Limited Distribution

TR-I-0261

Research Activities of the Speech Processing Department

January through December, 1991

ATR Interpreting Telephony Research Laboratories

May, 1992

ATR Interpreting Telephony Research Laboratories

This volume is the fifth in a series of the complete collection of technical papers from the Speech Processing Department, ATR Interpreting Telephony Research Laboratories:

1. Volume I: from April 1987 through December 1988 (TR-I-0010)
2. Volume II: from November 1987 through December 1988 (TR-I-0065)
3. Volume III: from January through October 1989 (TR-I-0115)
4. Volume IV: from November 1989 through December 1990 (TR-I-0230)
5. Volume V: from January through December 1991 (TR-I-0261)

The research areas of the Speech Processing Department include:

1. Large Vocabulary Continuous Speech Recognition
   (a) Hidden Markov Models
   (b) Neural Network Approaches
   (c) Feature-Based Approaches
2. Speaker Adaptation and Noise-Robust Speech Recognition
3. Language Source Modeling
4. Speech Synthesis by Rule
5. Voice Conversion
6. Speech Database

For additional information, contact:

Shigeki Sagayama, Head, Speech Processing Department
ATR Interpreting Telephony Research Laboratories
2-2, Hikari-dai, Seika-cho, Souraku-gun, Kyoto 619-02, JAPAN *
Telephone: +81 7749 5 1311,
Telefax: +81 7749 5 1308
E-mail: saga@atr-la.atr.co.jp

(* The older address "Sampei-dani, Inui-dani, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN" is also valid.)

Research Activities of the Speech Processing Department, January through December, 1991

Contents

1 Introduction

At ATR Interpreting Telephony Research Laboratories, speech research is being intensively pursued to realize automatic telephone interpretation that allows speech communication between telephone users speaking different languages. Automatic telephone interpretation or *speech translation*, to use more general terminology is one of the technologies most eagerly awaited by people all over the world.

The utterance of one party is to be translated into another language, which is heard by the person on the other end. The basic constituent technologies are speech recognition, language translation, and speech synthesis, which are not only essential for the realization of speech translation, but also widely applicable in various other areas. Although this concept of automatic telephone interpretation is rather new and research is still in its early stages, it is rapidly attracting the interest of researchers and research organizations throughout the world.

2 Automatic Telephone Interpretation Project

2.1 History and General Concept

In 1983, the first experimental demonstration of on-line speech translation was given at TELECOM-83 by NEC Research Laboratories. In 1987, speech translation between English and French was conducted at British Telecom Research Laboratories. That system was based on a set of more than 400 common business phrases. At Carnegie-Mellon University (CMU), a speech translation system with spoken input was developed in 1988 in a simple doctor-patient conversation domain.

In Japan, ATR Interpreting Telephony Research Laboratories, a subsidiary of Advanced Telecommunications Research International, was established in 1986 to initiate basic research for telephone interpretation. Stimulated by these earlier activities toward speech translation, a considerable number of research organizations became interested in this research area. Some of them have started active research in this particular field. In the United States, Carnegie-Mellon University is working on English-Japanese speech translation. AT&T Bell Laboratories is making an English-Spanish speech translation experiment with a relatively small vocabulary. In Germany, a new research project named "Verbmobil" has been started aiming at language translation with speech input. Karlsruhe University has begun a speech translation project between German and English. Korea Telecom also have a 10-year research project on automatic speech interpretation; its short-term target is Korean-Japanese speech translation. Mutual cooperation and collaborations among these research groups are gradually starting.

The basic components of an automatic telephone interpretation ("speech translation", in more general terminology) system are speech recognition, language translation, and speech synthesis. In realistic applications, speaker adaptation for input and speaker conversion for output are also required.

A simple combination of these elements of current technology, however, does not provide a satisfactory speech translation system. What is required of these components is somewhat different from how they are used in conventional applications. The major differences are as follows:

- High performance both in speech recognition and translation is required. Unlike conventional language translation, there cannot be any "pre/post-editing", since both input and output are speech.

- Real-time operation is a strong requirement.
- Spoken language is quite different from written language. Although, in spoken language, sentences are generally shorter and their structures are not particularly complicated, spoken dialogues include elliptic and anaphoric expressions. They may also include many syntactically ill-formed expressions.

Thus, speech translation is a new area that requires significant innovations in these constituent technologies.

Although the ultimate goal of the telephone interpretation system is universal dialogue in an unlimited domain, the immediate goal should be more feasible, such as a system which is limited to specific, task-oriented areas. ATR Interpreting Telephony Research Laboratories has selected an international conference registration task as a constrained task domain where the dialogue will be goal-directed and the expected vocabulary limited to approximately 1,500 words.

2.2 ATR's Experimental Speech Translation System

Figure 1 outlines the current experimental Japanese-English speech translation system at ATR Interpreting Telephony Research Laboratories. The input sentence speech uttered phrase-by-phrase is recognized with speaker adaptation capability under linguistic constraints and converted into several word sequence candidates. The number of candidates is narrowed down by dependency analysis and a final choice is determined by linguistic analysis. This is followed by language translation consisting of linguistic analysis, Japanese-to-English language transfer, and English sentence generation. The final stage is to convert English sentences into speech with the original speaker's characteristics. This stage is, however, not yet integrated into the system. The other direction, English-Japanese conversion, can be considered in the same way. As it seems inefficient to split our research efforts into Japanese and English aspects, our research topics mainly cover Japanese speech recognition, Japanese-English translation, and Japanese speech synthesis. Basically, the same approach can be applied to the English-Japanese speech translation.

3 Speech Research at ATR Interpreting Telephony Research Laboratories

3.1 Research Project

ATR Interpreting Telephony Research Laboratories (Akira Kurematsu, president) consist of three departments: Natural Language Understanding Department, Knowledge and Database Department, and Speech Processing Department. The laboratory was founded in April 1986 with the support of the Japan Key Technology Center, ATR International, NTT, KDD, NHK, and other Japanese enterprises.

The main target of our institute is fundamental research into speech and language processing and the integration of speech and language processing technologies to demonstrate the feasibility of an automatic telephone interpretation system. As of 1991, the laboratory had a research budget of 2.6 billion yen.

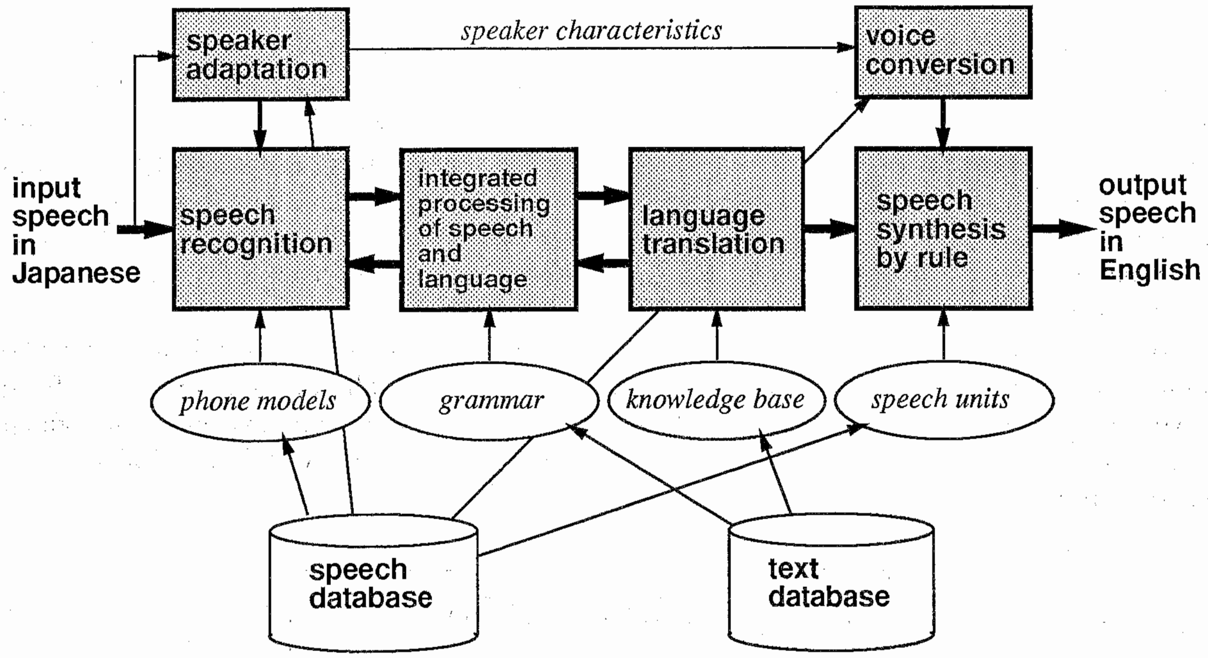


Figure 1: Conceptual diagram of SL-TRANS: an experimental speech translation system (Japanese-English)

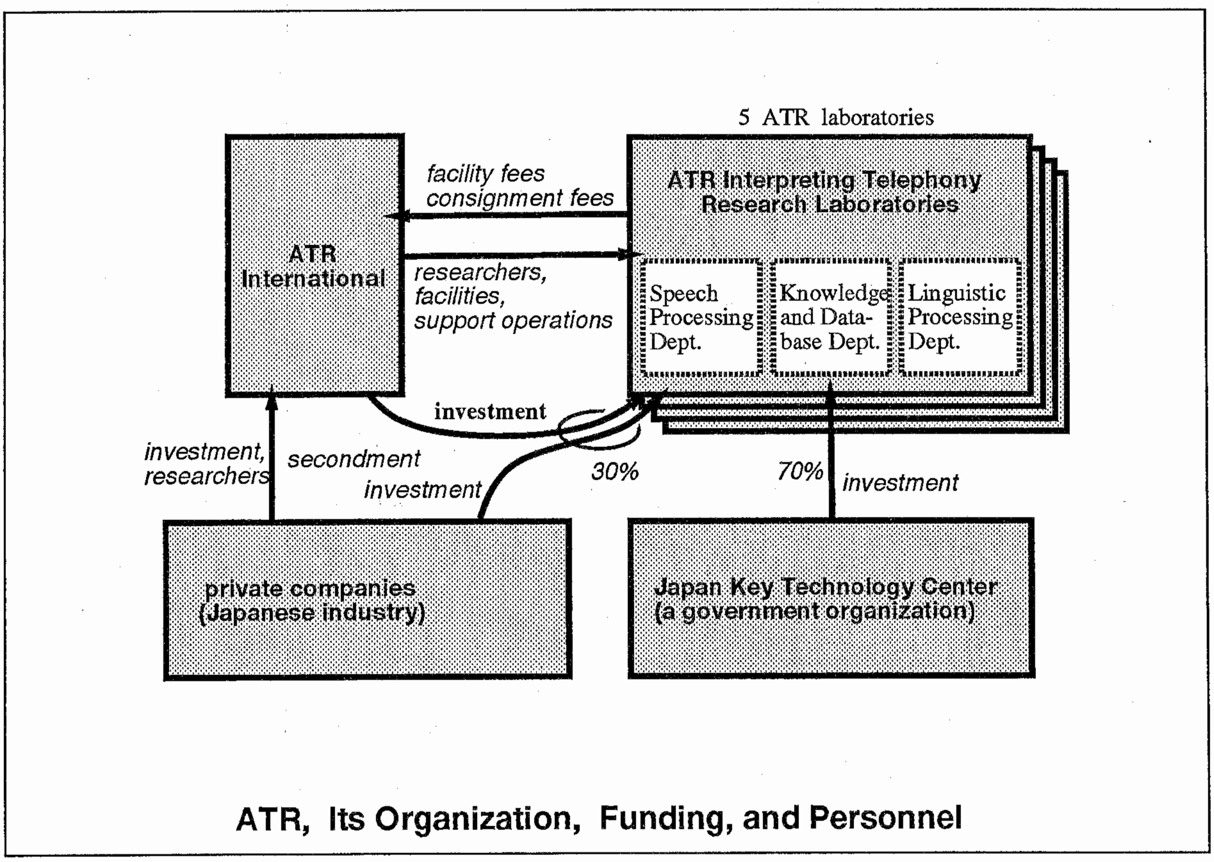


Figure 2: Funding and personnel of ATR Laboratories

3.2 Location

ATR Interpreting Telephony Research Laboratories is located in Kyoto Prefecture, although its geographical location is much closer to Nara than to Kyoto city. Nara was the capital of Japan more than 1,200 years ago. Promoted by Japanese government and industry, 12 separate areas in this district have been designated as Kansai Science and Technology Cities. They are expected to rapidly grow into Japan's new center of major scientific and technological research and education activities. ATR (including ATR International and its five subsidiary laboratories) was one of the earliest organizations to settle in this area.

ATR's transportation situation is not excellent at the moment. Figure 3 shows how to grt to ATR Interpreting Telephony Research Laboratories. The approximate time from Kyoto to Takanohara station of the Kintetsu Railway is 40 minutes by express train, which leaves Kyoto every 15 minutes. ATR shuttle buses or taxis are available at Takanohara station and take 10 minutes to reach ATR.

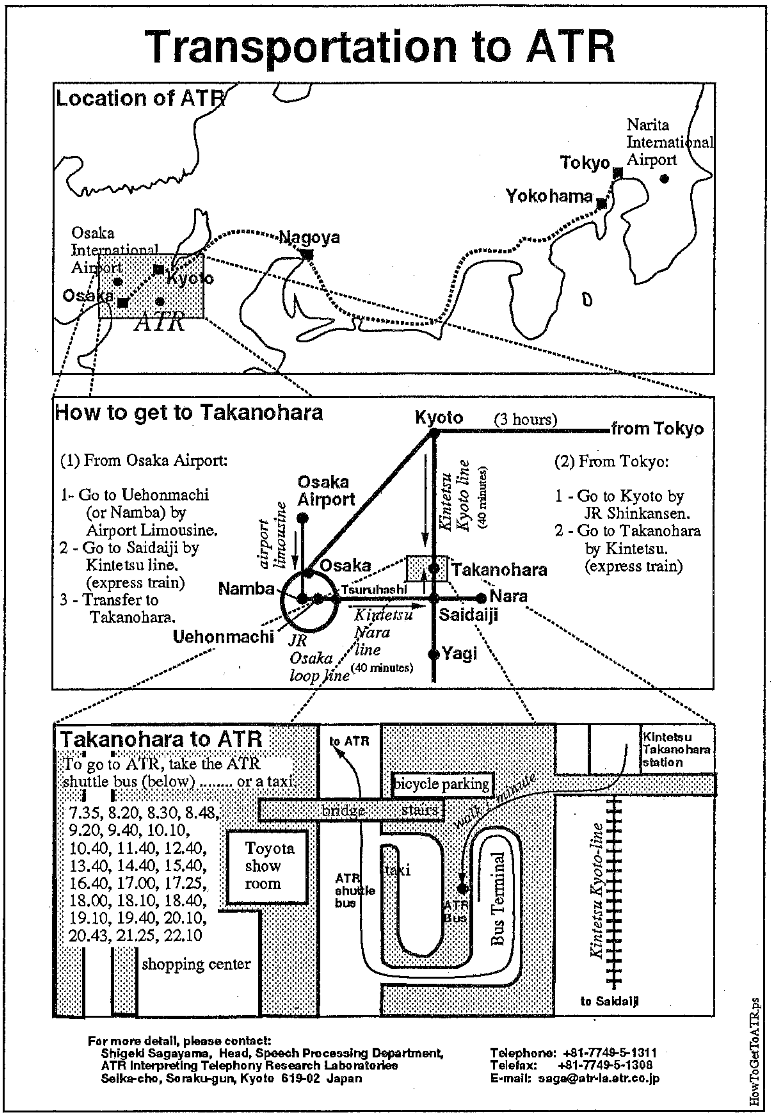
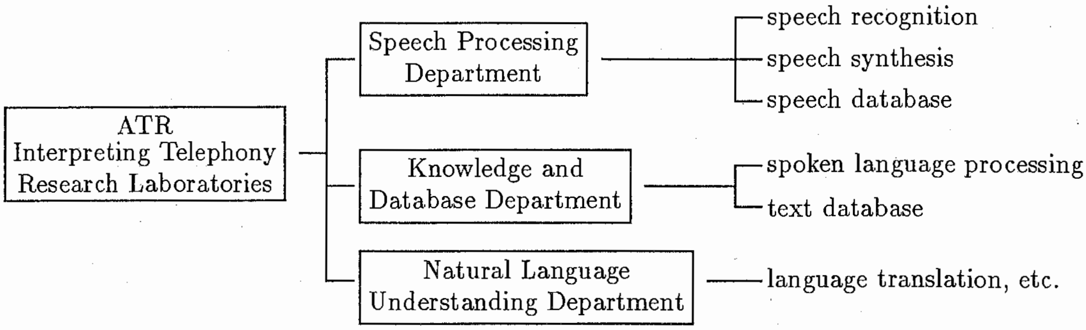


Figure 3: How to get to ATR Laboratories

3.3 Organization

ATR Interpreting Telephony Research Laboratories consist of three departments: Speech Processing Department (Shigeki Sagayama, Head), Knowledge and Database Department (Tsuyoshi Morimoto, Head), and Natural Language Understanding Department (Hitoshi Iida, Head).

An outline of the organization is shown below.



As of 1991, the whole laboratory had a research staff of about 40, and 235 academic publications, including 70 papers at international conferences.

3.4 Research Staff

The research staff is mainly composed of members temporarily seconded from research institutes, laboratories, and industrial companies that support ATR, and visiting/invited researchers.

Currently (as of May, 1992), the Speech Processing Department has 23 members including a department head, two supervisors, 15 researchers, four visiting/invited researchers, an engineer, and some students. Tables 1 and 2 shows former and current members of the research staff at the Speech Processing Department, including visiting/invited researchers. Table 3 shows the list of gradutate students who worked or are currently working as their internships at the department.

Current members can be reached through E-mail account names followed by '@atr-la.atr.co.jp'. (For example, Sagayama's E-mail address is 'saga@atr-la.atr.co.jp'.)

Table 1: Research Staff of the Speech Processing Department

name & title	Email	position	period
Akira Kurematsu, Dr	kurematu	President	4/1986 –
Kiyohiro Shikano, Dr		Department Head	6/1986 – 1/1990
Shigeki Sagayama	saga	Department Head	2/1990 –
Yoshinori Sagisaka, Dr	sagisaka	Superviser	4/1986 –
Hisao Kuwabara, Dr		Supervisor	10/1986 – 7/1989
Masahide Sugiyama, Dr	sugi	Supervisor	2/1990 –
Tetsuo Umeda		Supervisor	7/1989 – 6/1990
Takeshi Kawabata, Dr		Senior Researcher	9/1986 – 2/1990
Shin-ichi Tamura		Senior Researcher	9/1986 – 2/1990
Hidefumi Sawai, Dr		Senior Researcher	4/1988 – 3/1991

Table 2: Research Staff of the Speech Processing Department (cont'd)

name & title	Email	position	period
Kazuya Takeda		Researcher	8/1986 – 2/1990
Satoshi Nakamura		Researcher	9/1986 – 8/1989
Masanori Miyatake		Researcher	9/1986 – 3/1989
Kaichiro Hatazaki		Researcher	12/1986 – 3/1989
Toshiyuki Hanazawa		Researcher	3/1987 –2/1990
Katsuteru Maruyama		Researcher	3/1987 – 2/1990
Masanobu Abe		Researcher	4/1987 – 2/1991
Katsuo Abe		Researcher	3/1987 – 2/1990
Masami Nakamura		Researcher	9/1987 – 8/1990
Yasuhiro Komori		Researcher	9/1988 – 2/1992
Hiroaki Hattori	hiroaki	Researcher	5/1989 – 4/1992
Kazumi Ohkura	ohkura	Researcher	9/1989 –
Nobuyoshi Kaiki	kaki	Researcher	11/1989 –
Jun-ichi Takami	jun	Researcher	11/1989 –
Kazuki Katagishi, Dr	katagisi	Researcher	2/1990 –
Akito Nagai	nagai	Researcher	3/1990 –
Keiji Fukuzawa	fukuzawa	Researcher	4/1990 –
Katsuhiko Mimura	mimura	Researcher	5/1990 –
Shingo Fujiwara	fujiwara	Researcher	5/1990 –
Naoto Iwahashi	iwahashi	Researcher	10/1990 –
Jin-ichi Murakami	murakami	Researcher	3/1991 –
Yoshinaga Kato	kato	Researcher	4/1991 –
Kouichi Yamaguchi	yamaguti	Researcher	5/1991 –
Yasunaga Miyazawa	miyazawa	Researcher	7/1991 –
Tetsuo Kosaka	kosaka	Researcher	9/1991 –
Ryosuke Isotani	isotani	Researcher	4/1992 –
Alex Waibel, Dr		Invited Researcher	5/1987 – 8/1988, 6 – 9/1989
Hiroaki Saito, Dr		Visiting Researcher	2/1988 – 7/1989
William Poser, Dr		Invited Researcher	9/1988 – 2/1989
Alain de Cheveigné, Dr		Invited Researcher	2/1989 – 2/1990
Dieter Huber, Dr		Invited Researcher	3 – 6/1990
David Rainton, Dr	rainton	Visiting Researcher	4/1990 –
Wilhelm N. Campbell	nick	Visiting Researcher	8/1990 –
Harald Singer	singer	Visiting Researcher	3/1991 –
Helmut Lucke, Dr	lucke	Visiting Researcher	4/1992 –
Kouichi Murayama		Engineer	9/1987 – 8/1988
Takaharu Tanaka		Engineer	9/1988 – 8/1989
Kouji Kitagaito		Engineer	9/1989 – 8/1990
Shinobu Araki		Engineer	9/1990 – 8/1991
Yoshinori Ono	ono	Engineer	9/1991 –

Table 3: Research Staff (students and part-time) of the Speech Processing Department (cont'd)

name & title	Email	position	period
Hubert Segot		Student (ENST)	4 – 9/1987
Eiichiro Kitagawa		Student (Waseda Univ)	8/1987
Kenji Hashimoto		Student (Waseda Univ)	8/1987
Takahito Yamazaki		Student (Shizuoka Univ)	8/1988
Furati Mulad		Student (Shizuoka Univ)	8/1988
Patrick G. Haffner		Student (ENST)	3 – 11/1988
Yasuhide Hashimoto		Student (Toyohashi Univ of Tech)	7 – 8/1987
Takashi Endo		Student (Waseda Univ)	8/1987
Rei Furukawa		Student (Waseda Univ)	3 – 4/1989
Jean-Claude Dang		Student (ENST)	3 – 11/1989
Yoshimitsu Hirata		Student (Toyohashi Univ of Tech)	1 – 2/1989
Mitsuru Noda		Student (Toyohashi Univ of Tech)	1 – 2/1989
Yasuhiro Minami		Student (Keio Univ)	3–6/1989, 8/1989–2/1990, 7–10/1990
Katsunobu Ito		Student (Tokyo Inst of Tech)	10 – 12/1989
Yoshio Ueda		Student (Toyohashi Univ of Tech)	1 – 2/1990
Takeshi Hirado		Student (Toyohashi Univ of Tech)	1 – 2/1990
Satoru Nakamura		Student (Keio Univ)	7 – 9/1990, 3 – 4/1991
Toshiyuki Nomura		Student (Nagoya Univ)	7 – 8/1990
Mamoru Watado		Student (Waseda Univ)	7 – 8/1990
Hidehiro Inagaki		Student (Waseda Univ)	7 – 8/1990
Yasushi Maruyama		Student (Shinshu Univ)	3-4/1990
Alain Biem		Student (INT)	8/1990 – 1/1991
Hidefumi Kikuchi		Student (Waseda Univ)	8/1991
Masafumi Tamoto		Student (Tokyo Inst of Tech)	8/1991
Keisuke Doi		Student (Ryukoku Univ)	8/1991 – 9/1992
Tetsuya Yoshida		Student (Keio Univ)	8 – 9/1991
Kentaro Kurinami		Student (Keio Univ)	8 – 9/1991
Romain Brunias		Student (INT)	8/1991 – 1/1992
Richard Lengagne		Student (INT)	8/1991 – 1/1992
Tadashi Okamoto		Student (Toyohashi Univ of Tech)	1 – 2/1992
Satoshi Aki		Student (Toyohashi Univ of Tech)	1 – 2/1992
Hideyuki Watanabe		Student (Hokkaido Univ)	2 – 4/1992
Hiroki Yamamoto	xyama	Student (Waseda Univ)	3 – 9/1992
Yasuo Iwai		Part-time (Univ of Kyoto)	8/1989 – 3/1990
Toshiyuki Sadanobu		Part-time (Univ of Kyoto)	8/1989 – 3/1990
Natsuya Yoshida		Part-time (Univ of Kyoto)	8/1989 – 3/1990
Miwako Kurihara		Part-time (Doshisha Univ)	4/1990 –
Mechtild Tronnier		Part-time (Kansai Univ)	9/1990 – 4/1991
Ming-Yong Zhou	zhou	Part-time (Osaka City Univ)	10/1990 –
Kazue Kinugasa	xkazue	Part-time (Kansai Univ)	2/1992 –

3.5 Research Facilities

In 1991, we renovated our research facilities. Now, each member of the research staff is provided a DECstation5000 (speed 24MIPS) or DECstation3100 (speed 16MIPS), with 24 - 120 Mbyte memory and a 600M - 3G byte disk, which is connected to an Ethernet. We have introduced 2 additional DECsystem5000s (currently, 4 systems in total) with over 30G byte disk memory each which are used as file servers to contain the large scale speech database and are accessed from workstations through the Ethernet. Some workstations are equipped with speech input/output devices. Older computer systems, including VAX8600, VAX8700, and VAX8800, are also connected to the Ethernet and accessed from workstations. The department has two Hewlett-Packard HP9000 model 750s and three model 730s which run at 76 MIPS for high-speed computations.

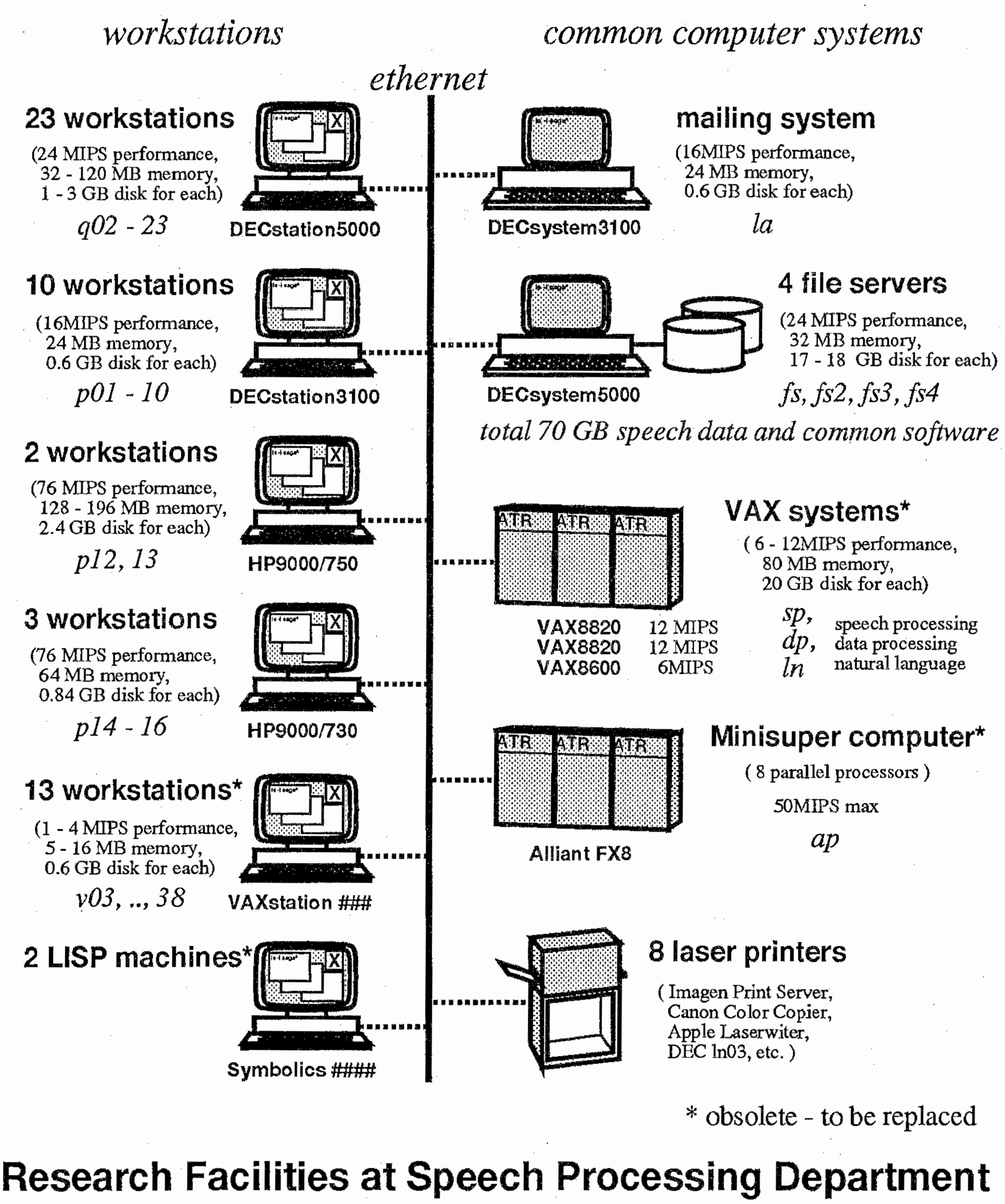


Figure 4: Computer Facilities of the Speech Processing Department, ATR Interpreting Telephony Research Laboratories

4 Speech Recognition Research

4.1 Phone Modeling

Reliable phoneme recognition and segmentation algorithms have been investigated, leading to considerable improvements over conventional approaches. We have been pursuing three approaches:

1. hidden Markov model approach
2. neural network approach
3. feature-based approach

These improvements have resulted in the successful implementation of continuous speech recognition systems such as HMM-LR and TDNN-LR. These systems were realized by combining HMM phoneme models and TDNN phoneme recognizers with the generalized LR parsing algorithm. Hardware implementation of HMM-LR has also been accomplished.

4.1.1 Hidden Markov Models

Hidden Markov models have been intensively studied for representing phone models in continuous speech recognition.

In the early stages of HMM research, studies on the number of states, tied probabilities of transition and output, probability smoothing techniques, initial probability settings, and duration control techniques were done based on the task of phoneme recognition with our large vocabulary speech database. Separate VQ (multiple codebook) and fuzzy VQ techniques improved the recognition rate from 86.5% to 95.7%. Major research topics have included the following items:

Research Subjects in 1991

- allophonic HMM (*Takami*)
- hidden Markov network topology generation: SSS (successive state splitting) algorithm (*Takami*)
- pitch-spectrum correlation exploited for improved phoneme models (*Singer*)
- tied rotation of observation vectors (*Rainton*)
- temporal characteristics of HMM (*Katagishi*)
- mixture density HMMs (*Yamaguchi*)
- miminimum classification error (MCE) training of HMM (*Rainton*)
- concatenated training (*Rainton*)
- HMM matrix parser (*Sagayama*)
- speaker independent HMM (*Kosaka*)
- speaker selection (*Miyazawa*)

Previous Research Subjects

- discrete HMM phone models: recognition rate of 86.5% (phoneme recognition rates for phonemes in word utterances)
- separate VQ + fuzzy VQ for HMM phone models: 95.7%

- HMM phone models based on fuzzy vector quantization
- continuous output probability density HMMs
- allophonic (context dependent) HMM phoneme models: context dependent phoneme recognition: 98.3% using single Gaussian models
- segment-based HMM
- HMM - neural network hybrid
- combining with a generalized LR parser
- word spotting in continuous speech
- phonetic typewriter
- English word recognition

HMM phone models are used in HMM-LR where the HMM phoneme models are combined with a generalized LR parser (a language source model) for efficiently recognizing a Japanese phrase input, word spotting in continuous speech, phonetic typewriter, and English word recognition.

A recent study on context-dependent phone models, represented by single Gaussian output probabilities, has attained a speaker-dependent phoneme recognition rate of 98.3% for phonemes in word utterances. For phonemes in phrase utterances, the models performed even better than the Gaussian mixture models.

4.1.2 Neural Network Approaches

ATR began neural network research in its early period and intensively investigated its application to speech recognition. The TDNN (time-delay neural network) is one subsequent development. Major topics in this area cover:

Research Subjects in 1991

- speaker-independent phoneme recognition neural network (*Sawai, Sugiyama*)
- incremental training of NN for phrase speech (*Kato*)
- error minimum criterion for NN training: 4-layer MLP (*Sugiyama*)
- multi-input/output units: fuzzy partition model (*Kato*)

Previous Research Subjects

- time delay neural network (TDNN)
- fast back-propagation algorithm: 1,000-fold speed-up
- neural network workbench for interactive design, training, and execution of neural networks
- phoneme spotting: 98% (frame correct)
- phoneme filter approach
- robustness across differences of speaking styles (solved by neighbor integration, KNIT (k-nearest neighbor interpolative training), pairwise discriminant TDNNs, fuzzy training)
- new architectures
- Boltzmann machine approach

TDNN attained a phoneme recognition rate of 98.6% for a /b/, /d/, /g/ task, and 96.7% for a full consonant task. Phoneme spotting by TDNN aimed at developing a continuous speech recognition approach by neural networks has been carried out and has resulted in a phoneme spotting rate of 98%.

Efforts to speed up the back-propagation algorithm resulted in a 1,000-fold speed-up of the TDNN. This speed-up and the use of a mini-supercomputer have made it possible to attempt larger scale neural networks.

Robustness across different speaking styles is one of the most important issues in neural networks for speech recognition. Although neural networks show very high phoneme recognition performance for word speech when they have been trained using word utterances, the performance degrades drastically for such different speaking styles as continuous speech. To solve this problem, intensive studies have been made and new schemes of neural network training (e.g. the KNIT method, fuzzy training) and new architectures (e.g. pairwise discriminant TDNNs) have been developed, and they perform much better than the original neural network approaches.

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

4.1.3 Feature-based Approaches

A feature-based approach is also taken for utilizing expert knowledge in spectrogram reading. Major topics in this area are:

Research Subjects in 1991

- feature based segmentation combined with HMM (*Fujiwara, Komori, Lengagne*)

Previous Research Subjects

- spectrum reading expert system
- hybrid system of knowledge based segmentation and neural network phoneme recognition

Phoneme segmentation knowledge for consonants is described by ART, a language for expert system description. A phoneme segmentation expert is integrated with a TDNN neural network for consonant discrimination. 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 syllabic nasals has been studied within a vowel spotting TDNN.

4.2 Speaker Adaptation and Noise-robust Speech Recognition

Aiming at a general preprocessor for speaker normalization, speaker adaptation research is being studied.

Research Subjects in 1991

- allophone adaptation (*Takami, Brunias*)
- a new framework of speaker independent speech recognition (*Miyazawa*)
- speaker-independent speech recognition (*Kosaka*)

- vector field approach to speaker adaptation using a small sized sample data (*Hattori*)
- segmental-VQ HMM for noisy speech (*Ohkura*)
- NN-based speaker adaptation (*Fukuzawa*)
- unsupervised speaker adaptation NN (*Fukuzawa*)

Previous Research Subjects

- codebook mapping for speaker adaptation
- supplemented codebook mapping
- speaker weighting for speaker adaptation
- HMM model modification "vector field smoothing"
- speaker mapping by a neural network
- speaker independent phoneme recognition by neural networks

Codebook mapping is a technique to map a pair of vector quantization codebooks of different speakers to each other using fuzzy vector quantization and DTW(dynamic time warping)-based word speech matching. This HMM speaker adaptation algorithm attains a phrase recognition rate of 78.6%, while the phrase recognition rates of speaker independence and speaker dependence are 59.6% and 88.4%, respectively. A supplemented HMM phoneme model training approach has been initiated to cope with speaker coarticulation variations. In the codebook mapping algorithm, a speaker adaptation approach has been applied with the concepts of fuzzy vector quantization and spectrum mapping.

Recently, adaptation of continuous output density HMM models, in place of such a frame-wise approach, have been found very successful and has attained a phoneme recognition performance higher than 90% after speaker adaptation. The wide applicability of this new principle, called "vector field smoothing", is being intensively investigated.

4.3 Language Source Modeling

Phone models are connected under the control of a language model to recognize continuous speech. We are taking both syntactic and stochastic approaches to language modeling. Some of our activities in this area are as follows:

Research Subjects in 1991

- allophone-based LR parser: 3 different approaches (*Nagai*)
- new search strategies for continuous speech recognition (*Yamaguchi*)
- spontaneous speech (*Murakami*)
- statistical grammar: generalized n-gram (*Murakami*)
- automatic grammar acquisition from data (*Murakami*)
- language discrimination (*Sugiyama*)

Previous Research Subjects

- generalized predictive LR parser
- HMM-LR continuous speech recognition
- phoneme context dependent LR parsing algorithm

- word/syllable/phoneme trigram models
- phonetic typewriter using phoneme sequence statistics
- word spotting based on HMM phone models
- two-stage LR parser for inter-*bunsetsu* syntax
- unknown word detection in HMM-LR
- the use of pitch patterns for phrase boundary detection
- NETgram: neural network approach for word category prediction
- TDNN-LR continuous speech recognition

A generalized LR parser has been combined with HMM phone models to search for the optimal grammatical path. Figure 5 shows an outline of the HMM-LR scheme for continuous speech recognition which attained a phrase recognition rate of 88% for a task having a phoneme perplexity of 5.9 with a vocabulary of 1,000 words. In this system, context-free rewriting rules are used to produce a better language source model. A recent enhancement of the LR parser with SSS algorithm has attained a phrase accuracy of 94.6%.

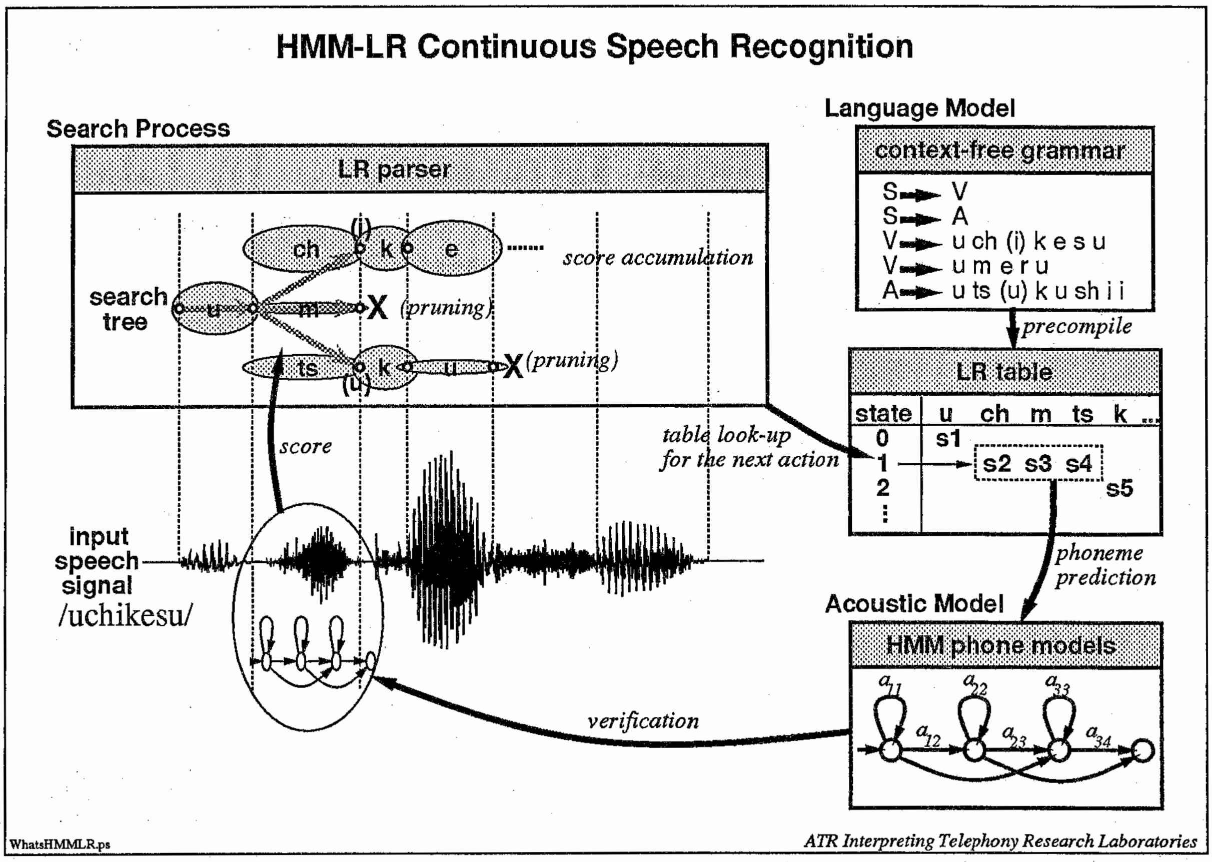


Figure 5: HMM-LR continuous speech recognition system

An experiment to recognize continuous utterances using TDNN phoneme spotting results has been conducted. This method sums up the phoneme spotting output by means of a DTW algorithm with an LR parser. Recently, it has been connected to both a PD-TDNN and a fuzzy trained TDNN, which has resulted in higher accuracies.

The NETgram approach for word category prediction using a neural network has also proven itself efficient when the training data size is relatively small compared with n-gram (bigram, trigram, etc.) approaches.

5 Speech Synthesis Research

5.1 Speech Synthesis by Rule

A high quality speech synthesis system by rule should be realized as an important part of the interpreting telephony system. A synthesis system based on flexible synthesis units has been developed. Moreover, the speech synthesis algorithm has been evaluated to realize a better synthesis system. A unit selection algorithm from a large-scale speech database and a unit concatenation algorithm have been developed. A prosody control algorithm has also been studied. Rules governing speaking styles and prosodic control have also been studied to realize various kinds of synthesized voices. Our research interests are:

Research Subjects in 1991

- prosody control for conversational speech (*Sagisaka, Kaiki*)
- statistical methods applied to duration and energy control (*Kaiki, Mimura, Campbell*)
- multi-speaker speech synthesis (*Kaiki, Iwahashi*)
- automatic unit generation (*Iwahashi*)
- optimal selection of units based on distortion minimization (*Iwahashi*)
- Automatic prosodic labeling for a speech database (*Campbell*)
- English speech synthesis (*Campbell*)
- hardware implementation of waveform generator

Previous Research Subjects

- speech synthesis using non-uniform speech units
- speaker conversion based on codebook mapping (*Abe*)
- segmental speaker conversion using HMM (*Abe*)

A prototype speech synthesis-by-rule system has been implemented.

5.2 Voice Conversion

Voice conversion from one speaker to another is an important aspect in realizing an automatic telephone interpreting system. A voice conversion algorithm based on codebook mapping by vector quantization was previously developed. Recently, our research interest has focused on two targets. One is the application to cross language voice conversion, where speech data from speaker S_1 of source language L_1 and from speaker S_2 of destination language L_2 are given for conversion into speaker S_1's voice in L_2. The other target is a segment-based approach to voice conversion, where the input speech is segmented using HMM phone models and converted at the segment level into another speaker's voice. Our research subjects cover:

Previous Research Subjects

- frame-wise voice conversion using VQ codebook mapping
- cross-language voice conversion
- segment-based voice conversion using HMM

6 Speech Database

A large-scale speech database with phonetic transcription, mainly for research in speech recognition and speech synthesis, has been developed with considerable support from other speech research laboratories. This database includes the following three categories:

(1) set-A: Large-scale database consisting of 5,240 Japanese common words, 216 phonetically-balanced words, 101 Japanese syllables, 25 numerics, 35 alphabetical letters, 9 foreign words, and 115 sentences in a conference registration task with 3 different speaking styles uttered by 20 professional broadcast announcers

(2) set-B: 503 phonetically balanced sentences uttered by 10 professional speakers

(3) set-C: a database of a large number of speakers (mainly for speaker-independent/speaker-adaptive speaker recognition research) containing 520 words (a subset of 5240 words from set-A), 216 phonetically-balanced words, 15 numerics, and 150 sentences (a subset of set-B), with an ultimate goal of 200-speaker collection, half of which will be phonetically labeled.

These databases are recorded and sampled at 20 kHz, digitized in 16 bits, and phonetically hand-labeled by well-trained workers.

The first two, (1) and (2), have already been completed and are gradually being made available to the public.

The speech database has been effectively used for speech research at ATR and other laboratories. Data of several speakers are already available on CD-ROMs. This speech database is now being accepted as the de facto standard in Japan.

7 Technical Publications from January through December 1991

The following pages are a list of technical publications, published or coauthored by the researchers of the Speech Processing Department, ATR Interpreting Telephony Research Laboratories, published during the period from January through December 1991, including reprints of these publications. Some of these works were done in their previous affiliations and published after their joining ATR.

Abbreviations used in the list are as follows:

- ASJ: the Acoustical Society of Japan
- IEICE: the Institute of Electronics, Information and Communications Engineers (Japan)
- ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing
- Eurospeech: Europian Conference on Speech Communications
- ICPhS: International Conference on Phonetical Sciences

List of Publications, January through December, 1991

[Fukuzawa91NNSP9] 107

Keiji Fukuzawa, Hidefumi Sawai, Masahide Sugiyama: "Segment-Based Speaker Adaptation by Neural Network," Proc. of Neural Networks for Signal Processings, pp. 442–451 (1991.10).

◆

[Hattori91ASJ3] 117

服部 浩明, 嵯峨山 茂樹: "少量データを用いたコードブックマッピングによる話者適応化," 日本音響学会平成 3 年度春季研究発表会講演論文集, 1-5-23, pp. 49–50 (1991.03).

(*Hiroaki Hattori, Shigeki Sagayama: "Speaker Adaptation with Small Size Data Based on Codebook Mapping Algorithm," ASJ Spring Meeting, 1-5-23, pp. 49–50 (1991.03).*)

[Hattori91ICASSP5] 119

Hiroaki Hattori: "Speaker Adaptation Based on Markov Modeling of Speakers in Speaker-Independndct Speech Recognition," Proc. of 1991 International Conference on Acoustics, Speech, and Signal Processing, S13.4, pp. 845–848, Toronto, Canada (1991.05).

[Hattori91SP12] 123

服部 浩明: "ニューラルネットワークを用いたテキスト独立な話者認識," 電子情報通信学会技術研究報告, SP91-90, pp. 71–78 (1992.12).

(*H. Hattori: "Text-Independent Speaker Recognition Using Neural Networks," IEICE Technical Report, SP91-90, pp. 71–78 (1992.03).*)

◆

[Hedelin90JSC9] 131

Per Hedelin, Dieter Huber: "The CTH Speech Database: An Intergrated Multilevel Approach," Journal Speech Communication, pp. 365–374, North-Holland (1990.09).

◆

[Huber90SST11] 141

Dieter Huber: "Speech Style Variations of F_0 in a Cross-Linguistics Perspective," Proc. of SST'90 (1990.11).

[Huber90EDR11] 147

Dieter Huber: "An Electronic Dictionary for Computer Speech Applications," The EDR International Workshop on Electronic Dictionaryes, pp. 1–11, Kanagawa (1990.11).

[Huber91ICPhS8] 158

Dieter Huber: "On the Discourse Function of Intonation," Proc. ICPhS'91, Vol. 5, pp. 190–193, Aix-en-Provence, France (1991.08).

◆

[Iwahashi91ASJ3] 162

岩橋 直人, 海木 延佳, 嵯峨山 茂樹, 匂坂 芳典: "音響的尺度を用いた合成単位素片の選択法," 日本音響学会平成 3 年度春季研究発表会講演論文集, 1-6-1, pp. 207–208 (1991.03).

(*Naoto Iwahashi, Nobuyoshi Kaiki, Shigeki Sagayama, Yoshinori Sagisaka: "Unit Selection with Acoustic Measure for Speech Synthesis," ASJ Spring Meeting, 1-6-1, pp. 207–208 (1991.03).*)

[Iwahashi91SP5] 164

岩橋 直人, 海木 延佳, 匂坂 芳典: "音響的尺度に基づく複合音声単位選択法," 電子情報通信学会技術研究報告, SP91-5, pp. 33–40 (1991.05).

(*Naoto Iwahashi, Nobuyoshi Kaiki, Yoshinori Sagisaka: "Non-uniform Unit Selection based on Several Acoustic Measures for Speech Synthesis," IEICE Technical Report, SP91-5, pp. 33–40 (1991.05).*)

[Iwahashi91ASJ10] 172

岩橋 直人, 藤原 紳悟, 小森 康弘, 杉山 雅英, 匂坂 芳典: "自動セグメンテーションによる音声合成単位の作成," 日本音響学会平成 3 年度秋季研究発表会講演論文集, 1-6-21, pp. 231–232 (1991.10).

(*Naoto Iwahashi, Shingo Fujiwara, Yasuhiro Komori, Masahide Sugiyama, Yoshinori Sagisaka: "Generation of Speech Synthesis Units by Automatic Labelling," ASJ Fall Meeting, 1-6-21, pp. 231–232 (1991.10).*)

◆

[Kaiki91ASJ3a] 174

海木 延佳, W. N. Campbell, 匂坂 芳典: "制御構造を考慮した音韻継続時間長の設定," 日本音響学会平成 3 年度春季研究発表会講演論文集, 1-6-9, pp. 223–224 (1991.03).

(*Nobuyoshi Kaiki, W. N. Campbell, Yoshinori Sagisaka: "Incorporating Segmental Duration Setting from the Mora Level Control," ASJ Spring Meeting, 1-6-9, pp. 223–224 (1991.03).*)

[Kaiki91ASJ3b] 176

海木 延佳, 匂坂 芳典: "種々の発話様式における音韻継続時間長の分析," 日本音響学会平成 3 年度春季研究発表会講演論文集, 1-6-10, pp. 225–226 (1991.03).

榑松 明: “音声認識,” 通信工業, 6 月号, pp. 14–19 (1991.06).
(*Akira Kurematsu: “Speech Recognition,” Tsushin Kogyo, Vol. 6, pp. 14–19 (1991.06).*)

[Kurematsu91DenkiTsushin10] 247

榑松 明: “自動翻訳電話技術の動向,” 電気通信, Vol. 54, No. 538, pp. 62–74 (1991.10).
(*Akira Kurematsu: “Overview of Automatic Telephone Interpretation Technology,” Denki Tsushin, Vol. 54, No. 538, pp. 62–74 (1991.10).*)

[Kurematsu91NCF10] 260

Akira Kurematsu: “A Perspective of Speech Translation Research,” Proc. of 1991 National Communication Forum (1991.10).

◆

[Ladd91ICPhS8] 281

D. R. Ladd, W. N. Campbell: “Theories of Prosodic Structure: Evidence From Syllable Duration,” Proc. ICPhS’91, Vol. 2, Aix-en-Provence, France, pp. 290–293 (1991.08).

◆

[Mimura91ASJ3] 285

三村 克彦, 匂坂 芳典: “音声パワーの動特性の分析と制御,” 日本音響学会平成 3 年度春季研究発表会講演論文集, 1-6-14, pp. 233–234 (1991.03).
(*Katsuhiko Mimura, Yoshinori Sagisaka: “Analysis of Temporal Patterns of Speech Power,” ASJ Spring Meeting, 1-6-14, pp. 233–234 (1991.03).*)

[Mimura91SP5] 287

三村 克彦, 海木 延佳, 匂坂 芳典: “統計的手法を用いた音声パワーの分析と動特性の制御,” 電子情報通信学会技術研究報告, SP91-4, pp. 25–32 (1991.05).
(*Katsuhiko Mimura, Nobuyoshi Kaiki, Yoshinori Sagisaka: “Analysis and Control of Temporal Patterns of Speech Power Using Statiatical Methods,” IEICE Technical Report, SP91-4, pp. 25–32 (1991.05).*)

[Mimura91Korea7] 295

Katsuhiko Mimura, Nobuyoshi Kaiki, Yoshinori Sagisaka: “Statistically Derived Rules for Amplitude and Duration Control in Japanese Speech synthesis,” 電子情報通信学会技術研究報告, SP91-41, pp. 19–24, (1991.07).
(*IEICE Technical Report; Proc. of Korea-Japan Joint Workshop on Advanced Technology of Speech Recongnition and Synthesis, pp. 151–156, Seoul (1991.07).*
)

David Rainton, Shigeki Sagayama: “Class Specific Observation Vector Orthogonalisation and Its Interpretation as a Form of Tied Continuous Mixture HMM,” 電子情報通信学会技術研究報告, SP91-56, pp. 7–13 (1991.09).
(*David Rainton, Shigeki Sagayama: “Class Specific Observation Vector Orthogonalisation and Its Interpretation as a Form Of Tied Continuous Mixture HMM,” IEICE Technical Report, SP91-56, pp. 7–13 (1991.09).*)

[Rainton91ASJ10] 400

David Rainton, 嵯峨山 茂樹: “A New Tied Continuous Mixture Density HMM via Orthogonalisation of the Full Covariance Observation Matrix,” 日本音響学会平成 3 年度秋季研究発表会講演論文集, 2-5-15, pp. 77–78 (1991.10).
(*David Rainton, Shigeki Sagayama: “A New Tied Continuous Mixture Density HMM Via Orthogonalisation of The Full covariance Observation Matrix,” ASJ Fall Meeting, 2-5-15, pp. 77–78 (1991.10).*)

◆

[Sagayama91Book] .... (not included in this collection)

Shigeki Sagayama: “Speech Analysis,” in *Speech Science and Technology* (S. Saito ed.), pp. 64-96 and pp. 318–326, Ohmsha Pub., Tokyo (in Japan) and IOS Press (other countries) (1992).

[Sagayama91SPREC2] 402

嵯峨山 茂樹: “HMM 尤度行列演算による音声認識アルゴリズム,” 隠れマルコフモデル・ニューラルネットワーク・シンポジウム (箱根強羅), pp. 39–42 (1991.02).
(*Shigeki Sagayama: “A Speech Recognition Algorithm Based on HMM Likelihood Matrix Computation,” SPREC Symposium (Hakone), pp. 39–42 (1991.02).*)

[Sagayama91TohokuSympo2] 406

嵯峨山 茂樹: “数理統計モデルに基づく音声認識,” 第 27 回東北大通研シンポジウム「パターンの認識・理解における諸問題とその実現」予稿集, 2-6, pp. 119–127 (1991.02).
(*Shigeki Sagayama: “Speech Recognition Based on Statistical Models,” Proc. of 27th Symposium, Institute of Electrical Communications, Tohoku University, 2-6, pp. 119–127 (1991.02).*)

[Sagayama91ISCIEcourse2] 415

嵯峨山 茂樹: “最近の音声認識研究,” 第 40 回システム制御情報講習会「ここまできた高度情報化技術」テキスト, システム制御情報学会, pp. 93–116 (1991.02).
(*Shigeki Sagayama: “Recent and Current Status of Speech Recognition Research,”*

The text of the 40th Tutorial Course of ISCIE: "The Recent Information Processing Technologies," pp. 93–116 (1991.02).)

鷹見 淳一, 甲斐 充彦, 嵯峨山 茂樹: “対判定型 TDNN と予測 LR パーザの結合による連続音声認識,” 日本音響学会平成 3 年度秋季研究発表会講演論文集, 1-5-22, pp. 43–44 (1991.10).
(*Jun-ichi Takami, Atsuhiko Kai, Shigeki Sagayama: “Continuous Speech Recognition by Combining PD-TDNNs and Predictive LR Parser,” ASJ Fall Meeting, 1-5-22, pp. 43–44 (1991.10).*)

[Takami91ASJ10b] 594

鷹見 淳一, 嵯峨山 茂樹: “逐次状態分割法 (SSS) による隠れマルコフネットワークの自動生成,” 日本音響学会平成 3 年度秋季研究発表会講演論文集, 2-5-13, pp. 73–74 (1991.10).
(*Jun-ichi Takami, Shigeki Sagayama: “Automatic Generation of the Hidden Markov Network by the Successive State Splitting Algorithm,” ASJ Fall Meeting, 2-5-13, pp. 73–74 (1991.10).*)

[Takami91SP12] 596

鷹見 淳一, 嵯峨山 茂樹: “音素コンテキストと時間に関する逐次状態分割による隠れマルコフ網の自動生成,” 電子情報通信学会技術研究報告, SP91-88, pp. 57–64 (1991.12).
(*Jun-ichi Takami, Shigeki Sagayama: “Automatic Generation of the Hidden Markov Network by Successive State Splitting on Contextual Domain and Temporal Domain,” IEICE Technical Report, SP91-88, pp. 57–64 (1991.12).*)

◆

[Takeda91JASA12] 604

Kazuya Takeda, Yoshinori Sagisaka, Hisao Kuwabara: “On Sentence-level Factors Governing Segmental Duration in Japanese,” The TeX Journal of the Acoustical Society of America, Vol.86(6), pp. 2081–2087 (1989.12).

◆

[Takezawa91IEICE9] 611

竹澤 寿幸, 森元 逞, 樽松 明: “日英音声言語翻訳実験システム SL-Trans2 における音声対訳処理,” 電子通信学会秋季全国大会講演論文集, D-53, 分冊 6, pp. 53–54 (1991.09).
(*Toshiyuki Takezawa, Tsuyoshi Morimoto, Akira Kurematsu: “Interaction Processing in SL-TRANS 2: an Experimental System for Translating Japanese Speech to English,” Proc. of the Institute of Electronics and Communication Engineers of Japan, Vol. 6, pp. 53–54 (1991.09).*)

[Takezawa91ASJ10] 613

竹沢 寿幸, 大倉 計美, 森元 逞, 嵯峨山 茂樹, 榑松 明: “日英音声言語翻訳実験システム SL-TRANS2,” 日本音響学会平成 3 年度秋季研究発表会講演論文集, 1-5-24, pp. 47–48 (1991.10).
(*Toshiyuki Takezawa, Kazumi Ohkura, Tsuyoshi Morimoto, Shigeki Sagayama, Akira Kurematsu: "SL-TRANS 2: an Experimental System for Translating Japanese Speech to English," ASJ Fall Meeting, 1-5-24, pp. 47–48 (1991.10).*)

◆

[Umezaki91IEICEtrans4] 615

梅崎 太造, Harald Singer, 板倉 文忠: “平滑化群遅延スペクトル距離尺度の不特定話者音声認識における評価,” 電子情報通信学会論文誌 A, Vol. J74-A, No. 4, pp. 610–618 (1991.04).
(*Taizo Umezaki, Harald Singer, Fumitada Itakura: "Evaluation of the Smoothed Group Delay Spectrum Distance Measure in Speaker-Independent Speech Recognition," Trans. of IEICE, Vol. J74-A, No. 4, pp. 610–618 (1991.04).*)

[Umezaki91ROC11] 624

Taizo Umezaki, Harald Singer, Fumitada Itakura: "Evaluation of the Smoothed Group Delay Spectrum Distance Measure in Speaker-Independent Speech Recognition," 1. ROC-Japan Seminar on New Speech Recognition Method, pp. 179–188, National Chiao Tung University, Taiwan (1991.11).