discrimination. There are currently two versions of LVQ; attention is here focused on LVQ2(McDermott-89-05). LVQ2 attained a phoneme recognition rate of 97.7% for an 18-consonant task, where the rate by K-means clustering was 91.5%.

LVQ2 is also successfully combined with HMM phoneme models described in section 2.2. The LVQ2 codebook can provide the HMM phonems models with high classification power at the phonemic level. This combined system is called *LVQ-HMM*(Iwamida-90-04). The LVQ-HMM attains a 97.4% phoneme recognition rate for the 18 consonant task. LVQ-HMM is being applied to continuous speech recognition combined with the language source model of the generalized LR described in section 4.1. The fuzzy VQ is also introduced in order to improve continuous speech recognition performance. Japanese phrase recognition experiments are now being carried out.

3. Speaker Adaptation

Aiming at a general preprocessor for speaker normalization, speaker adaptation research using *codebook mapping* techniques has been studied. Discrete spectrum space representation by vector quantization makes it possible to realize sophisticated speaker adaptation / normalization by codebook mapping. In paticular, the fine tuning to HMM has been investigated.

3.1. Speaker Adaptation by Vector Quantization

As a general preprocessor to a phoneme recognizer, speaker normalization algorithms have been developed. The algorithms adopt vector quantization as a discrete representation of spectral space. Discrete representation makes it possible to carry out sophisticated spectral normalization or mapping from one speaker to another. In previous studies, algorithms using a single codebook were developed, and a complex spectral distortion measure which is composed of a spectrum term, a differenced spectrum term and a differenced power term was adopted. The complex spectrum distortion measure improved the word recognition results, but also significantly increased spectrogram distortion. In order to reduce this degradation, algorithms with multiple codebooks (separate VQ) were investigated. *Fuzzy vector quantization* techniques have also been investigated to realize more accurate speaker adaptation. These algorithms are also successfully applied to voice conversion and cross-language voice conversion(M.Abe-90-04). These algorithms are also applied to neural network speech recognition and will be applied to feature-based speech recognition as a speaker normalization preprocessor.

These technologies lead us to an HMM speaker adaptation algorithm(S.Nakamura-89-05) greatly improved over previously reported algorithms. The HMM speaker adaptation algorithm attains a phrase recognition rate of 78.6%, where the phrase recognition rates of speaker independence and speaker dependence are 59.6% and 88.4%, respectively. The correct rates for the top-two and top-five phrase candidates are 92% and 97%, respectively. Moreover, a supplement HMM phoneme model training approach(Hattori-90-04) has been initiated to cope with speaker coarticulation variations.

A comparative study(S.Nakamura-90-04) between nonlinear neural network mapping and codebook mapping is performed, which clarify that codebook mapping is more accurate than neural network mapping.

These speaker adaptation algorithms will shortly be applied to noisy speech recognition, and will be compared to a noise reduction neural network ability(Tamura-89-05)(Tamura-90-04).

4. Language Source Modeling

The ATR Interpreting Telephony Laboratories also include two other departments which are studying language models more deeply than the Speech Processing Department. Nevertheless, the Speech Processing Department has been studying language source models based on bottom-up word/phoneme prediction and statistical word/phoneme prediction in collaboration with the other departments. We have been taking two approaches: an LR source modeling approach and a neural network approach (*Netgram* source modeling), to predict the next word sets or phonemes. Such bottom-up word/phoneme prediction approaches based on language source modeling should be combined with top-down approaches in deeper language processing.

Another approach to making a phoneme sequence source model has been initiated aiming at a phonemic typewriter.

4.1. Language Source Modeling by Generalized LR Parser

The generalized LR parser, called the LR parser for short, was introduced to predict next words/phonemes. The LR parser can be regarded not only as a parser for input word sequences but also a language source model for word/phoneme prediction/generation. The LR parser can deal with a context-free grammar. The LR parser is successfully integrated with HMM phoneme models to recognize a Japanese phrase utterance, where the LR parser very quickly predicts phoneme candidates according to the LR table. The LR parser will be developed for better language source modeling by introducing the probabilities of phoneme sequences and word sequences automatically estimated from the large-scale text database. The combined system of the LR language source model and HMM, called *HMM-LR*(Kita-89-05), attains a phrase recognition rate of 88%(Hanazawa-90-04) for a task with a phoneme perplexity of 5.9. The task includes 1,000 words and is written using a context-free grammar to deal with Japanese phrases(Bunsetsu). The cooccurrence of context-free rewriting rules will be used to produce a better language source model(Kita-90-04).

The LR parser is also successfully combined with the Sphinx system at Carnegie Mellon University(CMU) under the research collaboration between CMU and ATR.

Other low-level linguistic information is contained in phoneme sequences. The use of phoneme sequence information such as a syllable trigram has been studied aiming at source modeling of Japanese speech inputs.

4.2. N-Gram Word Prediction by Neural Networks

A neural network approach, *the NETgram*(M.Nakamura-89-05), has been developed to predict words using the Brown Corpus Text Database. The results have been compared to the results of statistical trigram modeling. The analysis of the internal representation of the NETgram reveals some correspondence to language categories. In order to realize this approach practically and quickly, further improvements to our learning network simulator have been made. The NETgram output values are successfully used to improve word recognition in English sentences. This approach will make the realization of a phonemic typewriter more feasible.

4.3. Phoneme Source Modeling

Aiming at a phonemic typewriter of Japanese continuous speech input, phoneme source modeling has just been initiated using a Japanese text database. First, a syllable trigram is calculated from the Japanese text database. The syllable trigram phoneme source model is then used as a phoneme prediction model, and is combined with the HMM speech recognition algorithm based on phoneme models. This results in a phoneme recognition rate of 90%. More accurate phoneme source modeling is being studied using Japanese text database.

(References)

(Dang-89-10) J-C. Dang, H. Sawai, **Deterministic Boltzmann Machines for Phoneme Recognition**, ASJ Fall meeting, 1-1-23, Toyama,(1989-10)

(Iwamida-90-04) H.Iwamida, E.McDermott, S.Katagiri, Y.Tohkura, **A Hybrid Speech Recognition System Using HMMs with an LVQ-Trained Codebook**, will appear at ICASSP90, (1990-04)

(Haffner-89-09) P. Haffner, H. Sawai, A. Waibel, K. Shikano, **Fast Back-Propagation Learning Methods for Large Phonemic Neural Networks**, European Conference on Speech Communica-tion and Technology, pp553-556, Paris, (1989-09)

(Hanazawa-90-04) T.Hanazawa, K.Kita, T.Kawabata, K.Shikano, **ATR HMM-LR Continuous Speech Recognition System**, will appear at ICASSP90, (1990-04)

(Hatazaki-89-05) K. Hatazaki, Y.Komori, T. Kawabata, K. Shikano, **Phoneme Segmentation Using Spectrogram Reading Knowledge**, ICASSP'89, S.8.2, pp393-396, Glasgow, (1989-05)

(Hattori-90-04) H.Hattori, S.Nakamura, K.Shikano, **Supplementation of Articulatory Speaker Variation for Speaker Adaptation on HMM**, will appear at ICASSP90, (1990-04)

(Kawabata-89-05) T.Kawabata, K. Shikano, **Island-Driven Continuous Speech Recognizer Using Phoneme-Based HMM Word Spotting**, ICASSP'89,S9.7, pp461-464,Glasgow, (1989-05)

(Kawabata-89-10) T. Kawabata, **Constructive Neural Network for Speech Recognition**, ASJ Fall meeting, 1-1-26, Toyama, (1989-10) (in Japanese)

(Kita-89-05) K. Kita, T. Kawabata, H. Saito, **HMM Continuous Speech Recognition Using Predictive LR Parsing**, ICASSP'89, S13.3, pp703-706,Glasgow, (1989-05)

(Kita-90-04) K. Kita, T. Kawabata, **HMM Continuous Speech Recognition Using Stochastic Language Models**, will appear at ICASSP90, (1990-04)

(Komori-89-09) Y. Komori, K. Hatazaki, T. Tanaka, T. Kawabata, K.Shikano, **Phoneme Recognition Expert System Using Spectrogram Reading Knowledge and Neural Networks**, European Conf. on Speech Communication and Technology, pp549-552, Paris, (1989-09)

(Komori-90-04) Y.Komori, K.Hatazaki, T.Tanaka, **Combining Phoneme Identification Neural Networks into an Expert System Using Spectrogram Reading Knowledge**, will appear at ICASSP90, (1990-04)

(Kuwabara-89-05) H. Kuwabara, K. Takeda, Y. Sagisaka, S. Morikawa, T. Watanabe, **Construction of a Large-Scale Japanese Speech Database and its Management System**, ICASSP'89S10b.12, pp560-563, Glasgow, (1989-05)

(M.Abe-89-05) M. Abe, S. Tamura, H. Kuwabara, **A New Speech Modification Method by Signal Reconstruction**, ICASSP'89, S11.9, pp592-595 Glasgow, (1989-05)

(M.Abe-90-04) M.Abe, K.Shikano, H.Kuwabara, **Cross-Language Voice Conversion**, will appear at ICASSP90, (1990-04)

(McDermott-89-05) E.McDermott, S.Katagiri, Shift-Invariant, **Multi-Category Phoneme Recognition using Kohonen's LVQ2**, ICASSP'89, S3.1, pp81-84, Glasgow, (1989-05)

(M.Nakamura-89-05) M. Nakamura, K. Shikano, **English Word Category Prediction Based on Neural Networks**, ICASSP'89, S13.10, pp731-734, Glasgow, (1989-05)

(Miyatake-90-04) M.Miyatake, H.Sawai, Y.Minami, K.Shikano, **Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks**, will appear at ICASSP90, (1990-04)

(S.Nakamura-89-05) S. Nakamura, K. Shikano, **Speaker Adaptation Applied to HMM and Neural Networks**, ICASSP'89, S3.3, pp89-92, Glasgow, (1989-05)

(S.Nakamura-90-04) S.Nakamura, K.Shikano, **A Comparative Study of Spectral Mapping for Speaker Adaptation**, will appear at ICASSP90, (1990-04)

(Sawai-89-05) H. Sawai, A. Waibel, M. Miyatake, K. Shikano, **Spotting Japanese CV-Syllables and Phonemes Using Time-Delay Nerual Networks**, ICASSP'89, S.1.7, pp25-28, Glasgow, (1989-05)

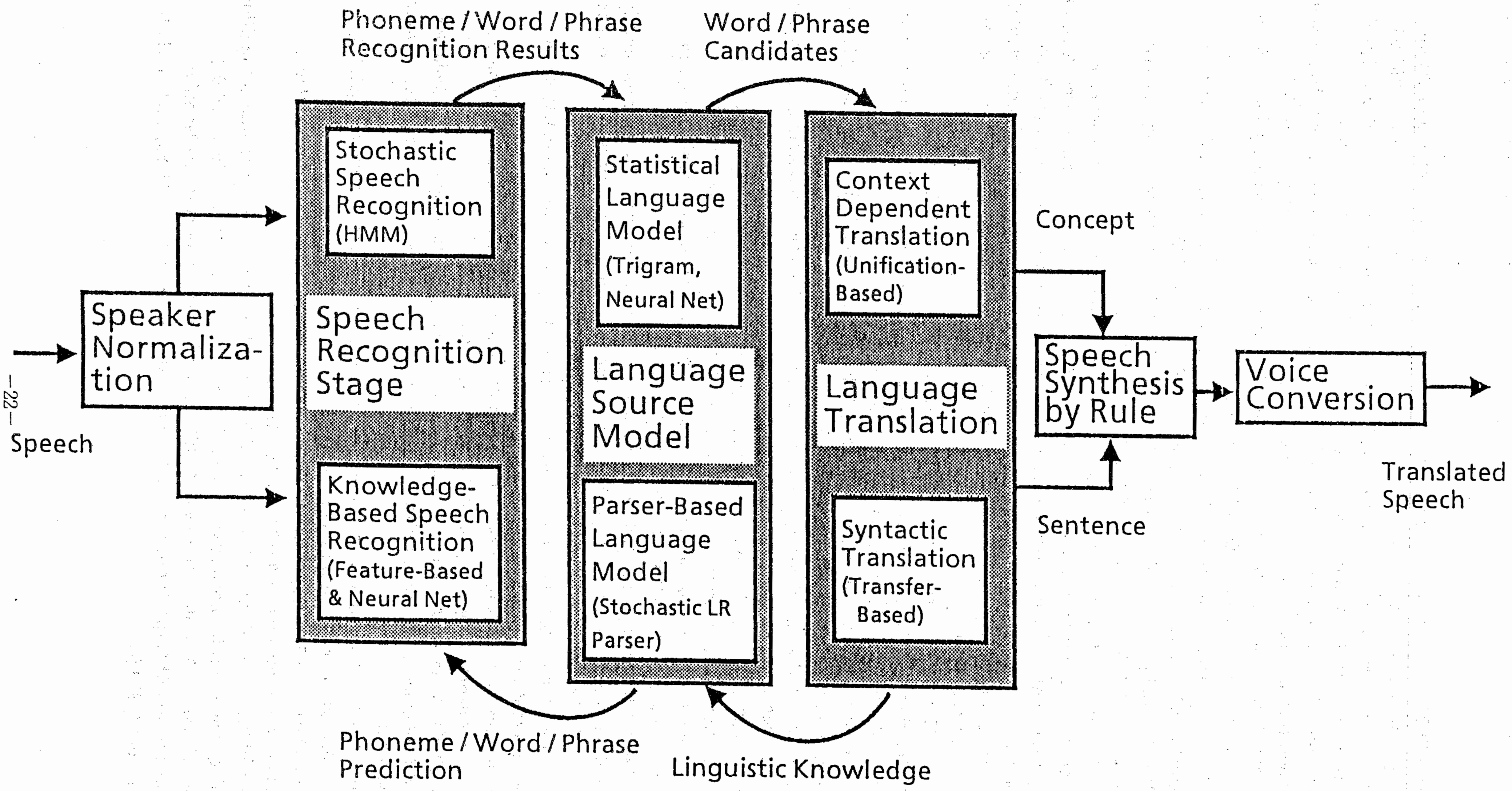
(Tamura-89-05) S.Tamura, **An Analysis of a Noise Reduction Neural Network**, ICASSP'89, A.1a.6, pp2001-2004, Glasgow, (1989-05)

(Tamura-90-04) S.Tamura, M.Nakamura, **Improvements to the Noise Reduction Neural Network**, will appear at ICASSP90, (1990-04)

(Waibel-89-03) A. Waibel, T.Hanazawa, G.Hinton, K.Shikano, K.Lang, **Phoneme Recognition Using Time-Delay Neural Networks**, IEEE Tr.ASSP, Vol.37, No.3, pp328-339, (1989-03)

(Waibel-89-05) A. Waibel, H. Sawai, K. Shikano, **Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks**, ICASSP'89, S.3.9, pp112-115, Glasgow, (1989-05)

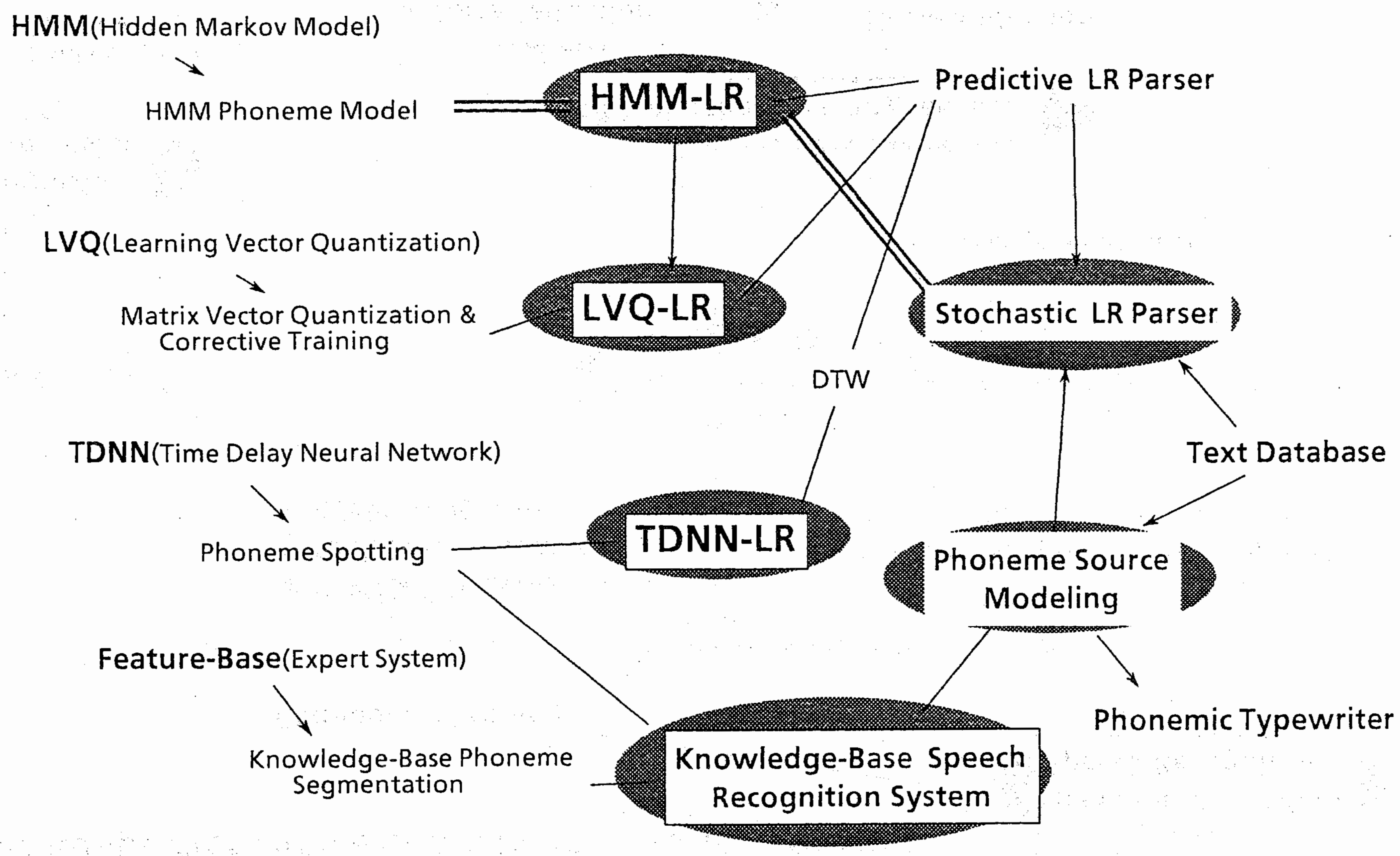
K. Shikano


Phoneme / Word / Phrase Recognition Results
Word / Phrase Candidates
Speech
Speaker Normalization
Stochastic Speech Recognition (HMM)
Speech Recognition Stage
Knowledge-Based Speech Recognition (Feature-Based & Neural Net)
Statistical Language Model (Trigram, Neural Net)
Language Source Model
Parser-Based Language Model (Stochastic LR Parser)
Context Dependent Translation (Unification-Based)
Language Translation
Syntactic Translation (Transfer-Based)
Concept
Sentence
Speech Synthesis by Rule
Voice Conversion
Translated Speech
Phoneme / Word / Phrase Prediction
Linguistic Knowledge

Proposed Interpreting Telephony System (as of July,1988)

Continuous Speech Recognition

Language Source Modeling



HMM(Hidden Markov Model)

HMM Phoneme Model

HMM-LR

Predictive LR Parser

LVQ(Learning Vector Quantization)

Matrix Vector Quantization & Corrective Training

LVQ-LR

Stochastic LR Parser

DTW

TDNN(Time Delay Neural Network)

Phoneme Spotting

TDNN-LR

Text Database

Phoneme Source Modeling

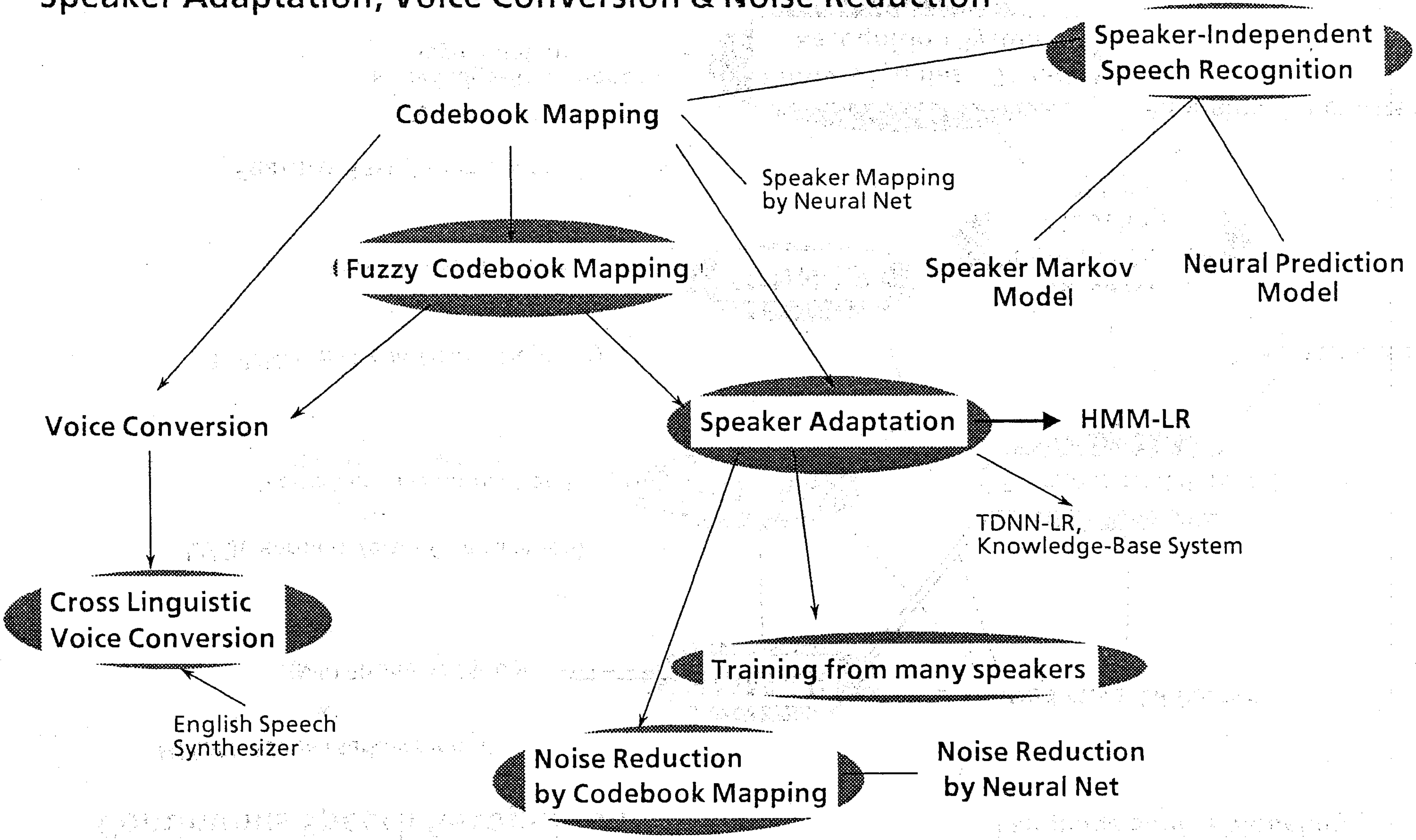
Feature-Base(Expert System)

Knowledge-Base Phoneme Segmentation

Knowledge-Base Speech Recognition System

Phonemic Typewriter

Speaker Adaptation, Voice Conversion & Noise Reduction



Speaker-Independent Speech Recognition

Codebook Mapping

Speaker Mapping by Neural Net

Speaker Markov Model

Neural Prediction Model

Fuzzy Codebook Mapping

Voice Conversion

Speaker Adaptation

HMM-LR

TDNN-LR, Knowledge-Base System

Cross Linguistic Voice Conversion

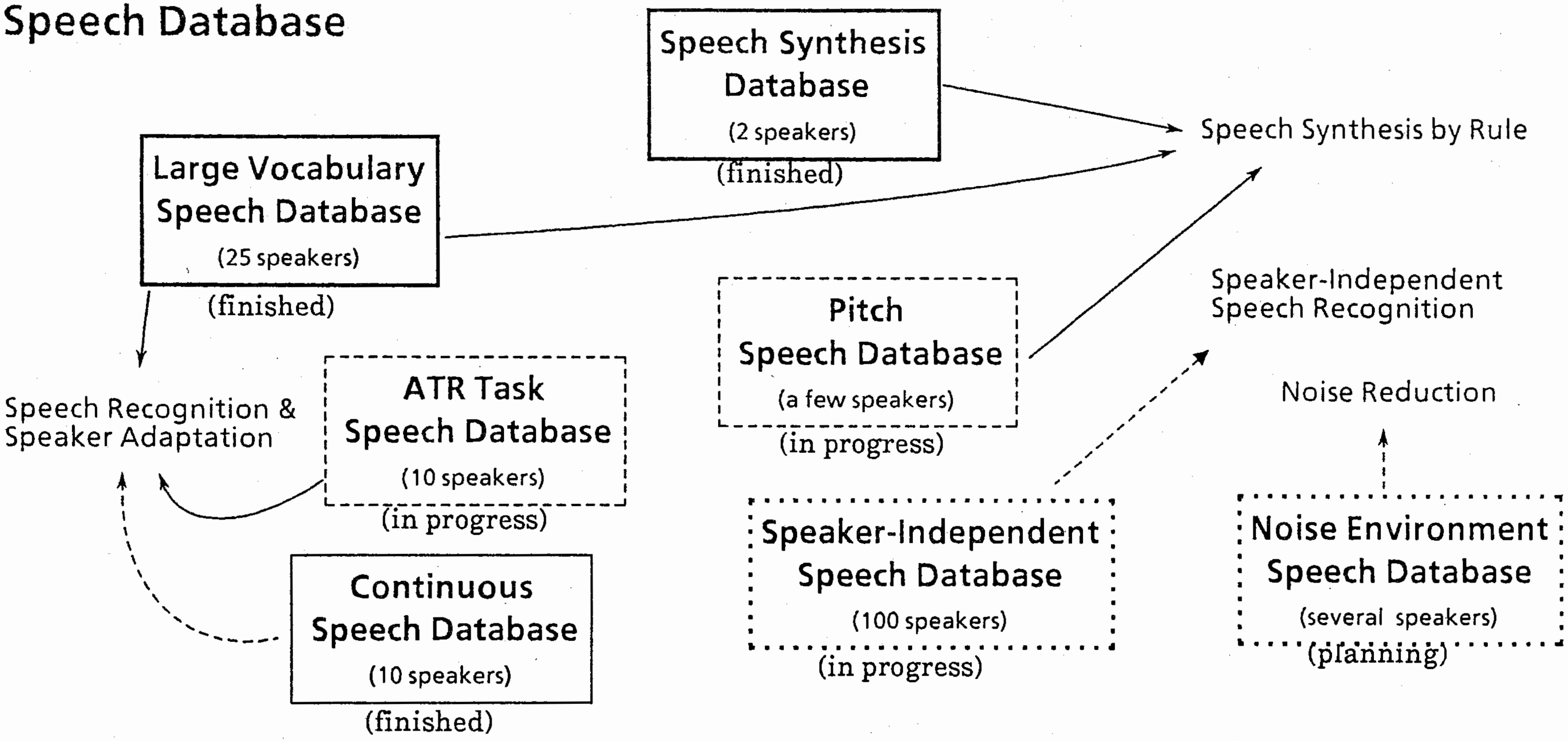
Training from many speakers

English Speech Synthesizer

Noise Reduction by Codebook Mapping

Noise Reduction by Neural Net

Speech Database



Speech Synthesis Database (2 speakers) (finished)

Large Vocabulary Speech Database (25 speakers) (finished)

Pitch Speech Database (a few speakers) (in progress)

ATR Task Speech Database (10 speakers) (in progress)

Continuous Speech Database (10 speakers) (finished)

Speaker-Independent Speech Database (100 speakers) (in progress)

Noise Environment Speech Database (several speakers) (planning)

Speech Synthesis by Rule

Speaker-Independent Speech Recognition

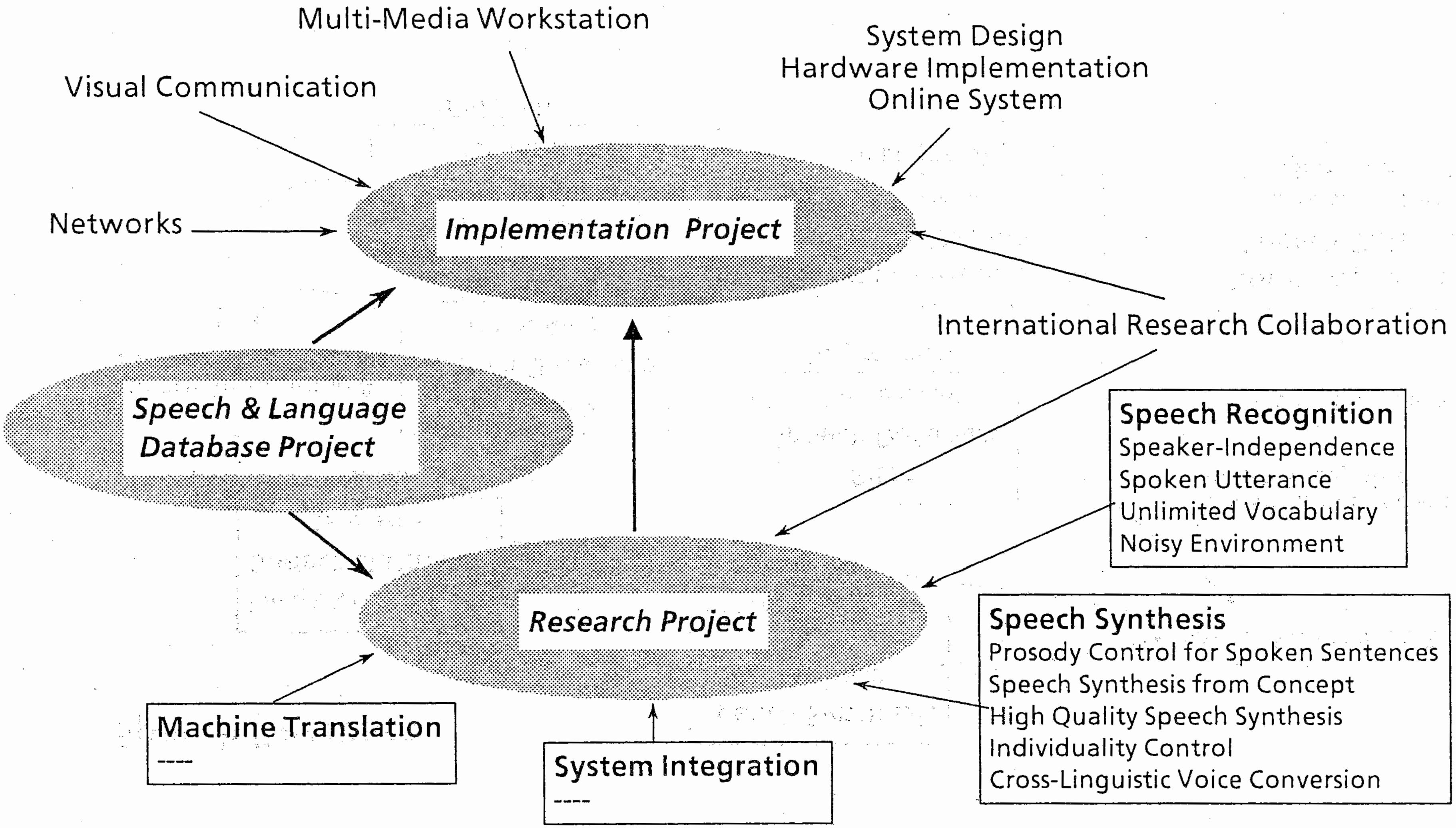
Noise Reduction

Speech Recognition & Speaker Adaptation

Interpreting Telephony Research Projects ⟷ Part of Human-Human Communication through Networks

Multi-Media Paradigm

Future Projects



Multi-Media Workstation

Visual Communication

System Design
Hardware Implementation
Online System

Networks

Implementation Project

International Research Collaboration

Speech & Language Database Project

Speech Recognition
Speaker-Independence
Spoken Utterance
Unlimited Vocabulary
Noisy Environment

Research Project

Speech Synthesis
Prosody Control for Spoken Sentences
Speech Synthesis from Concept
High Quality Speech Synthesis
Individuality Control
Cross-Linguistic Voice Conversion

Machine Translation

System Integration

·Language Model

*Phoneme source modeling
- Open vocabulary
- Unknown words

*Cache ---long time dependency
unification(semantics)
task knowledge.

·Spoken aspects

*Dialogue model--database

·Integration

·How to use global information.

Robustness ----- (Trained by word utterances
Tested by continuous speech

·Speaker adaptation

·Noise

·Acoustic modeling

·Spectral differences
·Speaker variations
·Prosody

·Understanding, intonation transfer

HMM Continuous Speech Recognition Using Predictive LR Parsing

HMM音韻認識と予測LRパーザを用いた連続音声認識

Takeshi KAWABATA

ATR Interpreting Telephony Research Laboratories

HMM Continuous Speech Recognition Using Predictive LR Parsing

Takeshi KAWABATA

ATR Interpreting Telephony
Research Laboratories

HMM (Hidden Markov Model) has the ability to cope with acoustical variation of speech by means of stochastic modeling, and has been used widely for speech recognition[1]. Because any word models can be composed of HMM phone models, it is easy to construct a large-vocabulary speech recognition system using this method.

Generalized LR parsing which was originally developed for programming languages has been extended by Tomita[2] to handle arbitrary context-free grammars. An LR parser is guided by LR parsing tables, which are created from context-free grammar rules, and proceeds left-to-right without backtracking. The primary mechanism for handling ambiguous natural language grammars is stack splitting. When the parser encounters a multiple entry, which is called "conflict", the parsing stack is divided into two stacks and each stack is processed in parallel.

Predictive LR parsing is a further extension of generalized LR parsing for sentence generation. For example, a phone-based predictive LR parser predicts next phones at each generation state, and generates many possible sentences as phone sequences. The predictive LR parser determines next phones using the LR tables of the specified grammar and splits the stack not only for grammatical ambiguity but also for phone variation. Because the predictive LR parser uses CFG rules whose terminal symbols are phone names, the phonetic lexicon for a specified task is embedded in the grammar.

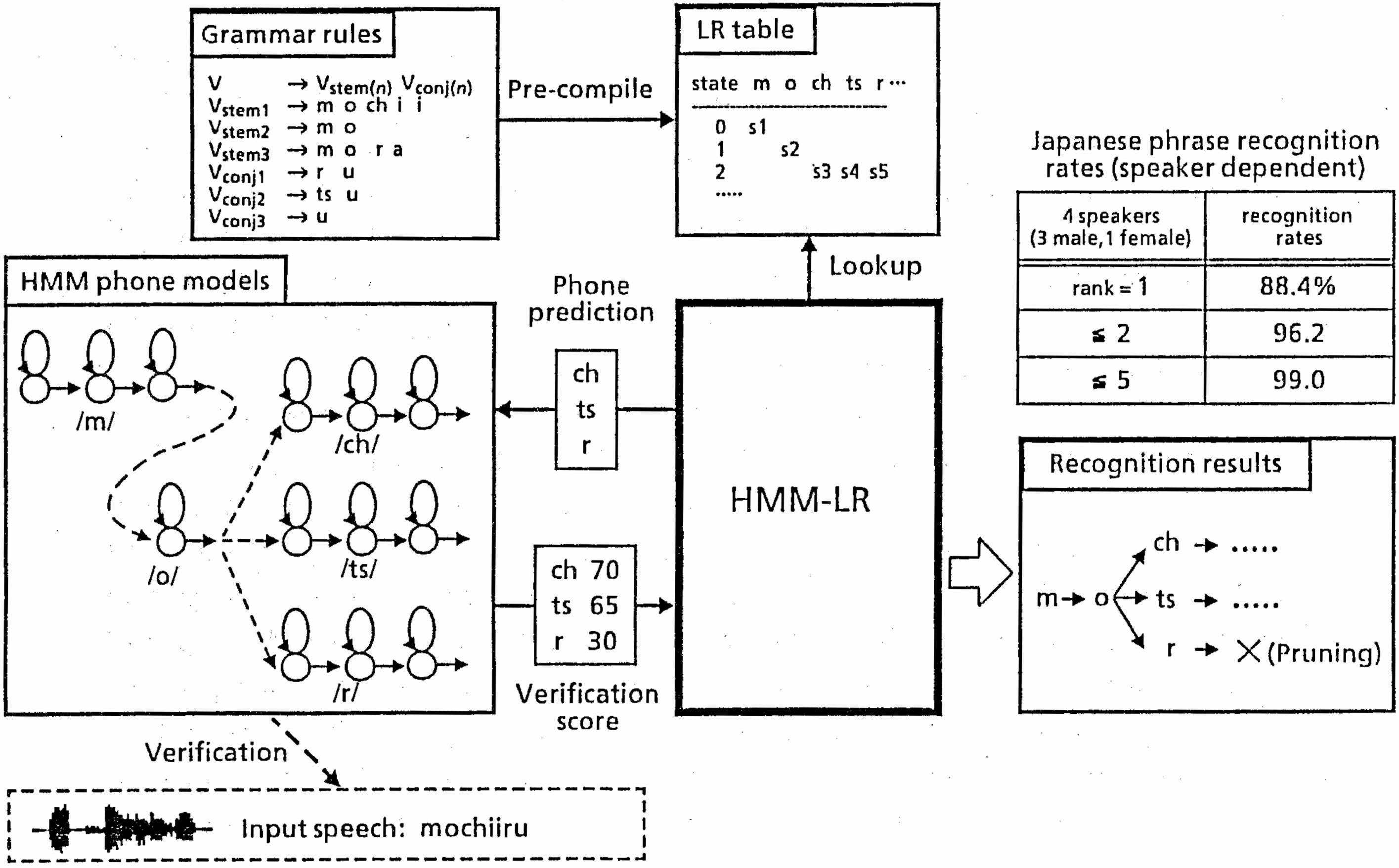
The HMM-LR continuous speech recognition system[3] consists of the predictive LR parser and HMM phone verification mechanism. First, the parser picks up all phones predicted by the initial state of the LR parsing table and invokes the HMMs to verify the existence of these predicted phones. During this process, all possible parsing trees are constructed in parallel. The HMM phone verifier receives a probability array which includes end point candidates and their probabilities, and updates it using an HMM probability calculation process. This probability array is attached to each partial parsing tree. When the highest probability in the array is lower than a threshold level, the partial parsing tree is pruned by threshold levels, and also by beam-search techniques. The parsing process proceeds in this way, and stops if the parser detects an accept action in the LR parsing table.

Thus, an accurate and efficient parsing mechanism is achieved through the integrated process of speech recognition and language analysis. The HMM-LR method was tested through a speaker-dependent Japanese phrase recognition experiment. The grammar used in the experiments describes a general Japanese syntax of phrases and is written in the form of context-free grammar. There are 1,461 grammar rules including 1,035 different words, and perplexity per phone is 5.87. Assuming that the average phone length per word is three, the word perplexity is more than 100. The average phrase recognition rate, for 4 speakers, is 88.4% for the top candidate and 99.0% for the best five candidates.

[1] Levinson, S.E., Rabiner, L.R. and Sondhi, M.M.: "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition", BSTJ, Vol.62, No.4, pp.1035-1074 (1983).

[2] Tomita, M.: "Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems", Kluwer Academic Publishers (1986).

[3] Kita, K., Kawabata, T. and Saito, H.: "HMM Continuous Speech Recognition Using Predictive LR Parsing", ICASSP89, S13.3, pp.703-464 (1989)



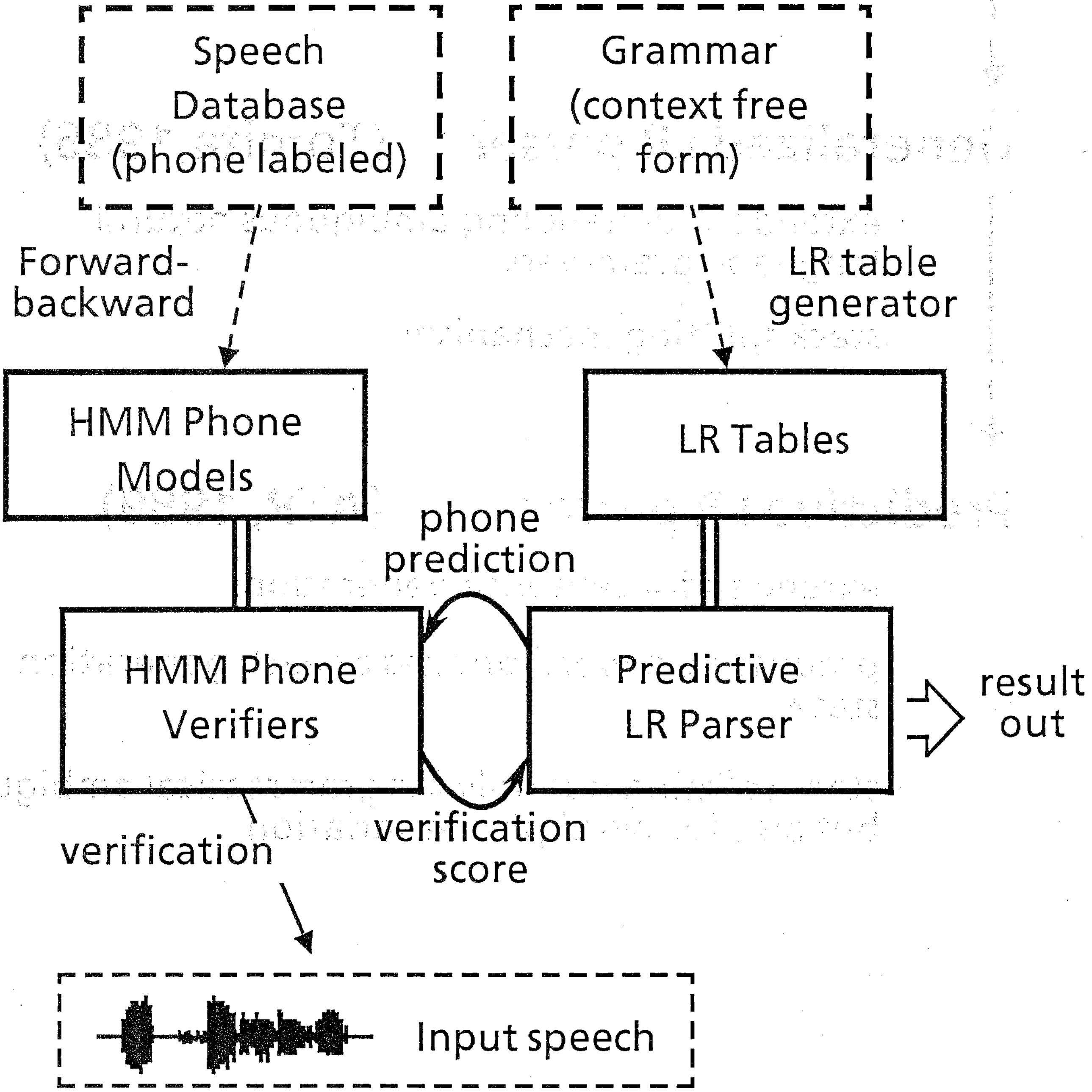
Japanese phrase recognition rates (speaker dependent)

4 speakers (3 male,1 female)	recognition rates
rank = 1	88.4%
≤ 2	96.2
≤ 5	99.0

Fig. HMM-LR continuous speech recognition system

[HMM-LR Concept]

Predictive LR parser predicts next phones and drives correspondent phone verifiers directly.



[Predictive LR Parser]

Traditional LR parser

- deterministic parser for programming languages

Generalized LR parser (Tomita,1986)

- extended for handling ambiguous natural language grammars
- stack splitting mechanism

Predictive LR parser (ATR, 1989)

- extended for sentence generation
- predicts next word/phones at each generation state
- stack splitting not only for grammatical ambiguity but also for word/phone variation

[LR Table Example]

Grammar

(1) S → NP V

(2) NP → N

(3) NP → N P

(4) N → m a m e

(5) N → a r e

(6) P → o

(7) V → o k u r e

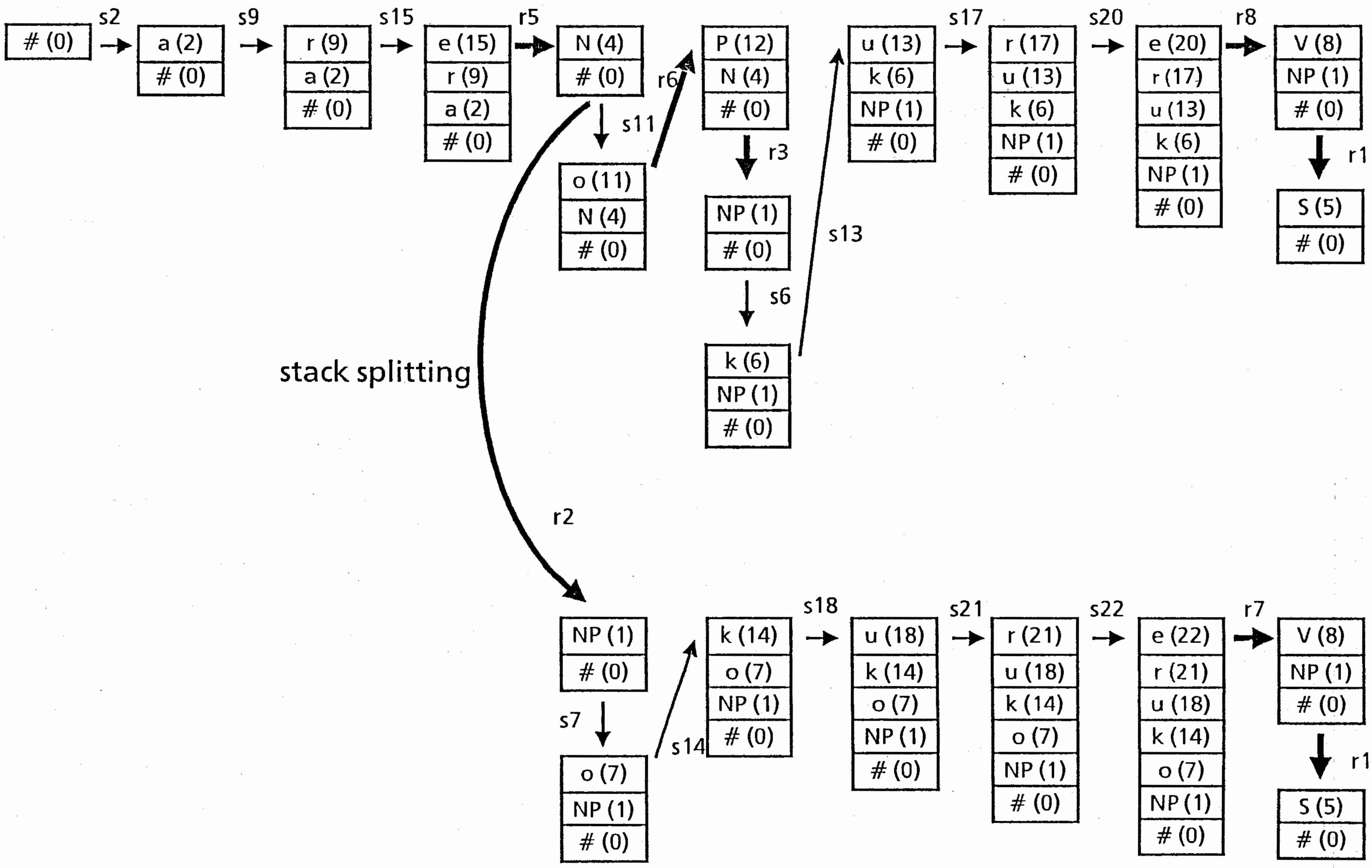
(8) V → k u r e

LR Table

	a	u	e	o	k	m	r	$	S	N	V	P	NP
0	s2					s3			5	4			1
1				s7	s6						8		
2							s9						
3	s10												
4				s11,r2								12	
5								acc					
6		s13											
7					s14								
8								r1					
9			s15										
10						s16							
11				r6									
12				r3									
13							s17						
14		s18											
15				r5									
16			s19										
17			s20										
18							s21						
19				r4									
20								r8					
21			s22										
22								r7					

[Stack Operation in Generalized LR Parsing]

input	a	r	e	o	k	u	r	e	$



(0)

s2

a (2)
(0)

s9

r (9)
a (2)
(0)

s15

e (15)
r (9)
a (2)
(0)

r5

N (4)
(0)

s11

o (11)
N (4)
(0)

r6

P (12)
N (4)
(0)

r3

NP (1)
(0)

s6

k (6)
NP (1)
(0)

s13

u (13)
k (6)
NP (1)
(0)

s17

r (17)
u (13)
k (6)
NP (1)
(0)

s20

e (20)
r (17)
u (13)
k (6)
NP (1)
(0)

r8

V (8)
NP (1)
(0)

r1

S (5)
(0)

stack splitting

r2

NP (1)
(0)

s7

o (7)
NP (1)
(0)

s14

k (14)
o (7)
NP (1)
(0)

s18

u (18)
k (14)
o (7)
NP (1)
(0)

s21

r (21)
u (18)
k (14)
o (7)
NP (1)
(0)

s22

e (22)
r (21)
u (18)
k (14)
o (7)
NP (1)
(0)

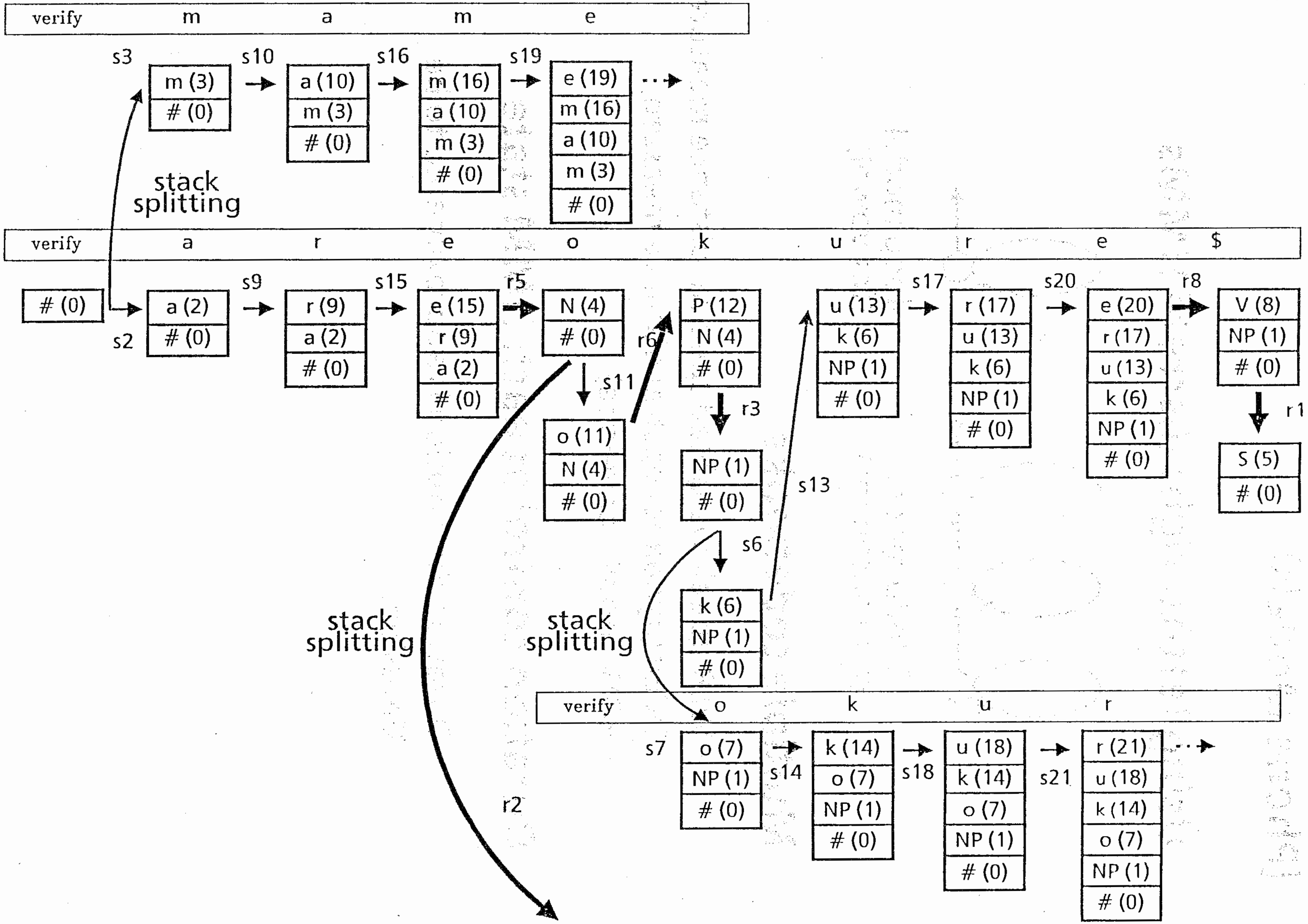
r7

V (8)
NP (1)
(0)

r1

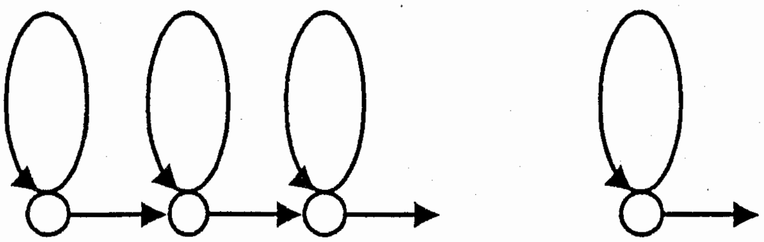
S (5)
(0)

[Stack Operation in Predictive LR Parsing]



[Phone Models]

Using left-to-right discrete HMMs



3-loop model
(for consonant)

1-loop model
(for vowel)

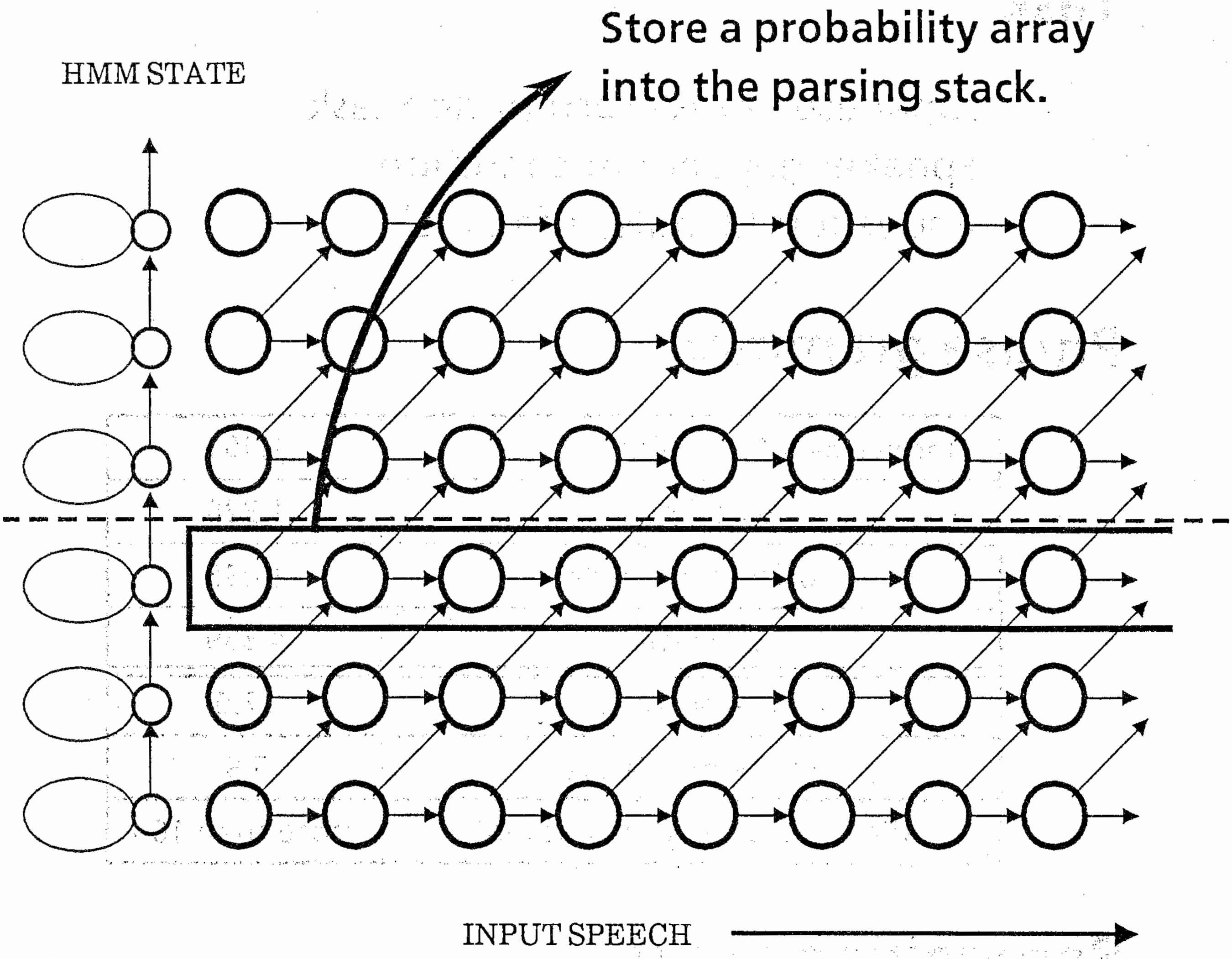
Multiple codebooks

- WLR (peak weighted spectral distance measure)
- DCEP (dynamic cepstral distance measure)
- POW (power envelope)

Duration control at each HMM state

- Duration distribution is approximated to be Gaussian.

[Stacking of Probability Array]



One hundred probability arrays are kept simultaneously in the system (Beam Search).

[Performance]

Task

- Japanese **phrase recognition** task
- Speaker dependent condition
- 4 speakers (3 male, 1 female)

Phrase grammar

The number of ... rules	1,461
lexical rules	1,320
different words	1,035
LR states	4,359
Task entropy	17.0
Phone perplexity	5.9
Word perplexity (estimated)	more than 100

Recognition result

rank	recognition rates
= 1	88.4%
≦ 2	96.2
≦ 5	99.0

[Conclusion]

Accurate continuous speech recognition is realized by combining phone-based HMMs and predictive LR parsing.

The HMM-LR recognizes Japanese phrase utterances with 88.4% accuracy. (speaker dependent)

The HMM-LR speech recognition system is implemented into SL-TRANS† as a speech input module.

† SL-TRANS: ATR Spoken Language TRANslation System

Intention Translation Method : A Spoken Dialog Translation System Using A Lexicon-Driven Grammar

Hitoshi IIDA
ATR Interpreting Telephony Research Laboratories

1. INTRODUCTION

Limited research in natural language communication understanding or interpretation between humans using telephones or keyboards has been done. It is important to retain intention transfer correctness and dialogue smoothness because there are many kinds of intention expressions in natural language dialogues, especially in inquiry dialogues.

This talk will present an intention translation method (ITM) for spoken natural dialogues. This method is characterized by two translating processes: one uses an interlingual strategy for utterance intentions by means of surface speech act types, and the other transfers propositional parts of utterances and aspects. This method can be placed between a 'transfer approach', such as GETA's approach[Boitet], which is a popular MT strategy, and an 'interlingual approach', such as 'script' or 'MOPS'[Lytinen], [Wilensky], [Tucker] using a knowledge representation scheme, which is a difficult approach.

This method allows a generation process to avoid generating an utterance from a logical expression[Appelt]. It integrates the two generation phases for the propositional part and the intentional part represented by surface speech act types into a single utterance. The strategy used by this method is also different from the strategy in which language expression patterns are prepared in advance[Sanford]. Moreover, a new method for interpreting special idiomatic expressions which indicate certain intentions is applied to a spoken dialogue translation system between Japanese and English .

2. GOAL-ORIENTED SPOKEN DIALOGUES

Simulated telephone dialogue data involving a Japanese, an interpreter (mother tongue: Japanese) and an American were collected under the topic of conference registration. An investigation of the data has revealed that[Arita], [Iida] ;

a) almost all the personal pronouns in the English translation correspond to zero pronouns in the original Japanese utterances,

b) the information which already appeared is seldom given in the following utterances,

c) there are many complicated sub-phrases consisting of various auxiliary verbs and particles which represent the speaker's honorific attitudes, for example politeness, euphemisims, etc., and

d) almost all turn-taking chunks correspond to each definite topic, and the discourse structure arising out of the dominance relationships between subgoals becomes clear and simple.

The first three characteristics can be regarded as the main characteristics of spoken natural dialogues. Under the constraint of the last characteristic d) an experimental translation system can be realized.

This talk will also report on the advantages of using both a Lexicon-driven Unification-based Grammar for analyzing such utterances, and an active chart parser with a unification.

2.1 Japanese Ellipses

Honorific expressions and zero-pronouns

Japanese has a rich grammatical system for honorifics. There are many conversational forms for indirect expressions which concern certain explicit expressions[Maeda], [Yoshimoto], for example (**EX-1**) "namae wa nani desu ka" ("what is (your/its) name?").

The following examples contain such an utterance. There are some polite ways of expressing the example **EX-1**.

(**EX-2**) "o-namae wo o-oshie kudasai" ("tell (me) (your) name, (please)")

(**EX-3**) "o-namae wo o-oshie negae masu ka" ("(could) (you) tell/give (me) (your) name?")

(**EX-4**) "o-namae wo o-oshie itadake masu ka" (same as the above)

In **EX-1** the name cannot be identified without a context, but all the rest of the names can be interpreted as names honored by the speaker. In particular, in a limited dialogue with two agents (speaker and hearer), there are two zero-pronouns. Two case roles, the agent and the recipient of the predicate 'tell' (or 'inform'), are identified.

Topics

Topics presented in Japanese utterances are usually marked by the case-particle 'wa' and sometimes by 'to ie ba' (nearly equal to 'as for'), etc. After that, they need no longer be expressed in the following utterances.

2.2 Intention Expressions

In the registration query task, the communicative act set, {"REQUEST (Speaker, Hearer, ACT(H))" , "INFORM (S, H, Proposition)"} is established. REQUEST calls for an action "INFORMIF (P)" or "INFORMREF (P)", which respectively demand either a yes/no answer of H or a referent identification of H. For example, the representation in **EX-2** has a surface speech act type REQUEST (S, H, INFORMREF (P)). In the following communicative act examples, EX-1 and EX-2, Japanese underlined words and phrases correspond to the English equivalents. Non-underlined phrases indicate certain intentions. The utterance in **EX-3** is an idiomatic expression.

(**EX-5**) REQUEST (S, H, INFORMIF (P)) :

"tourokuyoushi wa o-moti des-you ka" (" Do you have a registration form ? ")

(**EX-6**) INFORM (S, H, P) : "shouchi itashi mashi ta" (" All right .")

The communicative acts are arranged in speech acts. There are many conventional forms for indirect expressions which concern certain explicit expressions. **EX-1** is such an utterance. There is a polite way of expressing this example:

" o-namae wo o-kika-se kudasai masu ka " (" Could you give me your name?"). For example, 'negaeru' (a verb which means 'can-wish') is classified 'Indirect-request' and corresponds to the speech act type "CAN-REQUEST (S, H, P)." It must be clarified that such a Japanese function word has an intention. The interesting point is the focus on interpreting the forms from Japanese to English and vice versa.

3. METHOD FOR SPOKEN DIALOGUE TRANSLATION

Our interpretation approach is to construct and decompose feature structures by calculating feature structures using a unification-based approach. Handling dialogues by means of feature-based descriptions has various advantages over other methods. For example, it is easier to handle Japanese phrases indicating intentions and to compensate for the missing phrases. The objects under the interpretation process are uttereance semantic representations. A semantic representation is conceptually divided into a 'propositional-part' and an 'intentional-part'. The former is represented by a predicate and case roles, some of which occasionally have embedded propositional parts. The latter is represented by speech acts.

3.1 Unification-Based Utterance Analysis

The main characeristics of the unification-based approach are as follows[Kogure].

- Lexico-syntactic analysis by an active chart parser using Japanese Phrase Structure Grammar,
- Treatment of complex predicates, which in spoken Japanese are the expressions most indicative of the speakers' intentions,
- Zero-anaphora resolution.

3.1.1 Unification-based approach

Unification calculation usually permits integrated descriptions of information from various sources. In analyzing spoken utterances such as fragmental and various intentional phrases, constraints between syntax, semantics and pragmatics can be described in terms of feature structures. Therefore, this approach is suitable for analyzing a fragmental sentence or a sentence written in a contextual language such as Japanese.

Analyzing spoken utterances in a traditional rule-based approach is very likely to require numerous complicated grammatical rules because those utterances are fragmental and have various intention expressions. A lexico-syntactic approach can resolve the problem. In this approach a grammar has only a small number of general syntactic rule schemata, and most of the grammatical information must be specified in descriptions of lexical items. Therefore, it is easy to extend a grammar simply by adding new lexical items to the lexicon or addding new information to lexical items.

3.1.2 Head-driven Phrase Structure Grammar for Japanese utterance

In order to implement the unification-based approach, a complement-head grammatical structure based on a version of Head-Driven Phrase Structure Grammar (HPSG) [Shieber], [Karttunen] and Japanese PSG [Gunji] is adopted. The grammar describes not only syntactic and semantic information but also pragmatic information in an integrated way by using feature structure descriptions.

3.1.3 Intention expression treatment

The appearance and conjugation properties of sentence final position predicate constituents are generally restricted by the heads immediately following them. These restriction conditions are dealt with by means of the SUBCAT feature.

The lexical description of a sentence final particle "ka" expresses a question attitude. The sentence in **EX-5** is analyzed and its surface speech act types are derived from the SEMantic-feature value after ellipsis resolution. A lazy evaluation on an active chart parsing makes the supplement possible using the PRAGmatic-feature with pragmatic information, ex. default values for a speaker and a hearer,

3.1.4 Zero-pronoun supplement

Many zero-pronouns can be resolved by using pragmatic felicity conditions on uses of honorific expressions. To represent the pragmatic conditions, the PRAG feature is introduced. For example, the conditions on the use of auxiliary verb "itadakeru" (in **EX-4**) is described as the <PRAG RESTR> feature. The HONOR-UP relationship from the speaker to the subject agent of "itadakeru" is subcategorized, and the EMPATHY-DEGREE relationship between the subject and the object is required. Each slash element is examined to determine whether or not the constraints in its <PRAG RESTR> are compatible with previously introduced discourse objects.

3.2 Utterance Transfer

Propositional contents as analysis results are represented by predicates and some case roles in feature-based description. On the other hand, it is comparatively easy to translate a case-frame structure into one in another language. Therefore, the main job of the translating process is to map the case roles under one predicate in the source language onto case roles under other predicate in the target language, for example English. The load is reduced because it is not necessary for the processes to prepare as many transfer rules as syntactic-based translation systems require. Therefore, the transfer module accepts an analysis result represented by a feature-based description and aspect feature is, of course, an object of this process. Two different types of representations, propositional contents and intention expressions, are translated separately by a feature structure rewriting system. In an experimental translation system the latter is considered to be a language-independent representation and passes through the transfer process.*

3.3 Utterance Generation Considering Intention Expressions

The result of the translation process is a case-frame form. For example, the parser can get a semantic representation of an utterance,

"kaigi	ni	moushikom-i	tai	no	desu	ga"
(conference	ACC	register	want	COMPL	COPLA	MODERATE)

'I would like to make a registration for the conference'.

The contents of ga-MODERATE and tai-DESIRE features concerning surface speech act types are represented by other frames. The generation process has two phases: the first is to make up a target description on a surface linguistic structure concerning a propositional content and intentional contents. The second is to spell out the description according to the rules of target language writing style and morphology.

On the first stage, a case-frame dominated by a predicate is expanded into a sequence of each constituent, i.e. a surface case element, according to a sentence pattern. Moreover, an auxiliary phrase or an utterance form is determined according to the intention description. The second stage follows. The method is a very ordinary one.

4. OVERVIEW OF THE EXPERIMENTAL TRANSLATION SYSTEM

The experimental system consists of three main modules: Analyzer, Translator and Generator[Iida89]. The Analyzer has three main parts. The first is an 'input converter' which converts a Kanji-kana character input string into a sequence of candidate symbol lists. The second is an active chart parser which receives the sequence and makes the feature structures by unification operations. The third is the feature structure unification program which applies

* : We are developing an extended ITM. The method extracts illocutionary act types from input utterances and transfers them into target language illocutionary act relationships. [Kume]

grammar objects.

The second module, Translator, has a feature structure rewriting engine and a rewriting rule adaptation module which can select an appropriate rewriting rule for a dialogue situation from a rule set conerning context dependent meanings.

The Generator has two main modules, rewrite-engine and word- linearizer. It refers surface linguistic patterns written in a part of a description for a verb-type transfer dictionary, and morphological tables.

The system has some good features. One is an ability to accept lattice inputs, for example phoneme lattices and phrase lattices, from a speech processing system. Another is using a structure sharing technique to cut down duplicate phrase structures and to reduce unification times in the middle of parsing. The translation system has been implemented in a Symbolics Common Lisp from Japanese to English.

5. CONCLUSION

This paper presented a spoken dialogue translation method which mainly handles intention expression in each utterance, ellipsis resolution, in particular Japanese zero-anaphora, and utterance transfer. An experimental spoken dialogue translation system using a lexicon-driven grammar, case-frame transfer technique and utterance generation combined with intention expressions is also presented.

Currently we are designing unification based translation and generation modules and developing a totally consistent translation system. Our lexico-syntactic parsing method can be applied to many languages in addition to Japanese. An English-to-Japanese translation system using an English lexicon is planned.

REFERENCES

[Appelt] D.E.Appelt,"Planning English sentences", Cambridge U. P., 1985.

[Arita] H.Arita et al.,"Media-Dependent Conversation Manners - Comparison of Telephone and Keyboard Conversations - " (in Japanese), Tech. Rep. no NL61-5, Inf. Proc. Soc. of Japan, 1987.

[Boitet] Ch.P.Boitet et al.,"A case study in soft. evolution: from ARIANE-78 to -85", The Conf. on Theoretical and Method. Issues in MT, Hamilton, 1985.

[Gunji] T.Gunji,"Japanese Phrase Str. Grammar", Dordrecht, D. Reidel, 1987.

[Iida] H.Iida,"Pragmatic Characteristics of Natural Spoken Dialogues and Dialogue Processing Issues" (in Japanese), Journal of Artificial Intelligence of Japan, vol. 3, no. 4, 1988.

[Iida89] H.Iida,"An Experimental Natural Spoken Dialogue Translation System Using a Lexicon-Driven Grammar", Euro-Speech89, 1989.

[Karttunen] L.Karttunen,"D-PATR", Stanford U. Report no. CSLI-86-61, 1986.

[Kogure] K. Kogure et al.,"A Method of Analyzing Japanese Speech Act Types", 2nd International Conference on Theoretical and Methodological Issues in Machine Translation of NLs, 1988.

[Kume] M. Kume et al.,"A Descriptive Framework for Translating Speaker's Meaning", 4th European ACL, 1989.

[Lytinen] S.Lytinen &R.Schank,"Representation and Translation", Tech. Rep. 234, Yale U., 1982.

[Maeda] H. Maeda et al.,"Parsing Japanese Honorifics in Unification-Based Grammar", 26th Annual Meeting of the ACL, 1988.

[Sanford] D.L.Sanford & J.Roach,"Representing and using metacommunication to control speakers' relationships in NL dialogue", Man-Machine Sudies vol.26, no.3, 1987.

[Shieber] S.M.Shieber,"An Introduction to Unification-Based Approaches to Grammar", CSLI Lecture Notes, no. 4, 1986.

[Tucker] A.B.Tucker,"Current strategies in MT research and development", in 'Machine Translation' edited by S.Nirenburg, Cambridge U. P., 1987.

[Wilensky] R.Wilensky & M. Morgan,"One Analyzer for Three Languages", UCBerkeley TR-M 81-87, 1981.

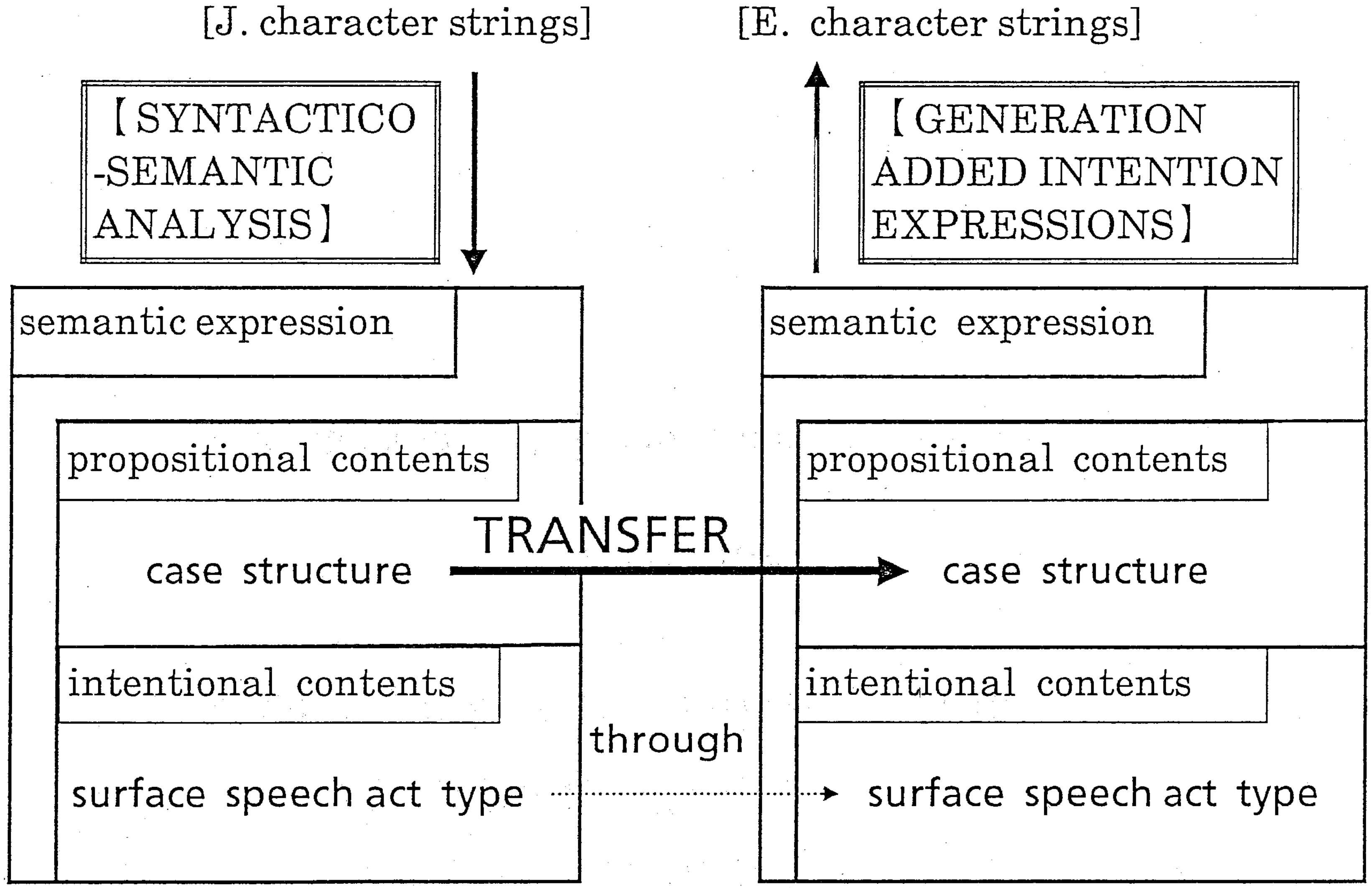
[Yoshimoto] K. Yoshimoto,"Identifying Zero Pronouns in Japanese Dialogues", The proceedings of COLING88, 1988.

Intention Translation Method: A Spoken Dialog Translation System Using A Lexicon-Driven Grammar

Hitoshi IIDA

ATR Interpreting Telephony Research Laboratories

Outline of the Dialogue Translation Method (DTM)



[J. character strings]

[E. character strings]

【SYNTACTICO -SEMANTIC ANALYSIS】

【GENERATION ADDED INTENTION EXPRESSIONS】

semantic expression

propositional contents

case structure

intentional contents

surface speech act type

TRANSFER

through

semantic expression

propositional contents

case structure

intentional contents

surface speech act type

Characteristics of linguistic Expressions in J. Dialogues

(**EX-1**) "(______) namae wa nani desu ka"

: → "what is (your/something)-name"

<<some polite ways>>

(**EX-2**) "o-namae wo (______-ni) o-oshie *kudasai*"

: → "tell (______) (______)-name, (please)"

(**EX-3**) "o-namae wo (______-ni) o-oshie *negae* masu ka"

: → "(   ) tell/give (______) (______)-name ?"

(**EX-4**) "o-namae wo (______-ni) o-oshie *itadake* masu ka"

: → "(   ) tell/give (______) (______)-name ?"

;;; parentheses are used to denote missing phrases

Intention Expressions

(EX-a) INDIRECT-REQUEST (S, H, INFORMIF (P)) :

"(____-ga) tourokuyoushi wa o-moti des-you ka"

: → *"Do you have a registration form ? "*

P : have (you, a-registration-form **)**

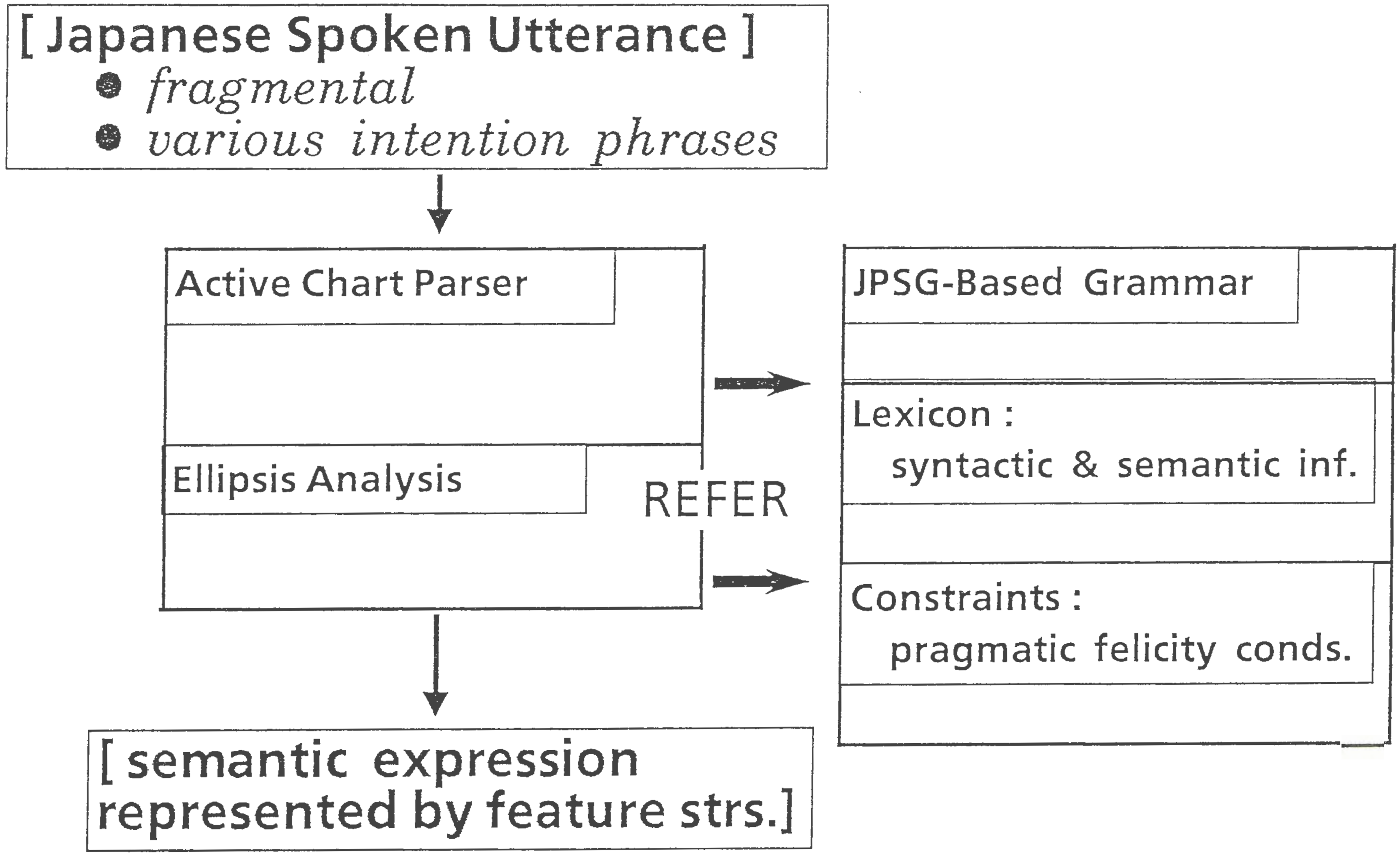
(EX-b/3) CAN-REQUEST (S, H, P) :

"(__-ga) (__-ni) o-namae wo o-oshie negae masu ka"

: → *"Could you give me your name ? "*

P : tell (you, I, your-name **)**

Unification-Based Utterance Analysis



[Japanese Spoken Utterance]

- *fragmental*
- *various intention phrases*

Active Chart Parser

Ellipsis Analysis

REFER

JPSG-Based Grammar

Lexicon :
syntactic & semantic inf.

Constraints :
pragmatic felicity conds.

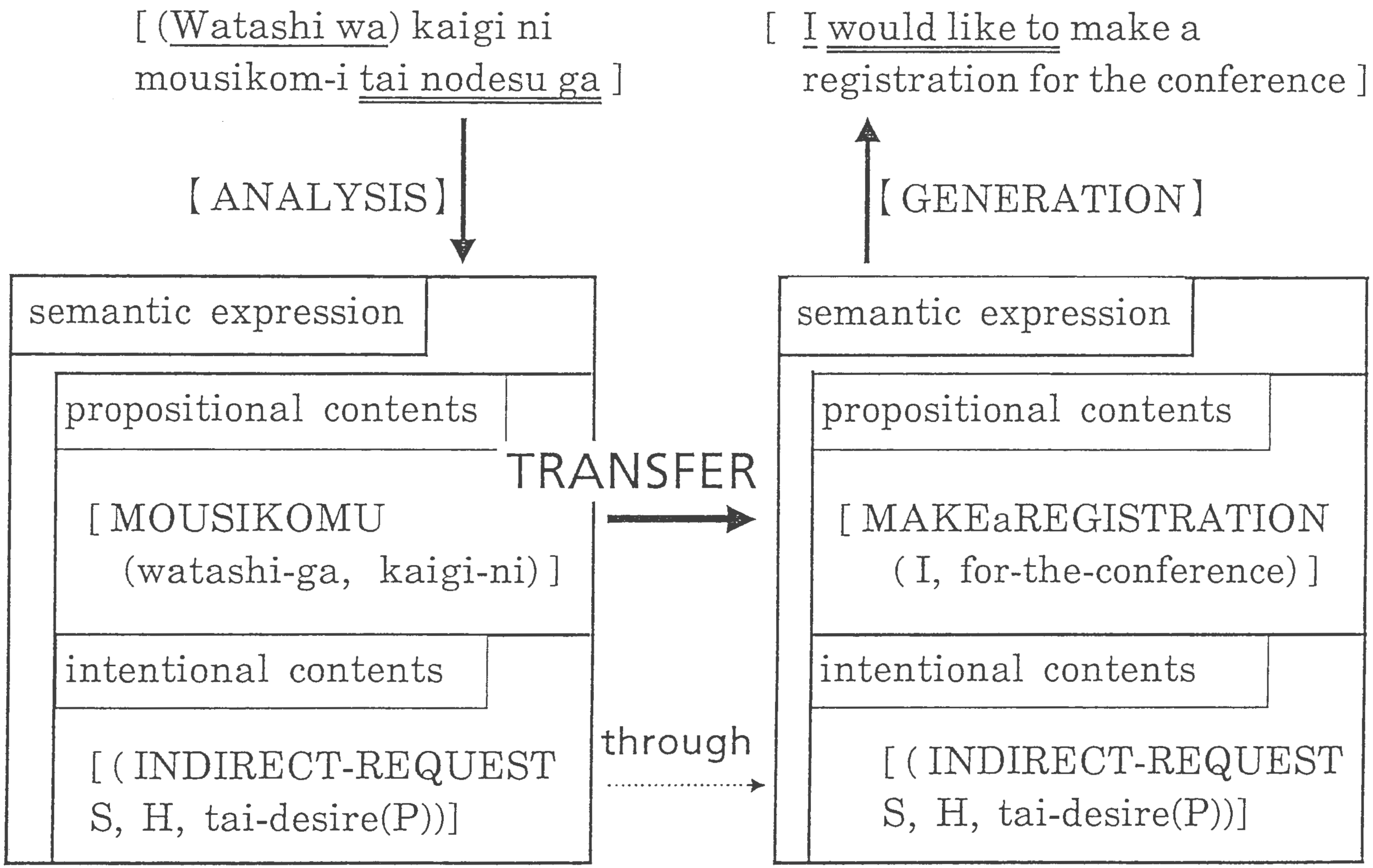
[semantic expression
represented by feature strs.]

Analysis Results (*'Do you have a registration form?'*)

```
[[SEM ?X14[[RELN  S-REQUEST]
       [AGEN ?X05[[LABEL *SPEAKER*]]]
       [RECP ?X02[[LABEL *HEARER*]]]
       [OBJE [[RELN  INFORMIF]
            [AGEN ?X02]
            [RECP ?X05]
            [OBJE [[RELN you-GUESS] ;;; you: aux. verb
                [EXPR ?X05]
                [OBJE [Proposition  motu-1]] ]]]]]  ;;; motu: verb (have)
 [PRAG [[RESTR  [[IN  [[FIRST  [[RELN  POLITE]
                          [AGEN  ?X05]
                          [RECP  ?X02]]]            . . . ]]
                         [OUT  ?X06]]]
           [SPEAKER  ?X05]
           [HEARER  ?X02]]]]
```

(for **EX-5** ' tourokuyoushi wa o-mochi des-you ka')

Outline of the DTM Transfer



[(Watashi wa) kaigi ni
mousikom-i tai nodesu ga]

[I would like to make a
registration for the conference]

【ANALYSIS】

【GENERATION】

semantic expression

propositional contents

[MOUSIKOMU
(watashi-ga, kaigi-ni)]

intentional contents

[(INDIRECT-REQUEST
S, H, tai-desire(P))]

TRANSFER

through

semantic expression

propositional contents

[MAKEaREGISTRATION
(I, for-the-conference)]

intentional contents

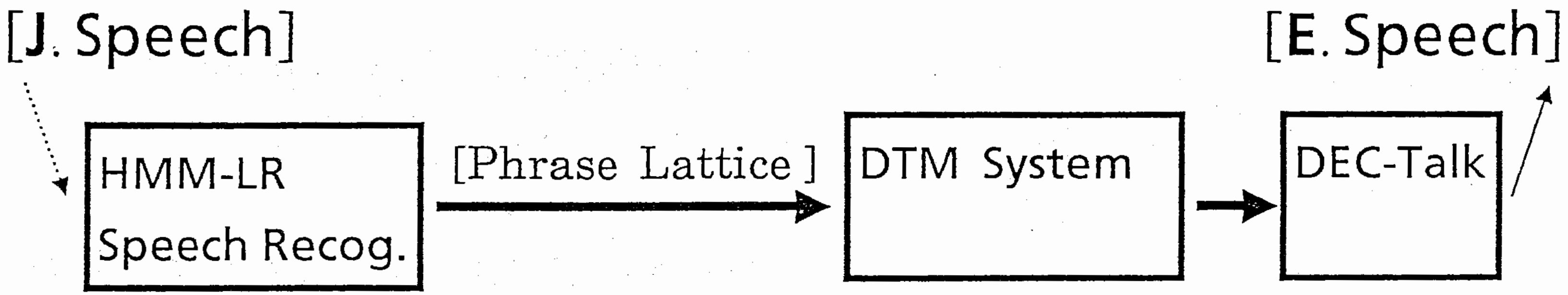
[(INDIRECT-REQUEST
S, H, tai-desire(P))]

Translation results & System Environment

	Japanese inputs:(kanji-characts.)		English outputs:
A:	もしもし そちらは会議事務局ですか	A:	Hello. Is that the office for the conference?
B:	はい　そうです	B:	Yes.　That is right.
A:	(私は)会議に申し込みたいのですが	A:	I would like to make a registration for the conference.
B:	(あなたは)登録用紙は既にお持ちでしょうか	B:	Do you already have a registration form?
A:	いいえ　まだです	A:	No.　Not yet.
B:	わかりました それでは (私は　あなたに) 登録用紙をお送りいたします	B:	I see. Then, I send you a registration form.

;;; where parentheses are used to denote missing phrases

;;; where double underlines are used to denote supplemented phrases



Advanced Dialogue Translation Techniques : Plan-Based, Memory-Based and Parallel Approaches

Hitoshi IIDA
ATR Interpreting Telephony Research Laboratories

I will focus my comments on advanced dialogue translation techniques and will discuss translation quality, robustness and processing cost. Recent work at ATR can be broken down as follows.

1. HIGH QUALITY TRANSLATIONS

Plan Recognition Model Using Three Types of Pragmatics for Dialogue Understanding :

A plan recognition model for resolving ellipses of phrases or choosing an appropriate translated word was proposed. Various predictions of the next utterance necessary for the integrated process of speech and language will also be performed in this framework.

The model consists of plans, objects, and inference rules. Four-typed plans for task-related knowledge and three-typed pragmatics are used in the current model : domain plans, which represent the structure of domain-dependent action hyerarchies; dialogue plans, which manage the global change of topics in the domain; communication plans, which represent a sequence of communicative acts for information exchange; and interaction plans, which manage the local structure of the dialogue, i.e. a demand-response pair in a goal-oriented dialogue. An analyzed utterance in a demand matches the decomposition of an interaction plan. Then the interaction plan matches the decomposition of a discourse plan which introduces an object and a domain plan. A response utterance matches the decomposition of an interaction plan which has already been instantiated. Plan chaining is done via Decomposition, Effects and Constraints described in the slots of a plan scheme. A prototype system of plan recognition based on this model has been implemented for the demand-response sample dialogues about the conference.

Interpreting Japanese utterances based on context information :

Shared goals and mutual beliefs between dialogue participants are mainly taken as the context in task-oriented Japanese dialogue. Communicative acts performed by dialogue participants can be interpreted based on such context information. Honorific relationships between dialogue participants, the speaker's point of view and the speaker's territory of information are also regarded as the context in Japanese dialogue.

In a developed interpretation model, the state of dialogue is composed of communicative act situation, mutual belief situation and shared goal situation. The speaker's communicative acts convey information on communicative act situations. Context information is included in mutual belief situation and shared goal situations. Constraints are defined between those situations. Communicative acts are interpreted under the context information through the constraints. The interpretation process is made on the basis of the Constraint Application Method and the Information Merging Method. Through the Information Merging Method, elliptical information in utterances can be filled based on the known context information.

A model to identify zero-pronouns referring to persons in Japanese dialogue has been developed. In this model, pragmatic constraints on honorific relationships between persons in a dialogue, the speaker's point of view and the speaker's territory of information are fetched from utterances. The context is a set of satisfactory conditions, which the utterance constraints must satisfy. The constraints fetched from utterances are interpreted under the current context. The interpretation of constraints under the context is a process of rewriting the constraints with satisfactory conditions and unified parameters for zero-pronouns with appropriate constants for persons in the context.

2. ROBUST TRANSLATIONS

In order to establish a robust translation technology which resolves rule-based concrete translation problems, an example-based paradigm has been adopted. Example based processing consists of finding preceding examples similar to the current object and adapting the examples to solve new problems. This new paradigm has three major advantages: (1) it allows the improvement of system performance simply by improving the example database; (2) it assigns reliability factor to the result; (3) it enables quick reasoning by indexing and parallel computing. The method is now being applied to translation of noun groups connected by the particle "NO".

A translation subsystem for the noun groups has been implemented. First, about 1,000 examples (noun groups and their English examples) have been extracted from ATR's Linguistic database. Secondly a lexicon with English equivalents and thesaurus code as semantic information has been built for about 2,500 Japanese nouns. The current translation mechanism is based on a simple similarity calculation and sort function and it will be modified in order to improve accuracy and speed. A retrieval mechanism based on the importance determined by rules is also developed.

3. HIGH SPEED TRANSLATIONS

An interpreting telephone system requires real-time or quasi real-time processing, particularly a Japanese dialogue analyzer which handles a lattice input. The performance of ATR's unification based parser is now being improved by using parallel processing.

A suitable parallel algorithm for a unification based parser has been designed, and its validity and performance on a parallel computer iPSC2 is being verified. An experimental parallel parser has been developed on iPSC2 (INTEL). The main characteristic of this parser is decomposition of the unification processees on distributed PE(Processing Element) with low communication overhead.

Advanced MT Techniques at ATR

Hitoshi IIDA

1. High Quality Translations

research topics — approaches

◇ dialogue understanding
◇ context processing

- ▸using pragmatics for dialogues
  - •cooperative dialogue developments
  - •constraints on honorific relationships
- ▸plan-based inference

2. Robust Translations

◇ context-dependent expressions
◇ idiomatic expressions
◇ ungrammatical word sequences

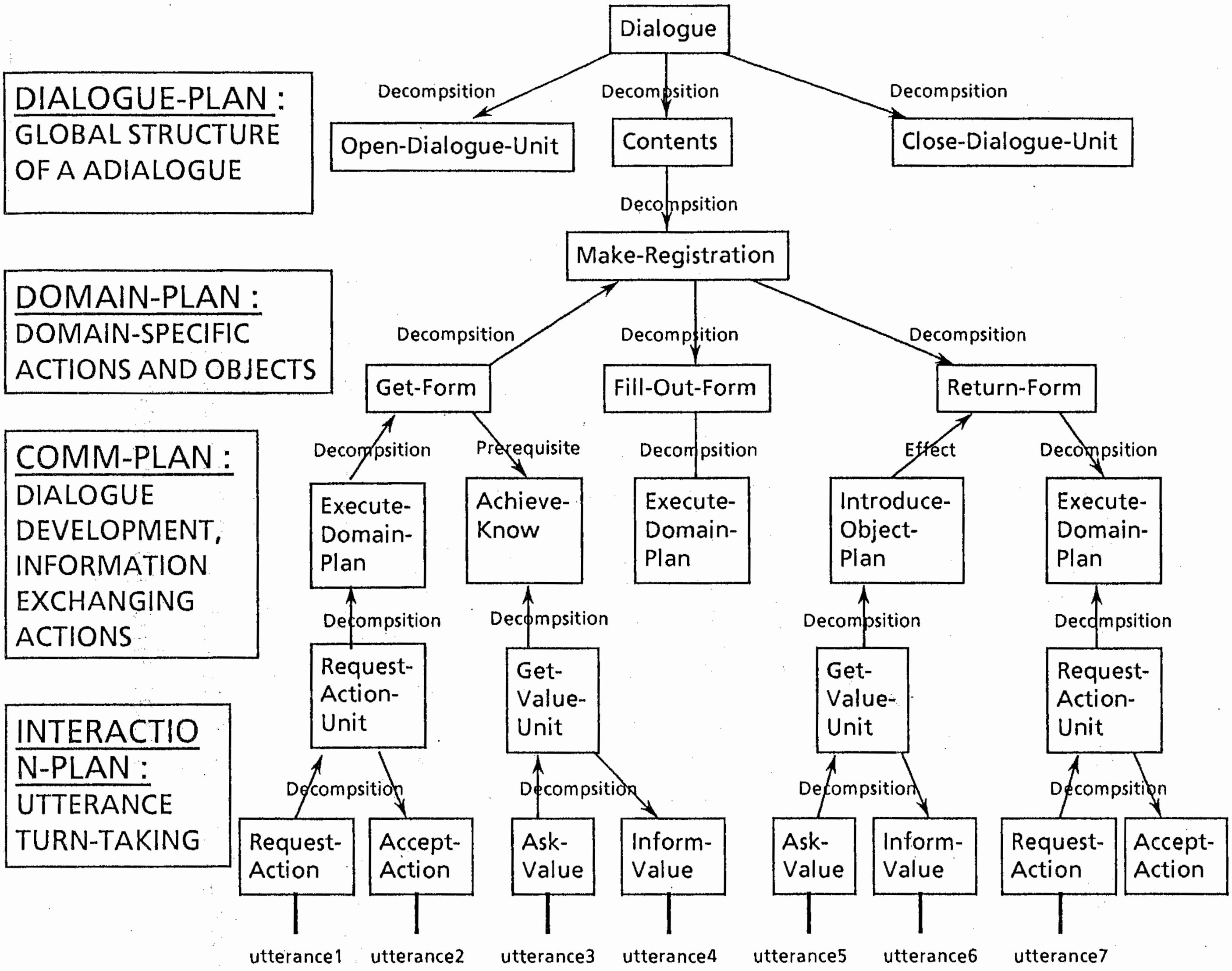
- ▸example-based MT
- ▸case-based MT
  - (•context processing)

3. High Speed Translations

◇ cutting down on the number of information combinations

- ▸parallel processing

Dialogue Structure Construction Process

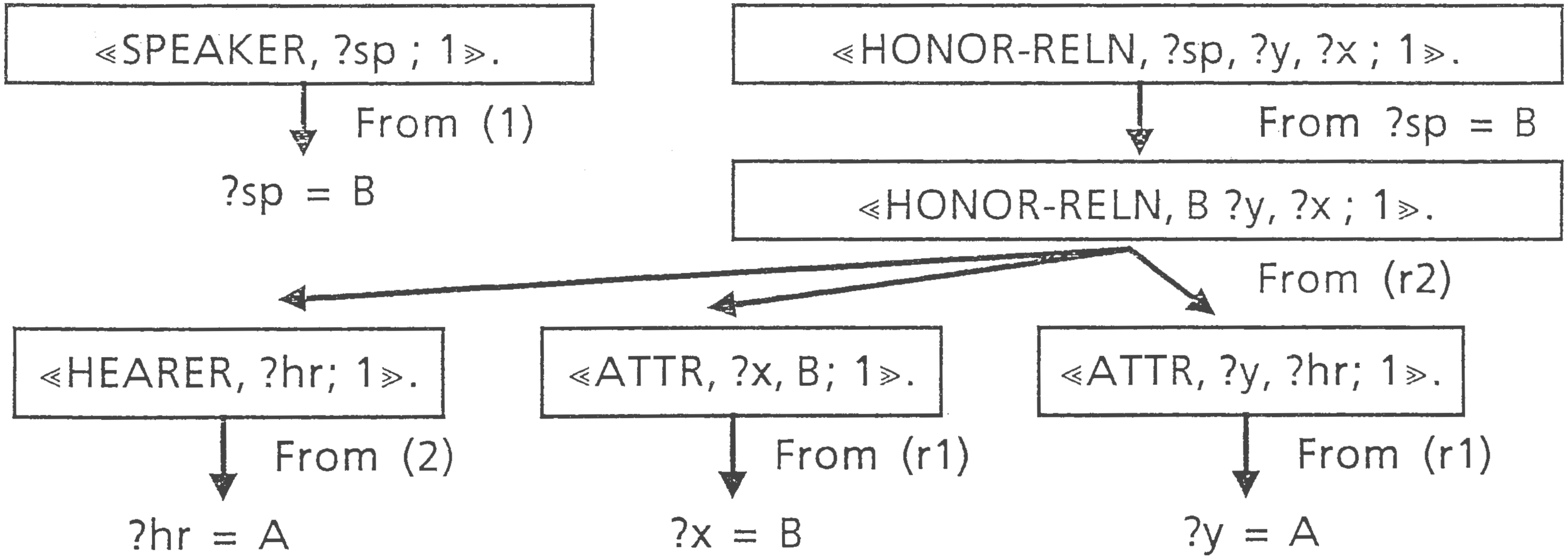


DIALOGUE EXAMPLE:

utterance1:	questioner:	Will you send me a registration form?
utterance2:	secretariat:	All right.
utterance3:		Will you give me your name and address?
utterance4:	questioner:	(My) name is Mayumi Suzuki, and (my) address is .....
utterance5:		When is the deadline?
utterance6:	secretariat:	December 1.
utterance7:		Please hurry. ;; ‘literal translation from Japanese’ (Please return the form as soon as possible.)

Context Processing Using Pragmatic Constraints

B: "Soredewa, (?x ga) (?y ni) touroku-youshi wo o-okuri-itasi-masu."
then ϕ SUBJ ϕ OBJ2 registration form OBJ1 send HONORIFICS
(*Then, I will send you a registration form.*)



Context

≪SPEAKER, B ; 1≫. (1)

≪HEARER, A ; 1≫. (2)

≪ATTR, *x, *x ;1≫. (r1)

≪SPEAKER, *sp ;1≫∧≪HEARER, *hr ;1≫∧ (r2)
≪ATTR, *y, *sp ;1≫∧≪ATTR, *x,*hr ;1≫
⇒≪HONOR-REL, *sp, *x, *y ;1≫.

Translation of “N_1 no N_2“

[Input]
京都での会議

[Retrieved Examples]

distance	Japanese	English	translation pattern
0.4	**東京 での 滞在**	**the stay in Tokyo**	**B in A**
0.4	**香港 での 滞在**	**the stay in Hong Kong**	**B in A**
0.4	**東京 での ご滞在**	**the stay in Tokyo**	**B in A**
1.0	大阪 の 会議	the conference in Osaka	B in A
1.0	東京 の 会議	the conference in Tokyo	B in A

[Output]
B in A → the conference in Kyoto

Distance calculation

Example Distance

$d(I,E) = \Sigma d(I_i,E_i) \times w_i$

ex. d(京都での会議,東京での滞在)

= d(京都,東京) × 0.49

+ d(での,での) × 1.0

+ d(会議,滞在) × 0.54 ≒ 0.4

Weight of attributes

$w_i = \sqrt{\Sigma(\text{freq. of translation pattern } k \text{ when } E_i = I_i)^2}$

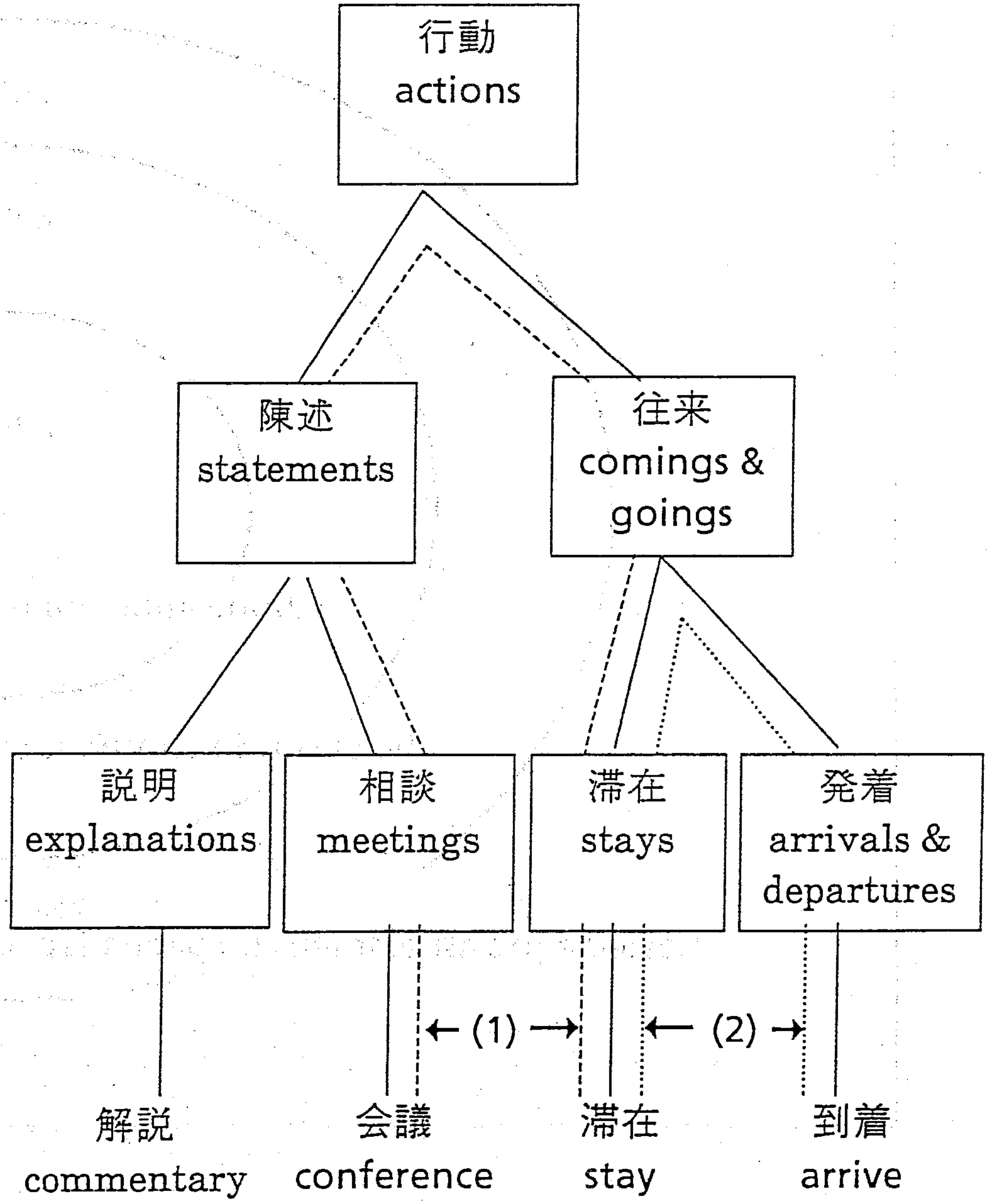
Semantic distance

Semantic distance d (0≦d≦1) is a real number proportional to the Most Specific Common Abstraction (MSCA).

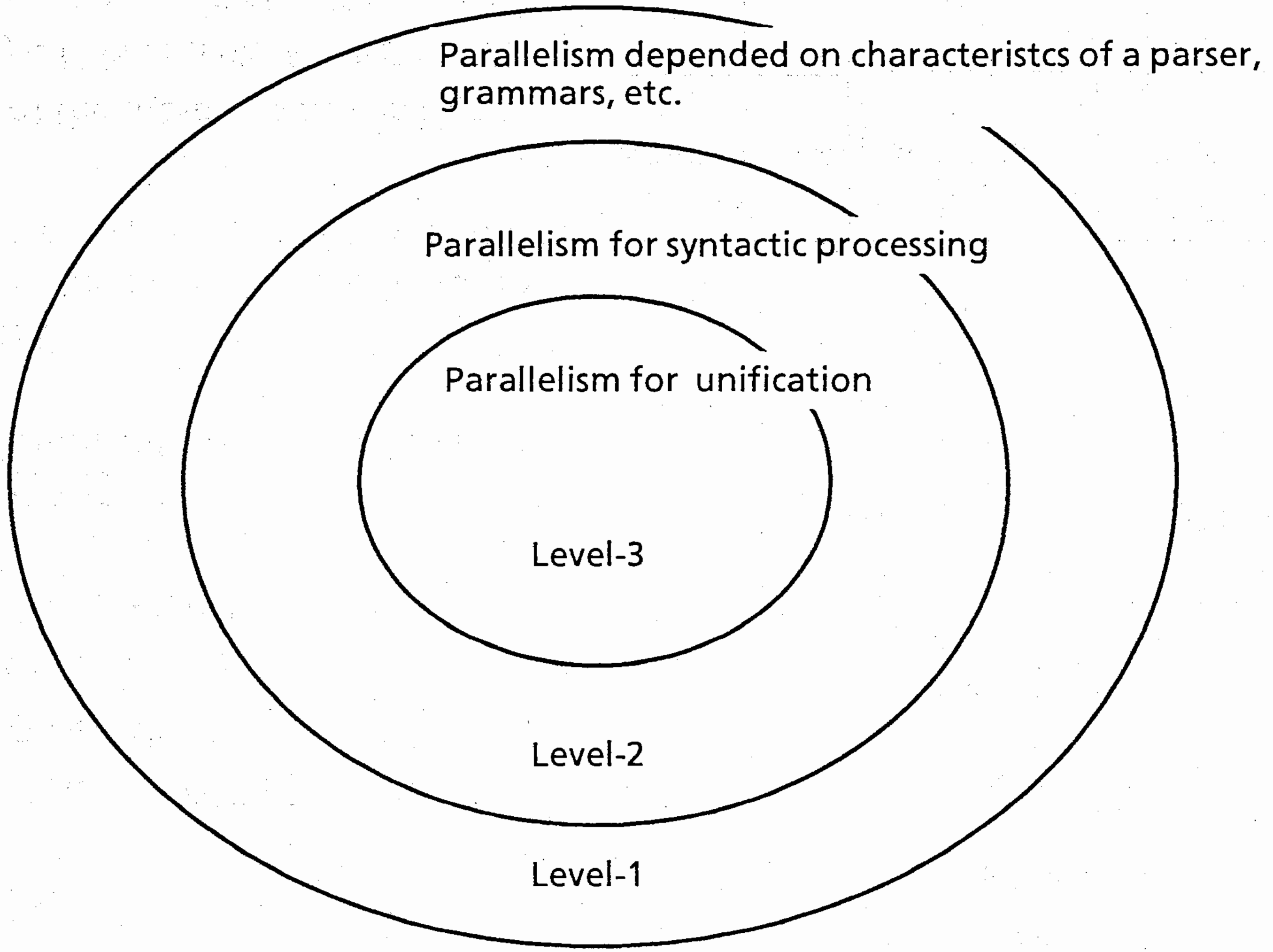
ex. [with the portion of the thesaurus shown right]

(1) MSCA (会議 , 滞在) = 行動 → d (会議 , 滞在) = 2/3

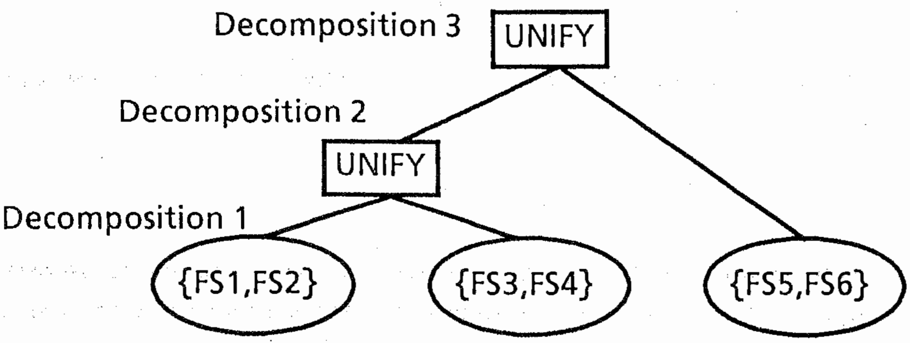
(2) MSCA (到着 , 滞在) = 往来 → d (到着 , 滞在) = 1/3



●Parallelism for UG-Based Parser


Parallelism depended on characteristcs of a parser, grammars, etc.
Parallelism for syntactic processing
Parallelism for unification
Level-3
Level-2
Level-1

● Parallel Processing For Unification Based Parser



Decomposition 3

UNIFY

Decomposition 2

UNIFY

Decomposition 1

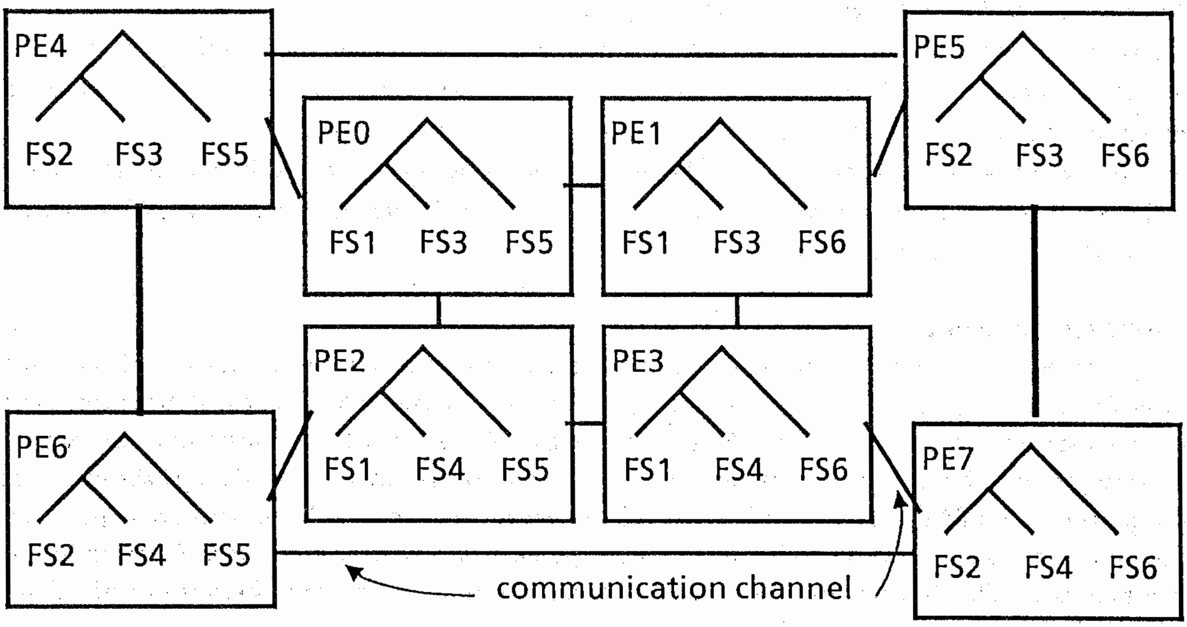
{FS1,FS2}

{FS3,FS4}

{FS5,FS6}

Decompose featuture structure set.

	PE0	PE1	PE2	PE3	PE4	PE5	PE6	PE7
Decomposition 1	FS1	FS1	FS1	FS1	FS2	FS2	FS2	FS2
Decomposition 2	FS3	FS3	FS4	FS4	FS3	FS3	FS4	FS4
Decomposition 3	FS5	FS6	FS5	FS6	FS5	FS6	FS5	FS6



PE4 FS2 FS3 FS5

PE5 FS2 FS3 FS6

PE0 FS1 FS3 FS5

PE1 FS1 FS3 FS6

PE2 FS1 FS4 FS5

PE3 FS1 FS4 FS6

PE6 FS2 FS4 FS5

PE7 FS2 FS4 FS6

communication channel

Spoken Language Processing in SL-TRANS

Tsuyoshi MORIMOTO(†), Kentaro OGURA(†), Kenji KITA(†), Kiyoshi KOGURE(†),
Koji KAKIGAHARA(‡)

(†)ATR Interpreting Telephony Research Laboratories, (‡)Matsushita Electric Industry Co.

Abstract

SL-TRANS is an experimental spoken language translation system from Japanese into English. Spoken language processing is composed of three stages, Bunsetsu speech recognition, Bunsetsu candidate filtering and language analysis. At these stages, linguistic information is used in a stepwise fashion: the Bunsetsu speech recognition module recognizes input speech by using the Bunsetsu syntax, the Bunsetsu filtering module filters out unplausible candidates by using Bunsetsu Kakariuke relationship and the language analysis module selects the unique sentence by using strict sentential syntactico-semantic or heuristic constraints. Experiment results show the final selection rate of sentences is 92% for a specific speaker.

1. Introduction

One of the most important and difficult problems in realizing an automatic interpreting telephony system is how to connect speech recognition and language processing. In,addition to the problems of syntactic or semantic ambiguities in written text processing, another dimensional ambiguity, that is ambiguity of input data itself, arises in spoken language processing. To resolve or remove this ambiguity, some linguistic information should be used.

Several methods using syntactical constraints [1,2,3] or co-occurrence relationship between words such as Japanese Kakariuke relationship[4] have been proposed.

However, to get a correct sentence effectively, it is not sufficient to use only one kind of such information. Moreover, it is necessary to use proper information at appropriate points to avoid an unnecessary increase in processing time to check out incorrect candidates.

In this paper, a new method of spoken language processing, implemented in SL-TRANS that is our experimental speech to speech translation system, is described. Experiment results are also indicated to see the effectiveness of this method.

2. General Schema

Fig. 1 shows the general schema of our method. Input spoken sentences are those which are uttered phrase (Japanese Bunsetsu) by phrase. This restriction is reasonable at present because complete continuous speech recognition leads to other difficulties such as caused by acoustical distortion, etc. In our system, recognition of utterance is performed for each Bunsetsu independently, and the output result is a sequence of candidates corresponding to each Bunsetsu. In this speech recognition stage, Japanese Bunsetsu syntax is used. Japanese Bunsetsu is composed of one independent word followed by several dependent words. This Bunsetsu syntax works well as a recognition constraint because within Bunsetsu composition rules are fairly deterministic.

In the next stage, the Kakariuke relationship, which is a kind of word co-occurrence relationship, is used to filter out unplausible candidates. In general, Japanese, especially spoken Japanese, has many ellipsis and flexibility in Bunsetsu order. It is thus to be expected that this Kakariuke relationship would work well as a sentential constraint. In fact, it could eliminate about 70% of the candidates in our experiment.

In the final stage, strict syntactico-semantic and heuristic analysis over these candidates is performed by the language analyzer and the most plausible sentence is selected.

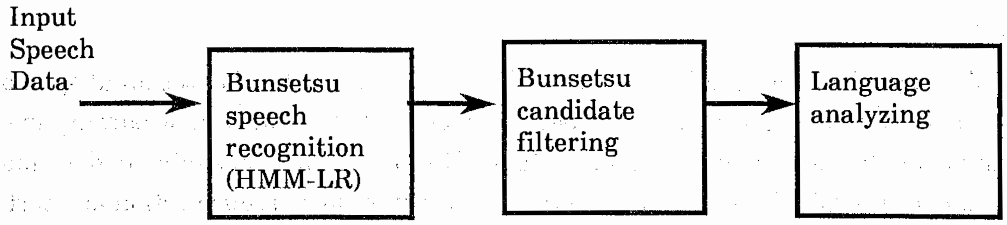
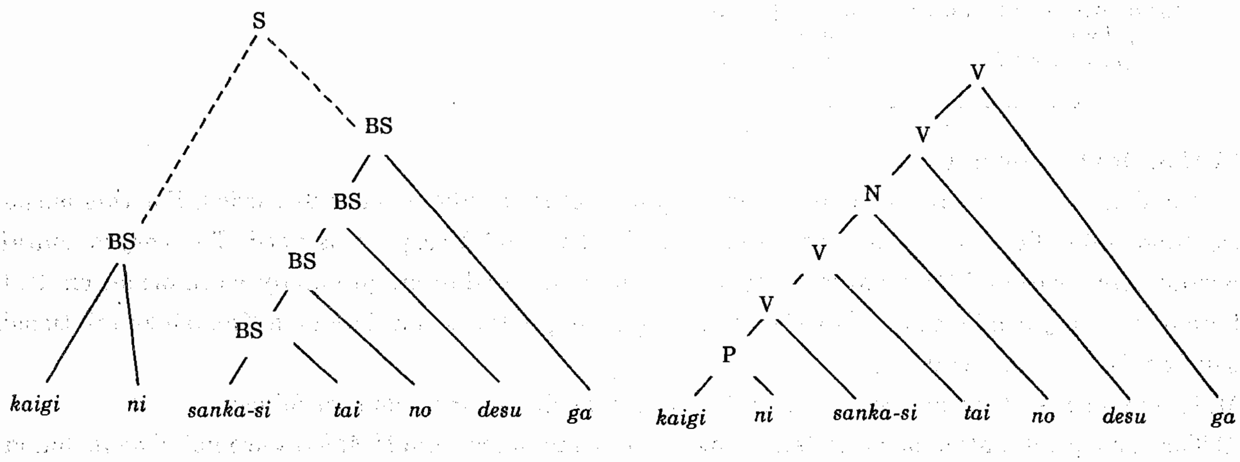


Fig.1 General processing schema

3. Speech Recognition Using Japanese Bunsetsu Syntax

3.1 Speech recognition and parsing

When applying syntax for speech recognition, as with processing texts, grammar definition and a parsing mechanism are needed. Then the problem to be considered is whether these components for speech recognition and for language analysis must be prepared independently or not. In the latter case, they will be used in common for speech recognition and language processing. This method seems to be better than the former one. However, a problem arises from the difference of processing units used in them. In speech recognition, the processing unit is Bunsetsu, thus the grammar should be defined based on Bunsetsu structure. On the other hand, the language analyzer must extract the total meaning of the sentence, so a grammar based on JPSG (Japanese version of HPSG) is desirable[8]. The difference between the two is apparent from the differences between their syntax trees (Fig.2). The crucial problem in the latter method is that a syntax tree for a sentence cannot be obtained constructively from Bunsetsu syntax trees. Accordingly, we adopted the former method.



(a) Syntax tree based on Bunsetsu structure

(b)Syntax tree based on JPSG

Fig.2 Difference between Bunsetsu syntax and JPSG syntax

Another problem to be considered is when the grammar should be applied. The following methods could be used:

(1) First, for all intervals of input signal, try to recognize all the recognition units defined in the system. Then, apply the grammar to the recognized results and select syntactically correct sequences.

(2) Predict the next possible recognition units by using the grammar, then try to recognize these units. The process is continued until the end of the input signal. All the results obtained at the completion of the processing eventually satisfy the syntax.

Method (2) is superior to (1) in both recognition rate and efficiency because in (2) the number of units to be recognized is smaller and information loss due to signal-symbol conversion can be minimized.

From the above consideration, we developed a new speech recognition mechanism called HMM-LR .

3.2 HMM-LR[7]

In HMM-LR, each phoneme is defined as a Hidden Markov Model[5]. The terminal symbols in the grammar are phonemes. The generalized LR parsing method[6] is adopted as a parsing mechanism. Grammar is pre-compiled and converted to an LR table. HMM-LR refers this table to determine which phonemes should be recognized next. According to the grammar rules, reduce action are performed at appropriate points. These actions proceed in parallel for multiple candidates. At each recognition stage, probabilities for the sequences of phonemes are calculated and only those candidates with high probability are kept (i.e, the beam search is performed). At the end of the input data, several candidates which were grammatically accepted and had higher probabilities are output as final candidates.

The details of HMM-LR will be reported in another session of this symposium.

4. Bunsetsu Candidate Filtering Using Japanese Kakariuke Semantic Relationship

4.1 Kakariuke Relationship

Kakariuke is a relationship between two Bunsetsu such as shown below:

(1) a predicate and its case-filler relationship

kaigi-ni *sankasuru*
(conference)(OBJ) *(attend)*

(2) modification relationship to a noun

kaigi-no *sankahi*
(conference)(of) *(registration fee)*

(3) modification relationship to a predicate.

sikyuu *okuru*
(immediately) *(send)*

4.2 Candidate Filtering

Candidates with no Kakariuke relationship with other candidates are discarded. For this purpose, the Kakariuke dictionary made from the analysis of the ATR corpus[9] is used. The corpus contains various conversational texts gathered by simulation and, in addition, pre-analyzed information. In the dictionary, every combination of two Bunsetsu appearing in the corpus having a Kakariuke relationship is defined, as is their frequency.

Matching the results from HMM-LR and the Kakariuke dictionary is done as follows:

(1) The same combination of Bunsetsu as the output candidates from HMM-LR are not always found in the dictionary. Then the expression levels for an each independent word, as described below, are defined:

(a)surface expression: a word exactly as it appeared

(b)standard expression: a standard word having the same meaning in the specific domain

(c)semantic feature expression: semantic feature for a word

All combinations of two candidates from the HMM-LR are checked using (a),(b) or (c) in this order. If several candidates with a Kakariuke relationship are found in the same level, the one having the highest matching score is selected according to equation (1).

$$K(X,Y)=F(X,Y)-w1\times D(X,Y)+w2\times S(Y) \quad \cdots\cdots \quad (1)$$

where

K(X,Y): matching score between Bunsetsu X and Y

F(X,Y): frequency of appearance in the Kakariuke dictionary

D(X,Y): distance between X and Y in the input data
S(Y): speech recognition score
w1,w2: weight

An example is shown below.

Original Sentence;

tourokuyoushi-wa	*sudeni*	*omochi-deshou-ka*
(registration form)(TOP)	*(already)*	*(have)* (POL) (INTR)

(Do you already have a registration form?)

Input (Output from HMM-LR);

tourokusi-ta	*sudeni*	*omoi-mashou-ka*
(register)(PAST)	*(already)*	*(think)*(POL,INT)(INTR)
tourokuyoushi-ga	*itsu-ni*	*omochi-deshou-ka*
(registration form)(SBJ)	*(when)*(ATT)	*(have)* (POL) (INTR)
tourokuyoushi-wa	*sen-ni*	*omochisi-mashou-ka*
(registration form)(TOP)	*(thousand)*(DEG)	*(bring)* (POL,INT) (INTR)

Output;	*tourokuyoushi-wa*	*sudeni* *sen-ni*	*omoti-deshou-ka*

5. Sentence Preference during Language Analysis

If there are several sentential candidates resulting from the Kakariuke filtering, their syntactico-semantical propriety is checked by the language analyzer. Nevertheless, if their are still several candidates, the sentence which has the lowest penalty calculated by equation (2) is selected.

$$P(X)=a1\times Nt(X)+a2\times Nu(X)-a3\times S(X) \quad \cdots\cdot \quad (2)$$

where

Nt:number of nodes of syntax tree
Nu:number of unfilled obligatory elements
S(X):speech recognition score
a1,a2.a3:weight

Nt and Nu reflect the heuristics that a simpler sentence is more plausible.

An example is indicated below:

Input;

(1-1) *soredewa*	(2-1) *saremasu*
then	*( ) do ( )* or *( ) is done*
(1-2) *sureba*	(2-2) *sitsureisimasu*
if( ) do ( )	*goodbye*

Evaluation of Penalty;

Among all, the combination of (1-1) and (2-2) has the fewest nodes and unfilled obligatory elements. Thus, this is selected.

Output;

soredewa   *sitsureisimasu*

6. Experiment Results in SL-TRANS

SL-TRANS (Spoken Language Translation System), can recognize input Japanese speech, translate it into English and output synthesized English voice. Fig.3 shows the configuration of this system. As a machine translation system, NADINE [10] is used with a functional extension of the analyzer as described above, and the DECtalk is used as an English speech synthesizer.

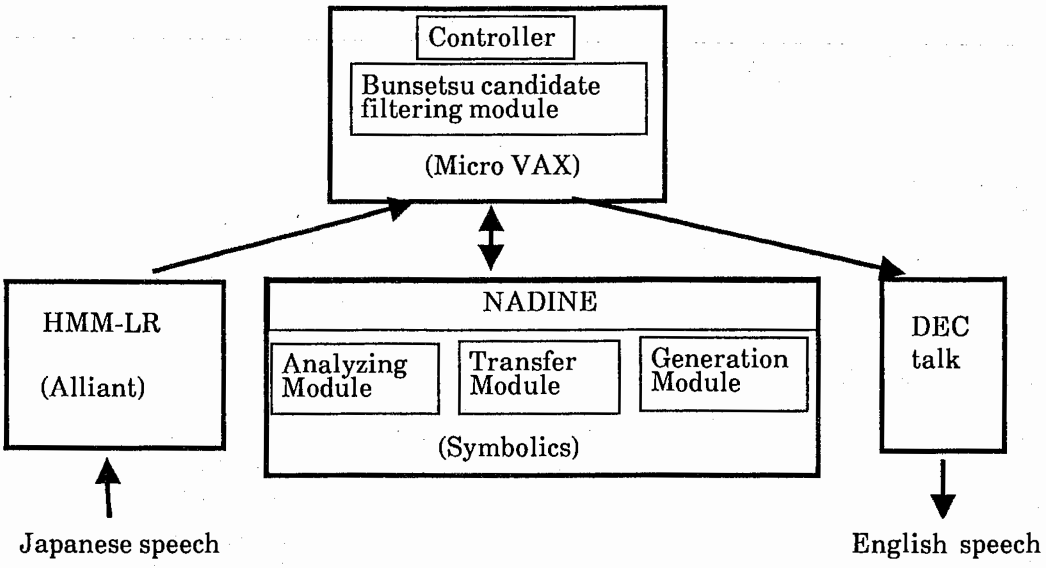


Fig.3 Configuration of SL-TRANS

Experiments have been performed for dialogue containing 37 sentences about a secretarial service for an international conference. Results are summarized in Table 3. Bunsetsu recognition rate of HMM-LR is 87% for the first rank. This means that the sentence recognition rate of HMM-LR is $0.87^{2.2} = 0.74$ (i.e. 74%) for the first rank. To compensate for this rate, HMM-LR outputs the top 5 candidates. On the other hand, this increases the possible number of sentential candidates to $4.6^{2.2} = 28.7$ for one sentence. However, this number was reduced to $1.5^{2.2} = 2.4$ by the Bunsetsu filtering. Finally, 34 out of 37 sentences, that is 92%, were selected correctly after language analysis.

Table 3. Experiment results of SL-TRANS

Input	Specific speaker Number of sentences: 37 Number of Bunsetsu: 83 Average number of Bunsetsu/ sentence: 2.2
HMM-LR	Bunsetsu recognition rate 87% for the 1st rank 96% for the top 5 ranks Average number of output candidates/ Bunsetsu: 4.6 →number of sentential candidates: $4.6^{2.2} = 28.7$
Bunsetsu filtering	Average number of selected candidates/ Bunsetsu: 1.5 →number of sentential candidates: $1.5^{2.2} = 2.4$
Language processing (System total)	Number of sentences selected correctly: 34 Sentence selection rate: 92%

7. Concluding Remarks

In this paper, a new method for connecting speech recognition and language processing is proposed. The experiment results show this method is effective.

We are currently working on enhancing the capability of this method by introducing statistical characteristics of grammar or heuristics of sentence structure. Utilization of higher level information such as dialogue structure should also be investigated.

References

[1] Nakagawa,S.: Spoken Sentence Recognition by Time-Synchronous Parsing Algorithm of Context-Free Grammar, ICASSP 87, 1987

[2] Ney,H.: Dynamic Programming Speech Recognition Using a Context-Free Grammar, ICASSP-87, 1987

[3] Chow,Y., Roukos,S.: Speech Understanding Using Unification Grammar, ICASSP-89, 1989

[4] Ozeki,K.: A Multiple-Stage Decision Algorithm for Optimum Bunsetsu Sequence Selection Based on the Degree of Multi-Bunsetsu Kakariuke-Dependency, Trans. IEICE, Vol. J71-D, No.4, 1988 (In Japanese)

[5] Levinson,S., Rabiner,L., Sondhi,M.: An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition, Bell Syst. Tech. J.,Vol.62,No.4, pp1035-1074, 1983

[6] Tomita,M.: Efficient Parsing for Natural Language,: A Fast Algorithm for Practical Systems, Kluwer Academic Publishers,1986

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

[8] Gunji,T.: Japanese Phrase Structure Grammar,Reidel, 1987

[9] Morimoto,T., Ogura,K., Iida,J.: Constructing Linguistic Database for Automatic Telephone Interpreting Research, IPSJ Spring Meeting, 1988 (In Japanese)

[10] Iida,J., et al.: An Experimental Spoken Natural Dialogue Translation System Using a Lexicon-Driven Grammar, EUROSPEECH-89,1989

Spoken Language Processing in SL-TRANS

Tsuyoshi MORIMOTO
Kentaro OGURA
Kenji KITA
Kiyoshi KOGURE
Kji KAKIGAHARA

Problems in Connecting Speech Recognition(SR) and Language Processing(LP)

(1)SR using only acoustical characteristics is not accurate

(2) How should linguistic information be used to get a correct sentence accurately and effectively ?

- What kind of linguistic information should be used ?

  syntax

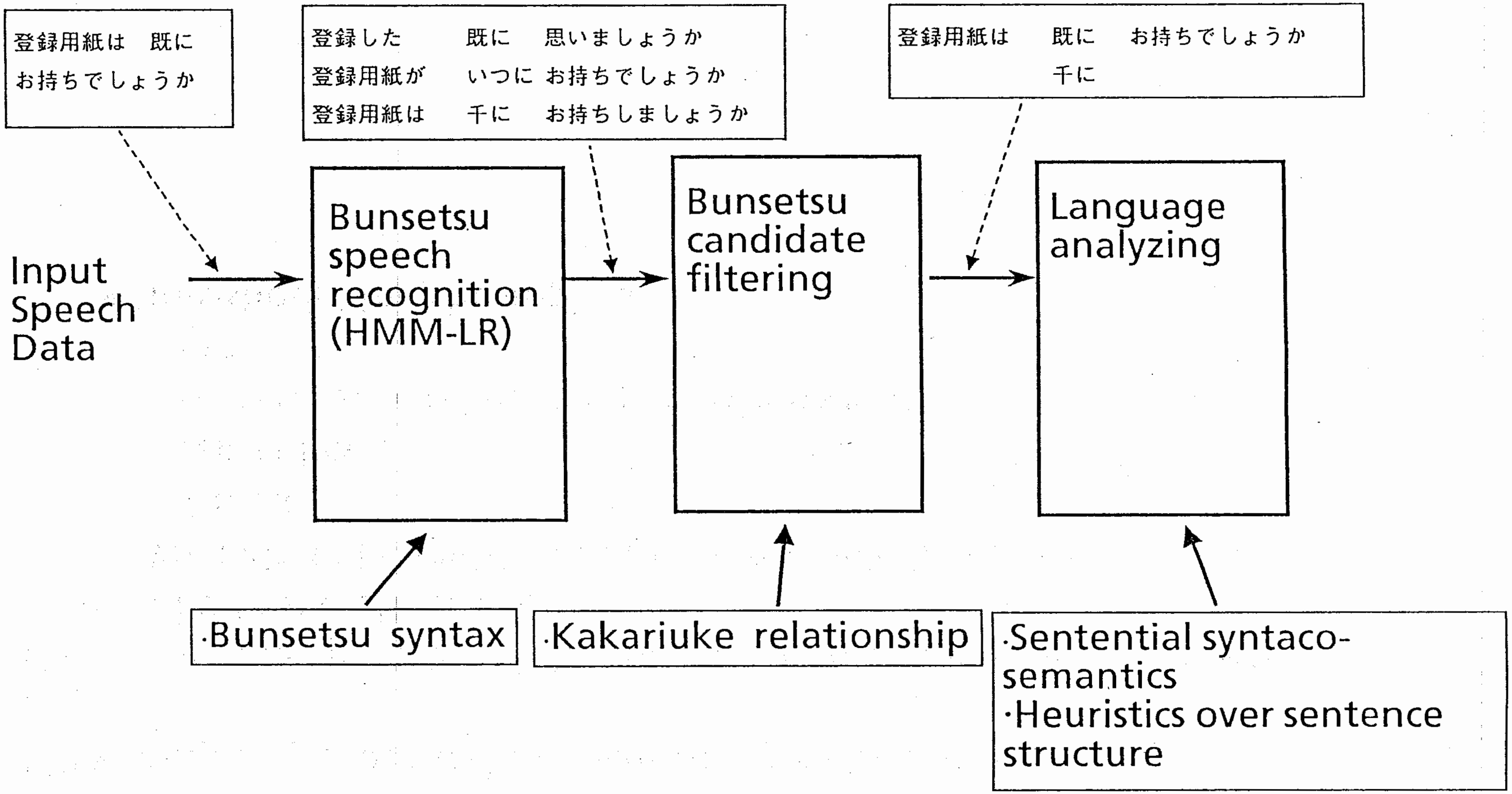
  semantics

  co-occurence relationship between words

  etc.

- How should they be applied ?

OHP2



登録用紙は　既に
お持ちでしょうか

Input Speech Data

Bunsetsu speech recognition (HMM-LR)

登録した	既に	思いましょうか
登録用紙が	いつに	お持ちでしょうか
登録用紙は	千に	お持ちしましょうか

Bunsetsu candidate filtering

登録用紙は	既に	お持ちでしょうか
	千に	

Language analyzing

·Bunsetsu syntax

·Kakariuke relationship

·Sentential syntaco-semantics
·Heuristics over sentence structure

General schema

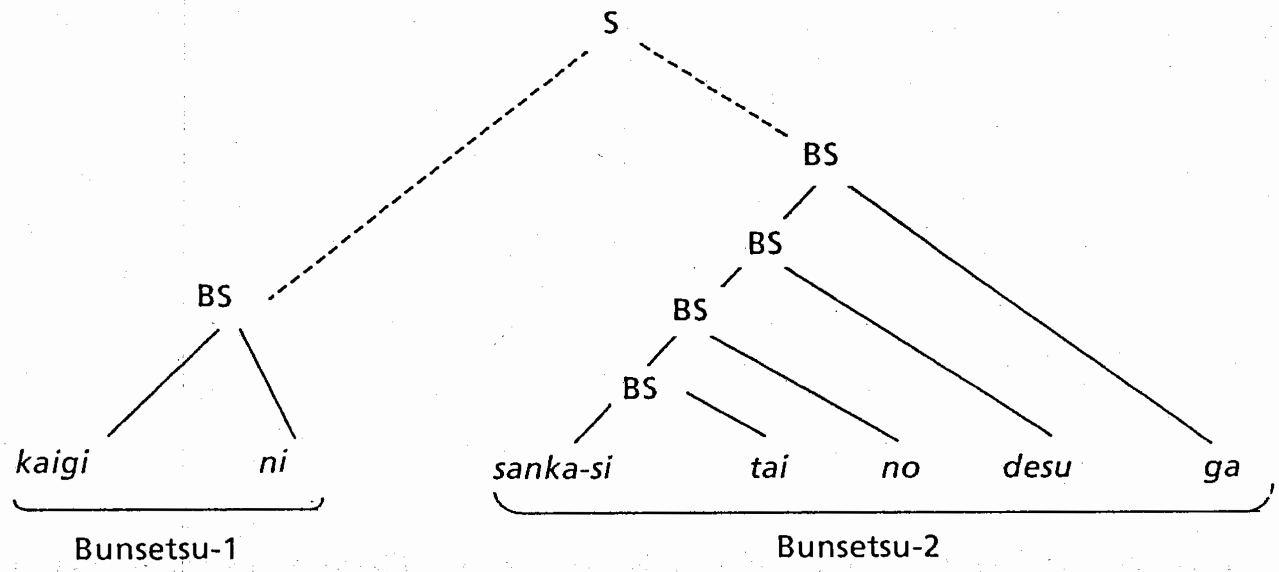
Grammar and Parsing mechanism for SR and LP

(1) Prepare both independently

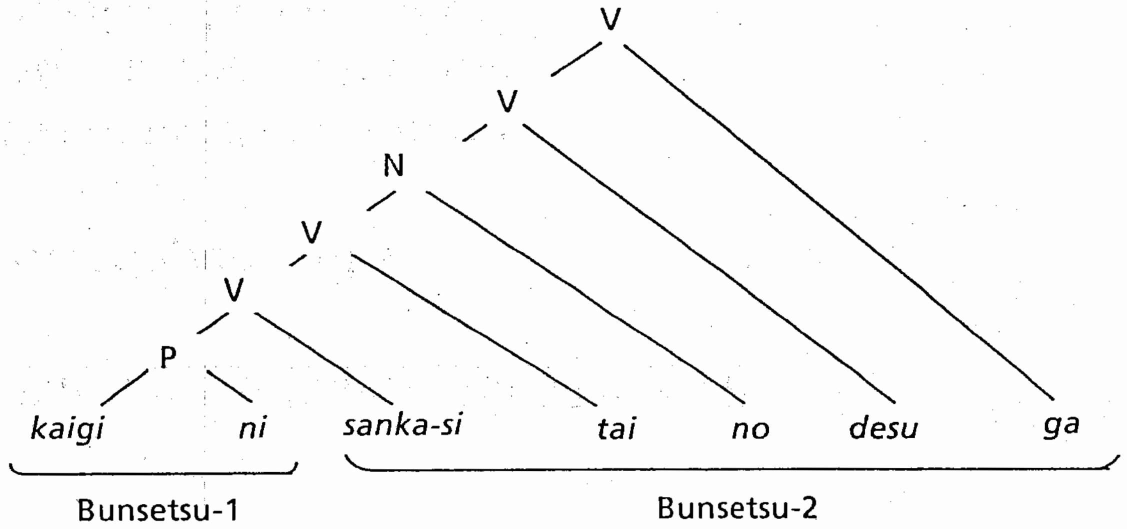
(2) One or both are used in common

- Difference in processing unit
- Difference of their purpose
- Modularity

between the speech recognition language processing modules



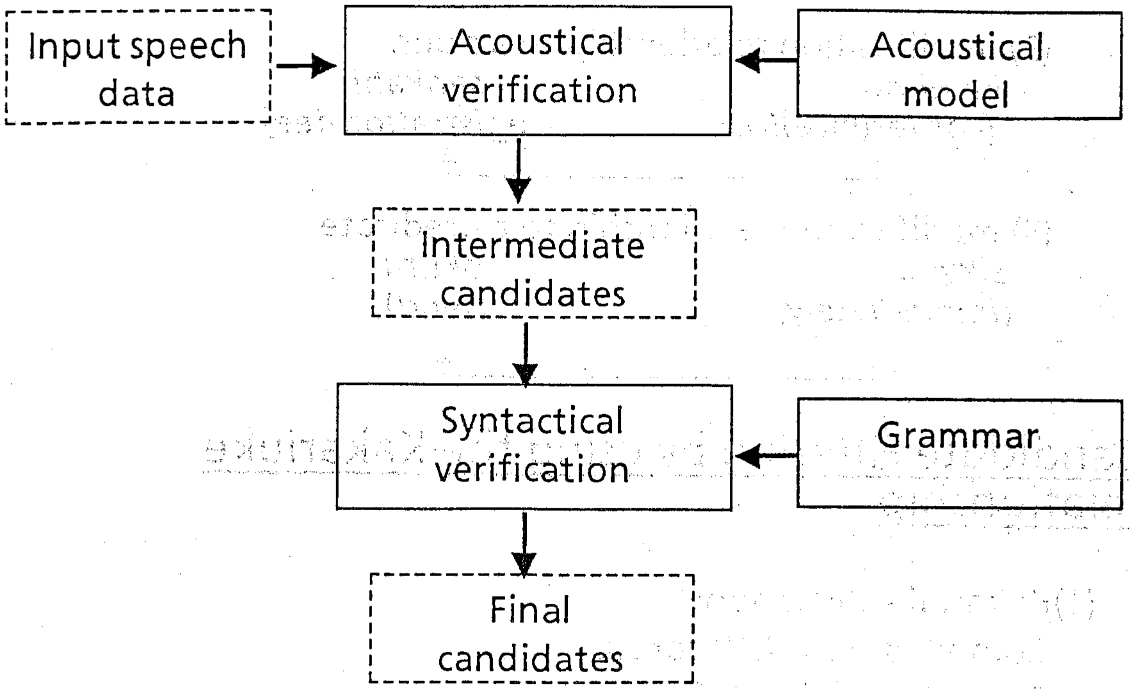
(a) Syntax tree based on bunsetsu structure



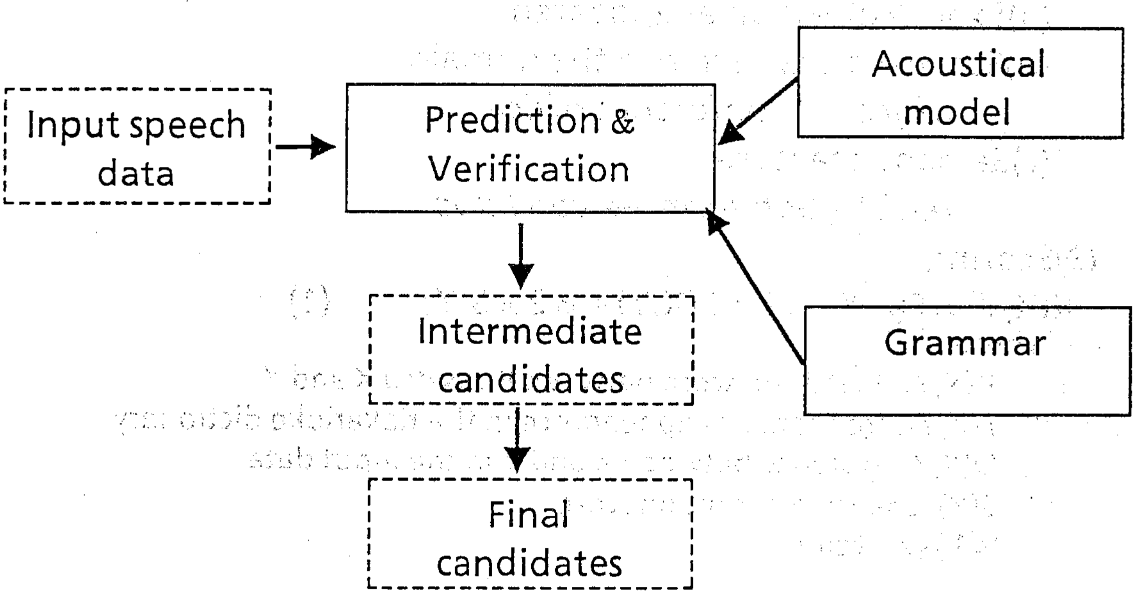
(b)Syntax tree based on JPSG

When should the grammar be applied ?

(1)After acoustical verification



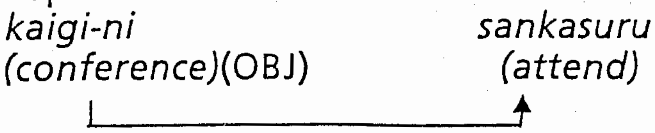
(2)Concurrent with acoustical verification



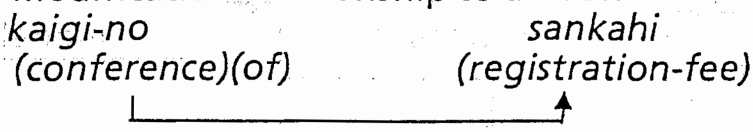
OHP6

Kakariuke relationship

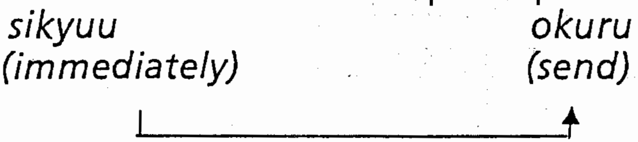
(1) A predicate and its case-filler relationship



(2) modification relationship to a noun



(3) modification relationship to a predicate.



Candidate Filtering by using the Kakariuke relationship

(1)Kakariuke dictionary

Compiled from ATR corpus

Frequency is also defined

(2)Three level matching

(a)Exact expression as appeared

(b)Standard expression in the domain

office → conference office

(c)Semantic feature

registration form → con / doc

(3)Scoring

$$K(X,Y) = F(X,Y) - w1 \times D(X,Y) + w2 \times S(Y) \quad \cdots\cdot \quad (1)$$

where

K(X,Y): matching score between Bunsetsu X and Y

F(X,Y): frequency of appearance in the Kakariuke dictionary

D(X,Y): distance between X and Y in the input data

S(Y): speech recognition score

w1,w2: weight

(4)Including dependent words of modifier

OHP7

Example

Original Sentence;

tourokuyoushi-wa	sudeni	omochi-deshou-ka
(*registration-form*)(TOP)	(*already*)	(*have*) (POL)(INTR)

(*Do you already have a registration form?*)

Input (Output from HMM-LR);

tourokusi-ta	sudeni	omoi-mashou-ka
(*register*)(PAST)	(*already*)	(*think*)(POL,INT)(INTR)
tourokuyoushi-ga	itsu-ni	omochi-deshou-ka
(*registration-form*)(SBJ)	(*when*)(ATT)	(*have*) (POL) (INTR)
tourokuyoushi-wa	sen-ni	omochisi-mashou-ka
(*registration-form*)(TOP)	(*thousand*)(DEG)	(*bring*)(POL,INT)(INTR)

Output;

tourokuyoushi-wa	sudeni sen-ni	omoti-deshou-ka

Sentence Selection During Language Analyzing

(1) Strict syntactico-semantic check based on JPSG

(2) Evaluating the complexity of a sentence (Penalty evaluation)

$$P(X) = a1 \times Nt(X) + a2 \times Nu(X) - a3 \times S(X)$$

where
Nt:number of nodes of syntax tree
Nu:number of unfilled obligatory elements
S(X):speech recognition score
a1,a2,a3:weight

Example

Input;

(1-1) soredewa
then

(1-2) saremasu
( ) do ( ) / ( ) is done

(2-1) sureba
if ( ) do ( )

(2-2) sitsureisimasu
goodbye

Evaluation of Penalty;

[1]Number of nodes

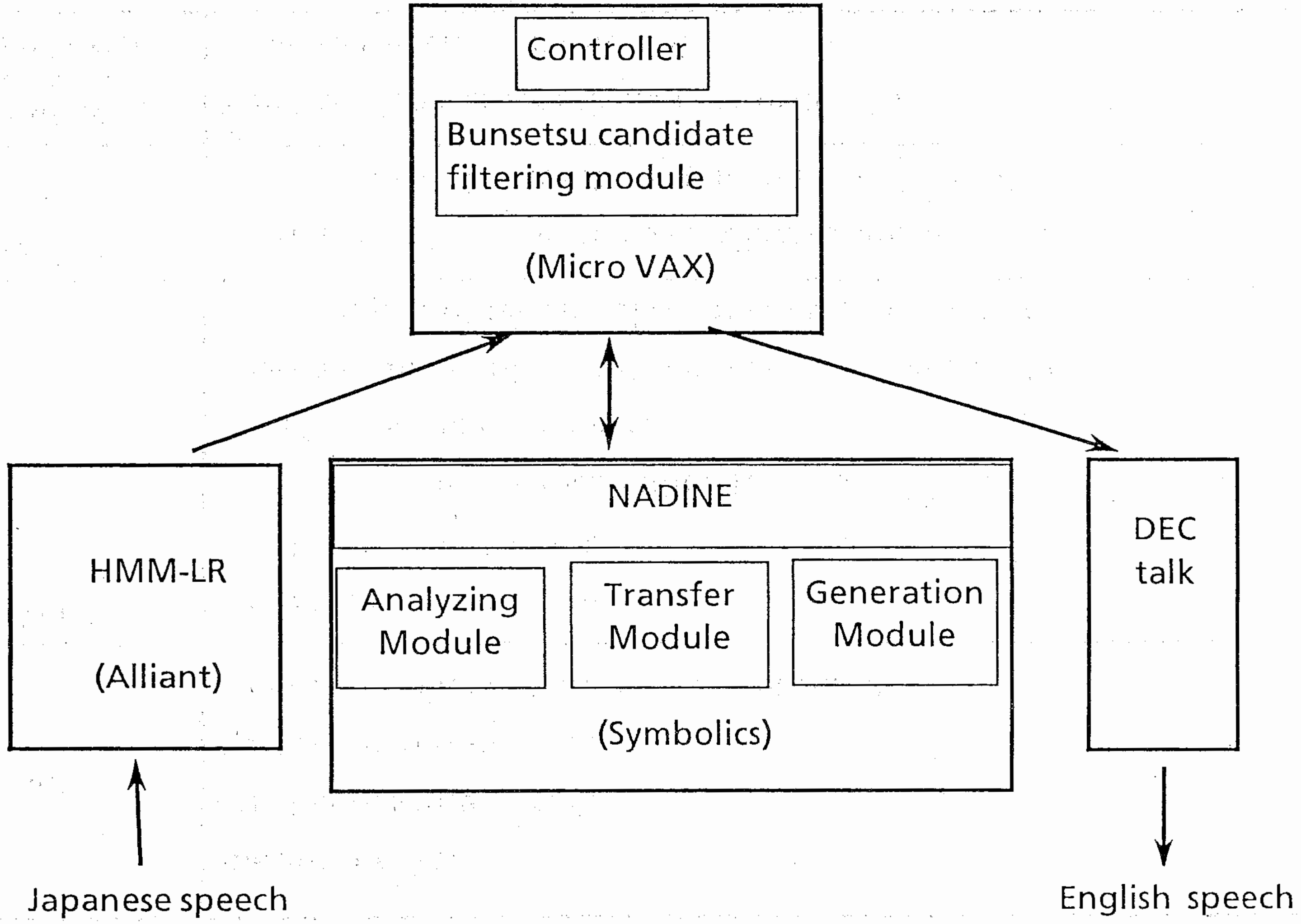
Among all combinations of input, the combination of (1-1) and (2-2) has the fewest nodes.

[2]Number of unfilled obligatory elements

In (1-2), there are two unfilled obligatory elements, and one or two in (2-1). Hence, the combination of (1-1) and (2-2) is more probable

→ the combination of (1-1) and (2-2) is selected

OHP9



Controller

Bunsetsu candidate filtering module

(Micro VAX)

HMM-LR

(Alliant)

NADINE

Analyzing Module

Transfer Module

Generation Module

(Symbolics)

DEC talk

Japanese speech

English speech

Configuration of SL-TRANS

OHP10

Experiment results of SL-TRANS

Input	Specific speaker Number of sentences: 37 Number of Bunsetsu: 83 Average number of Bunsetsu/ sentence: 2.2
HMM-LR	Bunsetsu recognition rate 87% for the 1st rank Sentence recognition rate: $0.87^{2.2} = 0.74$ 96% for the top 5 ranks Average number of output candidates/ Bunsetsu: 4.6 → number of sentential candidates: $4.6^{2.2} = 28.7$
Bunsetsu filtering	Average number of selected candidates/ Bunsetsu: 1.5 → number of sentential candidates: $1.5^{2.2} = 2.4$
Language processing (System Total)	Number of sentences selected correctly: 34 Sentence selection rate: 92%

Conclusion

(1) A new method for connecting speech recognition and language processing is proposed

(2) The experiment result shows the effectiveness of this method

Current work / Future plans

(1) Introduction of statistical characteristics of grammar or heuristics over sentence structure

(2) Utilization of higher level information such as dialogue structure