The ATR Multilingual Speech-to-Speech Translation System

Satoshi Nakamura, *Member, IEEE*, Konstantin Markov, *Member, IEEE*, Hiromi Nakaiwa, Genichiro Kikui, Hisashi Kawai, *Member, IEEE*, Takatoshi Jitsuhiro, *Member, IEEE*, Jin-Song Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto, *Fellow, IEEE*

***Abstract*—In this paper, we describe the ATR multilingual speech-to-speech translation (S2ST) system, which is mainly focused on translation between English and Asian languages (Japanese and Chinese). There are three main modules of our S2ST system: large-vocabulary continuous speech recognition, machine text-to-text (T2T) translation, and text-to-speech synthesis. All of them are multilingual and are designed using state-of-the-art technologies developed at ATR. A corpus-based statistical machine learning framework forms the basis of our system design. We use a parallel multilingual database consisting of over 600 000 sentences that cover a broad range of travel-related conversations. Recent evaluation of the overall system showed that speech-to-speech translation quality is high, being at the level of a person having a Test of English for International Communication (TOEIC) score of 750 out of the perfect score of 990.**

***Index Terms*—Example-based machine translation (EBMT), minimum description length (MDL), multiclass language model, speech-to-speech translation (S2S), statistical machine translation (SMT), successive state splitting (SSS), text-to-speech (TTS) conversion.**

I. Introduction

SPEECH-TO-SPEECH translation (S2ST) is a pipe dream for human beings that enables communication between people speaking in different languages. Since our world is becoming borderless day by day, the importance of S2ST technology has been increasing. ATR began its S2ST research in order to overcome the language barrier problem in 1986. So far, we have been working on speech recognition, machine translation, speech synthesis, and integration for an S2ST system. The history of our S2ST research can be divided into three phrases. The first phase focused on a feasibility study of S2ST that only allowed limited vocabulary and clear read-style speech. In the second phase, we extended the technology to handle "natural" conversations in a limited domain. We are currently in the third phase, which began in 2000. Its target is to develop technologies to make the S2ST system work in real environments.

The drastic increase in demand for translingual conversations, triggered by IT technologies such as the Internet and an expansion of borderless communities as seen in the increase in the number of EU countries, has boosted research activities on S2ST technology. Many research projects have addressed speech-to-speech translation technology, such as VERBMOBIL [1], C-STAR,[1] NESPOLE!,[2] [2], and BABYLON.[3] These projects mainly focused on the construction of prototype systems for several language pairs.

S2ST between Western languages and a non-Western language, such as English-from/to-Japanese, or English-from/to-Chinese, requires technologies to overcome the drastic differences in linguistic expressions. For example, a translation from Japanese to English requires 1) a word separation process for Japanese because Japanese has no explicit spacing information, and 2) transforming the source sentence into a target sentence with a drastically different style because their word order and their coverage of words are completely different, among other factors.

The other factor for S2ST is that the technology must be portable for every domain because S2ST systems are often used for applications in a specific situation, such as supporting a tourist's conversations in nonnative languages. Therefore, the S2ST technique must include (semi-) automatic functions for adapting to specific situations/domains and specific language pairs in speech recognition, machine translation, and speech synthesis [3].

Multilingual speech-to-speech translation devices are vital for breaking the language barrier, which is one of the most serious problems inherent in globalization. In S2ST, machine translation is the core technology for generating natural translation from original input. Therefore, the performance of S2ST relies heavily on the performance of the machine translation system. There are many machine translation systems on the market. However, most of the currently available ones are handcrafted rule-based translation systems designed for written text, mainly because it is difficult to gather data that exhaustively cover diverse language phenomena. In rule-based systems, efforts have been made to improve rules that abstract the language

Manuscript received June 18, 2004; revised January 7, 2005. This work was supported in part by a contract with the National Institute of Information and Communications Technology entitled "A study of speech dialog translation technology based on a large corpus" and in part by a contract with the Ministry of Public Management, Home Affairs, Posts, and Telecommunications entitled "Multilingual speech-translation systems using mobile terminals." The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Timothy J. Hazen.

The authors are with the ATR Spoken Language Translation Research Laboratories, 619-0288, Kyoto, Japan (e-mail: satoshi.nakamura@atr.jp; konstantin.markov@atr.jp; hiromi.nakaiwa@atr.jp; genichiro.kikui@atr.jp; hisashi.kawai@atr.jp; takatoshi.jitsuhiro@atr.jp; jinsong.zhang@atr.jp; hirofumi.yamamoto@atr.jp; eiichiro.sumita@atr.jp; seiichi.yamamoto@atr.jp).

Digital Object Identifier 10.1109/TSA.2005.860774

[1]C-STAR. Consortium for Speech Translation Advanced Research. http://www.c-star.org.

[2]NESPOLE! Negotiating Through Spoken Laguage Through E-Commerce. http://nespole.itc.it.

[3]BABYLON. http://darpa-babylon.mitre.org/index.html.

phenomena by using human insight. In taking this type of approach, however, it is difficult to port a particular system to other domains, or to upgrade the system to accommodate new expressions. Portability is one of the most important factors for S2ST, because S2ST systems are often designed for a specific domain and situation for various language pairs depending on their users. Therefore, customization for their domains and situations and for their language pair is obligatory work for S2ST.

With the increased availability of substantial bilingual corpora by the 1980s, corpus-based machine translation (MT) technologies such as example-based MT and stochastic MT were proposed to cope with the limitations of the rule-based systems that had formerly been the dominant paradigm. Since that time, we have conducted research on applying corpus-based methods to speech translation and have developed several technologies. Our research experience shows us that corpus-based approaches are suitable for speech translation technology. This is because corpus-based methods:

1) can be applied to different domains;
2) are easy to adapt to multiple languages;
3) can handle ungrammatical sentences, which are common in spoken language.

One of our research themes is to develop example-based translation technologies that can be applied across a wide range of domains, and to develop stochastic translation technologies that can be applied to language pairs with completely different structures, such as English and Japanese. Example-based methods and stochastic methods each have different advantages and disadvantages, so we plan to combine them into a single, more powerful system.

At present, however, corpus-based methods can only be applied to narrow domains due to the lack of sufficiently large bilingual spoken language corpora. Therefore, one of our subthemes is to establish a methodology for gathering large volumes of data to enable us to translate various expressions at high quality. For this subtheme, we have started to conduct research on several methods, including paraphrasing, to create huge bilingual corpora, and on methods for evaluating the coverage of the collected corpora.

A speech recognition system should be robust enough to recognize speech in noisy environments with various speaking styles. The machine translation system needs to be domain portable and to be good at translating a wide variety of topics. Speech synthesis must realize more natural and expressional speech quality. In this project, all the researchers—including speech processing researchers and natural language researchers—are working collaboratively and closely to achieve the S2ST system. For the S2ST system to be successful, the speech recognition system should recognize speaker-independent, continuous, spontaneous conversational speech. Back in 1986, the state-of-the-art technology of speech recognition could only recognize speaker dependent connected words from a small vocabulary. Thanks to many efforts made so far based on statistical modeling technologies like hidden Markov models (HMMs) and N-grams and large amounts of speech and text corpora, the recognition of speaker-independent, continuous conversational speech will soon be available. The next point for consideration in developing speech recognition for the S2ST system is to make the speech recognition system multilingual.

Corpus-based technologies are undoubtedly a major trend in contemporary text-to-speech (TTS) systems. In contrast to the conventional rule-based approach where experts' linguistic and acoustic knowledge is manually implemented in TTS systems, the corpus-based approach makes it possible to extract the knowledge from corpora and encode it in TTS systems in an automatic manner. Consequently, corpus-based systems are easy to build and have higher quality than rule-based systems in general. A drawback of the corpus-based systems, if any, is that they require a large memory size to store corpus data. For this reason, rule-based systems are often used in embedded applications.

Among corpus-based approaches, waveform concatenation techniques are widely adopted in commercial and experimental TTS systems for their natural-sounding output speech. Many studies have been conducted on these technologies since the early 1990s [4]–[6]. Among them, ATR, as one of the pioneers in corpus-based speech synthesis technology, has made major contributions to the progress of the technology through various studies, which led us to the development of two TTS systems, ν-talk [7] and CHATR [8].

Sagisaka proposed a novel synthesis scheme [9] in which nonuniform phoneme sequences were used as synthesis units. A unique point in this scheme was that it was able to make full use of a large speech corpus, whereas in conventional schemes a set of syllables, such as CV or VCV syllables (C: Consonant, V: Vowel), was extracted from a speech corpus and stored in the system as a unit inventory. The work was the first step toward the corpus-based speech synthesis. Iwahashi *et al.* [10] proposed a segment selection algorithm that searches for an optimal sequence of speech segments in terms of an acoustic criterion by using the dynamic programming algorithm. As a result of subsequent work by Sagisaka and his coworkers, they developed a TTS system named ν-talk [7]. Issues left unresolved included 1) vocoder-like speech quality caused by cepstral parametrization of speech segments and 2) poor correlation between the acoustic criteria and perceptual measures.

After ν-talk, a new TTS system named CHATR was developed [8]. Although CHATR was originally designed to be a workbench for speech synthesis research, it later became known as a TTS system based on waveform concatenation. Another important feature of CHATR was that it was designed so that it is easily applied to different languages. Indeed, TTS systems of several languages, including Japanese, English, German, Chinese, and Korean, were built into the framework of CHATR. Although CHATR produced very natural speech in limited domains, the quality was unstable for unrestricted domains. The identified problems included the following:

1) a weakness in text processing;
2) a weakness in prosody modeling;
3) small corpora (the largest corpus available was a 2-h Japanese corpus, and CHATR was capable of handling up to a 10-h corpus by design),
4) the cost function for segment selection was purely based on acoustical distances and it was not perceptually justified.

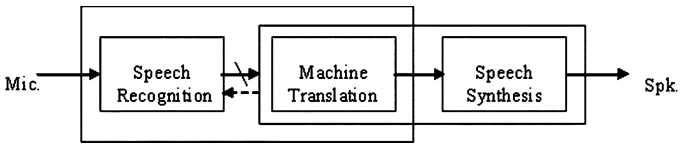


Fig. 1. Block diagram of the ATR S2ST system.

The rest of this paper is organized as follows. Section II introduces briefly the speech-to-speech translation task as a statistical problem. Section III gives an overview of the ATR S2ST system. Each of our system modules, i.e., speech recognition, machine translation, and speech synthesis is described in detail in Sections IV –VI, respectively. The Basic Travel Expression Corpus (BTEC) and MT-assisted dialogs (MAD) corpora that were used to build our system are presented in Section VII. Details about the evaluation experiments and achieved results are reported in Section VIII. Finally, we draw some conclusions in Section IX.

II. Speech-to-Speech Translation Background

The goal of automatic speech-to-speech translation is to generate a speech signal in one (target) language that conveys the linguistic information contained in a given speech signal of another (source) language.

A statistical approach to the speech-to-speech translation task gives the following formal solution:

$$S_T^* = \arg\max_{S_T} P(S_T|S_S) \tag{1}$$

where S_S and S_T are the speech signals in the source and target languages. As direct evaluation of the conditional probability $P(S_T|S_S)$ is intractable, it can be factorized as

$$\begin{aligned} P(S_T|S_S) &= \sum_{T_T,T_S} P(S_T,T_T,T_S|S_S) \\ &= \sum_{T_T,T_S} P(S_T|T_T,T_S,S_S)P(T_T|T_S,S_S)P(T_S|S_S) \\ &\approx \sum_{T_T,T_S} P(S_T|T_T)P(T_T|T_S)P(T_S|S_S) \end{aligned} \tag{2}$$

where T_S and T_T are the text transcriptions of the source and target speech signals. Then, the maximization of $P(S_T|S_S)$ can be further simplified to

$$\begin{aligned} \max_{S_T} P(S_T|S_S) = \max_{S_T} P\left(S_T|T_T^*\right) \\ \times \max_{T_T} P\left(T_T|T_S^*\right) \max_{T_S} P(T_S|S_S) \end{aligned} \tag{3}$$

where T_T^* and T_S^* are arguments maximizing the second and third terms. This equation suggests that the S2S translation problem can be decomposed into three independent parts: $P(T_S|S_S)$, which represents speech recognition; $P(T_T|T_S)$, which is a text-to-text translation model; and $P(S_T|T_T)$, which corresponds to speech synthesis.

III. Overview of the S2S Translation System

The overall speech-to-speech translation system is shown in Fig. 1. The system consists of three major modules, i.e., a multilingual speech recognition module, a multilingual machine translation module, and a multilingual speech synthesis module. Those modules are designed to process Japanese, English, and Chinese using a corpus-based method. The term "corpus" means a database with linguistic information. Thus, "A corpus-based method" means a method that uses those corpora. In our case, statistical methods utilizing those corpora are deployed, such as HMMs and N-grams for speech recognition, statistical and example-based translation for language translation, and waveform concatenation for speech synthesis. In the following sections, those methods are introduced in detail.

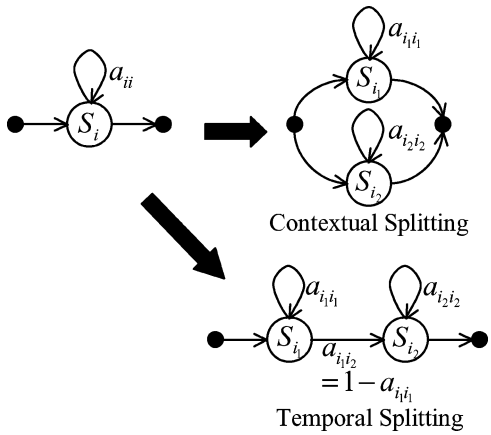


Fig. 2. Contextual splitting and temporal splitting.

IV. Multilingual Speech Recognition

A. Successive State Splitting (SSS)-Based Acoustic Modeling

For acoustic modeling in speech recognition, context-dependent phone models can obtain much better performance than context-independent phone models. While context-dependent phone models have many parameters, the most important problem to solve has been how to efficiently capture contextual and temporal variations in training speech and properly model them with fewer parameters.

Phonetic decision tree clustering [11] was proposed as a method for generating tied-state structures of acoustic models for speech recognition, while the SSS algorithm was originally proposed by ATR to create a network of HMM states of speaker-dependent models [12]. The SSS was subsequently expanded to the ML-SSS algorithm to create speaker-independent models [13] by data-driven clustering with contextual information.

However, since these methods are based on the maximum likelihood (ML) criterion, the likelihood value for training data increases as the number of parameters increases. To overcome this problem, we have recently proposed the ML-SSS algorithm based on the minimum description length (MDL) criterion as the splitting and stop criteria [14]. This algorithm is referred to as "the MDL-SSS algorithm." We will describe the ML-SSS algorithm and the MDL-SSS algorithm in the following sections.

1) ML-SSS Algorithm: The ML-SSS algorithm iteratively constructs the appropriate context-dependent model topologies by finding a state that should be split at each iteration. It then reestimates the parameters of HMMs based on the ML criterion in the same way as in phonetic decision tree clustering. This algorithm supposes the two types of splitting shown in Fig. 2.

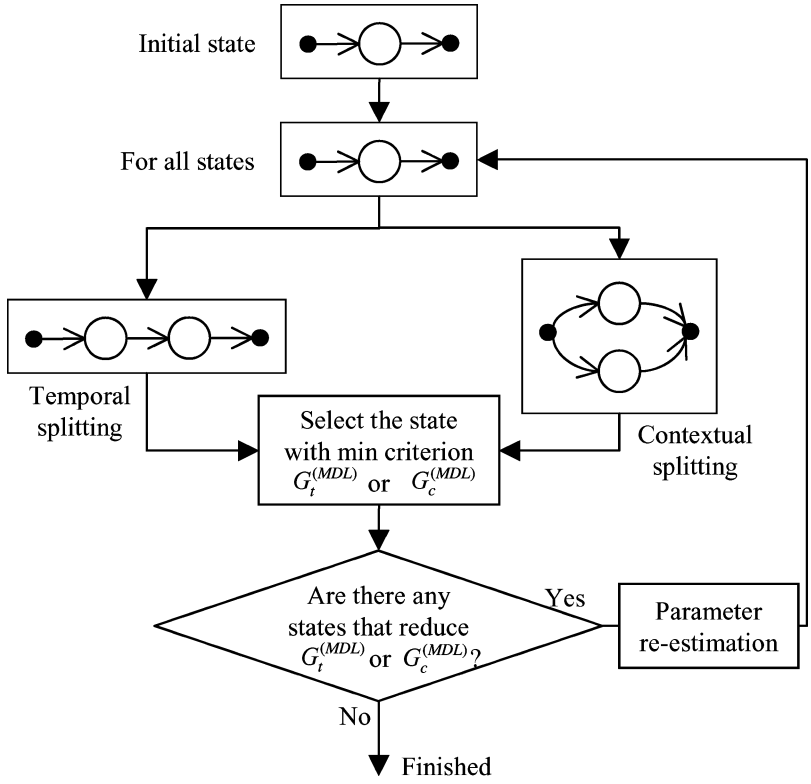


Fig. 3. Flow chart of MDL-SSS algorithm.

After contextual splitting, the total expected gain is calculated. For temporal splitting, the ML-SSS algorithm creates one more state and connects it to the original state. The parameters of the two distributions are estimated by the forward–backward algorithm, and the total expected gain of temporal splitting is also calculated for the temporal split states. Since it is computationally expensive to reestimate all of the parameters of a network at every splitting, approximated likelihood values are used. Next, the gains of both contextual and temporal splitting are calculated for all states. Finally, these expected gains are compared with each other and the split with the best gain among all states is selected. The total number of states and the maximum temporal length of states for each triphone model are stop criteria and must be given before splitting begins. Nonetheless, it is difficult to find the optimal values of these parameters. Accordingly, a sequence of experiments needs to be done to find the optimal values by changing parameters heuristically.

2) SSS Algorithm Using MDL Criterion: Fig. 3 shows the flow of the MDL-SSS algorithm. The differences in the MDL values for both contextual and temporal splitting are calculated for each state, and the split with the smallest difference value is chosen. Splitting is finished when there is no state that can be split and reduce the criterion by splitting. The total number of states and the maximum number of states per triphone are not required as stop criteria.

We define the criteria for contextual splitting and temporal splitting, $G_c^{(\mathrm{MDL})}$ and $G_t^{(\mathrm{MDL})}$, respectively, as follows:

$$G_c^{(\mathrm{MDL})}(S_i) = -G_c^{(\mathrm{ML})}(S_i) + C_c K \log \Gamma(S) \quad (4)$$

$$G_t^{(\mathrm{MDL})}(S_i) = -G_t^{(\mathrm{ML})}(S_i) + C_t \frac{(2K+1)}{2} \times \{(M+1)\log \Gamma'(S) - M \log \Gamma(S)\} \quad (5)$$

where the order of features is K, and the total number of states is M. The first terms, $G_c^{(\mathrm{ML})}$ and $G_t^{(\mathrm{ML})}$, in the right-hand sides are the negative values of the expected gains in the ML-SSS algorithm, while C_c and C_t are the scaling factors of the second terms. $\Gamma(S) = \sum_{i=1}^{N_s} \Gamma(S_i)$ represents the expected frequency of the number of samples for all states, whereas $\Gamma'(S)$ is the value after temporal splitting. Equation (5) compensates the total number of samples $\Gamma(S)$ because segments that are shorter than the lengths of state sequences are discarded. Moreover, $\Gamma(S)$ will be decreased to $\Gamma'(S)$ if a temporal split is selected. The MDL-SSS algorithm selects the state with the smallest $G_c^{(\mathrm{MDL})}$ or $G_t^{(\mathrm{MDL})}$, and stops splitting when $G_c^{(\mathrm{MDL})} > 0$ and $G_t^{(\mathrm{MDL})} > 0$ for all states.

B. Advanced Language Modeling

1) Multidimensional Word Classes: In the conventional word class N-gram defined by

$$P(w_i|w_{i-N+1}\dots w_{i-1}) = P\left(C(w_i)|C(w_{i-N+1})\dots C(w_{i-1})\right) P\left(w_i|C(w_i)\right) \quad (6)$$

only one-dimensional word classes are used [15]. Both the left- and right-context Markovian dependencies are used together. Only words having the same left- and right-context Markovian dependence belong to the same word class. This word class definition is not adequate for representing the Markovian dependence for words that have only the same left- or right-context Markovian dependence, such as "a" and "an." The left context of "a" and "an" is almost equivalent; however, the right context is significantly different. The difference between left and right context is more serious in languages with inflection, such as French and Japanese. For example, the Japanese inflection form

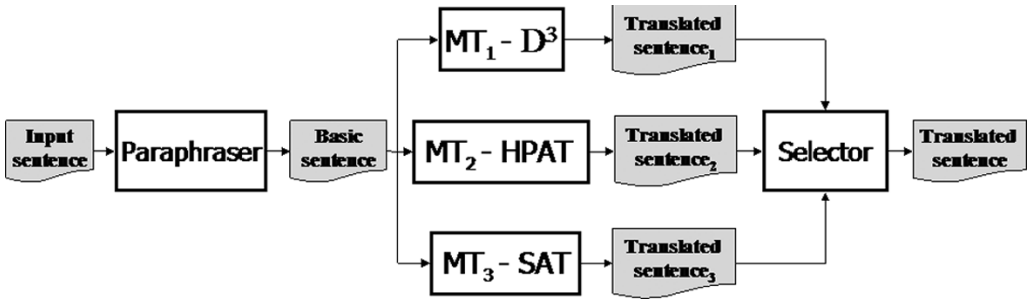


Fig. 4. Overview of the machine translation system developed by the C-cube project.

has an influence only on the right context, while the left-context Markovian dependence can be shared between the same words with different inflection forms.

We have introduced the idea of multidimensional word classes to represent left- and right-context Markovian dependence separately [16]. The multidimensional word classes can assign the same word class to "a" and "an" to represent the left-context Markovian dependence (left-context class), and assign them to different word classes to represent the right-context Markovian dependence (right context class). Each multidimensional word class is automatically extracted from the corpus using statistical information, rather than grammatical information such as part-of-speech (POS).

2) Class N-Grams Based on Multidimensional Word Class: Applying multidimensional word classification to formula (6), the following formula is obtained:

$$P(w_i|w_{i-N+1}\dots w_{i-1}) = P(C^l(w_i)|C^{rN-1}(w_{i-N+1})\dots C^{r2}\left(w_{i-2}C^{r1}(w_{i-1})\right)P\left(w_i|C^l(w_i)\right) \quad (7)$$

where the suffix for class C is used to represent position-dependent (left- and right-context) Markovian dependence. Here, $C^l(w)$ represents the left context class to which the word w belongs, and $C^{ri}(w)$ represents the right context class to which the ith word w belongs. Hereafter, we refer to these class N-grams based on multidimensional classes as multiclass N-grams.

3) Word Clustering for Multiclass Bi-Grams: For clustering, we adopt vectors to represent left- and right-context Markovian dependence, i.e., which words will appear in a left or right context with what probability. These Markovian dependence vectors are defined as follows:

$$v^l(x) = \left[P^b(w_1|x), P^b(w_2|x), \dots, P^b(w_V|x)\right] \quad (8)$$

$$v^r(y) = \left[P^f(w_1|y), P^f(w_2|y), \dots, P^f(w_V|y)\right] \quad (9)$$

where $v^l(x)$ represents the left context Markovian dependence vector of x. This vector is used for left-context class clustering. $P^b(w_i|x)$ is the value of the probability of the backward bi-gram from x to w_i (ith word in the lexicon), while $v^r(y)$ represents the right-context Markovian dependence vector of y. This vector is used for right context class clustering. $P^f(w_i|y)$ is the value of the probability of the forward bi-gram from y to w_i, and V is the size of the vocabulary.

For clustering, the distance between the word pair's vectors is used, since word pairs with similar vectors also have similar Markovian dependence. We use Euclidean distance as a distance measure. Word clustering is, thus, performed in the following manner, called the uni-gram weighted Ward method [17].

4) Use of Frequent Word Successions: Furthermore, multiclass N-grams are extended to multiclass composite N-grams. In this model, higher-order word N-grams are partially introduced by regarding frequent variable-length word sequences as new word succession entries. In this way, for frequent word sequences with length L, an L order word N-gram can be estimated reliably, even if the training corpus size is insufficient to estimate the N-grams of other words. After the introduction of higher-order word N-grams, the increase in parameters only corresponds to a uni-gram of word succession. Therefore, multiclass composite N-grams can maintain a compact model size in multiclass N-grams.

V. Multilingual Machine Translation

Development of a machine translation system requires lots of time and expenditure. If the development is for a multilanguage system, the cost is multiplied by $N^*(N-1)$ for N languages. Therefore, a drastic cost reduction is required. It is well known that spoken languages are different in many aspects from written languages; therefore, porting existing systems for written language is not promising, but new development does pay. Under this background of S2ST, we have decided to adopt a corpus-based approach and have been developing multilingual corpora and machine-learning approaches by using the corpora.

We named our project "Corpus-Centered Computation (C-cube)." C-cube places corpora at the center of the technology. Translation knowledge is extracted from corpora, translation quality is gauged by referring to corpora, the system quality is optimized automatically by gauging, and the corpora themselves are paraphrased or filtered by automated processes. Fig. 4 shows an overview of our machine translation system developed in the C-cube project.

There are two main types in corpus-based machine translation: 1) example-based machine translation (EBMT) [18], [19], and 2) statistical machine translation (SMT) [20]–[26]. C-cube is developing both technologies in parallel and blending them. In this paper, we introduce three different machine translation systems: D-cube, HPAT, and SAT.

1) D-cube (Sentence-based EBMT): This retrieves the most similar example of the input and example sentences by dynamic programming-based matching, and adjusts the gap between the input and the retrieved example by using dictionaries [27]. Most EBMTs use translation examples in

phrase level, while D-cube uses translation examples in sentence level, thus, D-cube realizes very natural translation when it finds a good example in the bilingual corpus, while it suffers from translation coverage.

2) HPAT (Phrase-based EBMT): Based on phrase-aligned bilingual trees, transfer patterns are generated. According to the patterns, the source phrase structure is obtained and converted to generate target sentences [28]. HPAT features a feedback cleaning method that evaluates translation quality automatically and based on that it filters out bad transfer patterns [29], which improves translation quality drastically.
3) SAT (Word-based SMT): SAT deals with Japanese and English on top of a word-based SMT framework. SAT is a developing series of SMT, which includes phrase-based translation [30], chunk-based translation [50], and sentence-based greedy decoding [32].

No single system can achieve complete translation of every input. The translation quality changes sentence-by-sentence and system-by-system. Thus, we could obtain a large increase in accuracy if it were possible to select the best one of different translations for each input sentence. With this idea, a multiengine machine translation approach has been taken by several research groups [33], [34]. In contrast to these previous studies, we utilize language and translation models simultaneously and a multiple comparison test for checking significance [31].

VI. Text-to-Speech Conversion

A block diagram of our speech synthesis module, which is called XIMERA, is shown in Fig. 5. Similar to most concatenative TTS systems, XIMERA is composed of four major modules, i.e., a text processing module, a prosodic parameter generation module, a segment selection module, and a waveform generation module.

The target languages of XIMERA are Japanese and Chinese. Language-dependent modules include the text processing module, acoustic models for prosodic parameter generation, speech corpora, and the cost function for segment selection. The search algorithm for segment selection is also related to the target language via the cost function. XIMERA is currently focused on a normal reading speech style suitable for news reading and emotionless dialogs between man and machine.

The prominent features of XIMERA are as follows:

1) its large corpora (a 110-h corpus of a Japanese male, a 60-h corpus of a Japanese female, and a 20-h corpus of a Chinese female);
2) HMM-based generation of prosodic parameters [35];
3) a cost function for segment selection optimized based on perceptual experiments [36].

VII. BTEC and MAD Corpus Description

ATR has been constructing two different types of corpora in the travel domain: 1) a large-scale multilingual collection of basic sentences that covers many domains [37], and 2) a small-scale bilingual collection of spoken sentences that reflects the characteristics of the spoken dialogs [38]. The former is used to train the multilingual translation component, while the latter is used to link spoken sentences to basic sentences.

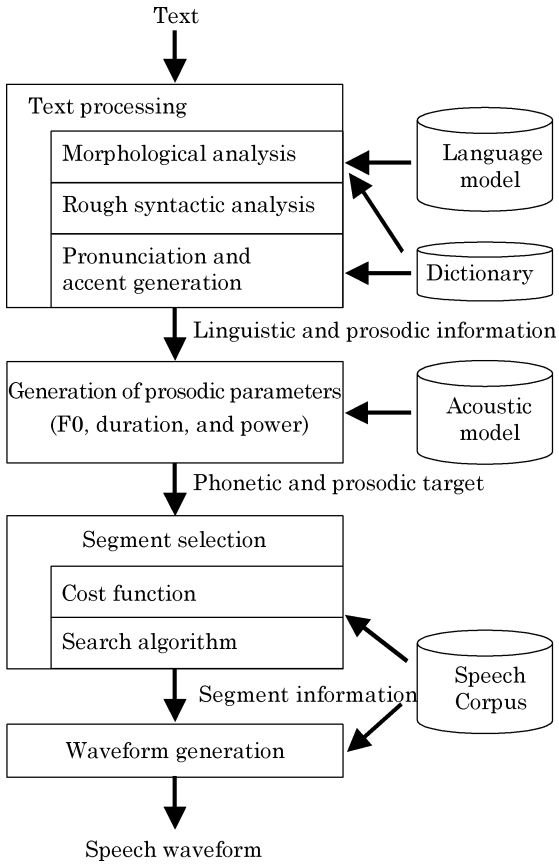


Fig. 5. Block diagram of TTS module.

The BTEC was planned to cover large-scale utterances for every potential subject in travel conversations, together with their translations. Since it is almost infeasible to collect them through transcribing actual conversations or simulated dialogs, we decided to use sentences from the memories of bilingual travel experts. We started by investigating phrasebooks that contain Japanese/English sentence pairs that those experts consider useful for tourists traveling abroad. We collected these sentence pairs and rewrote them to make translations as context-independent as possible and to comply with our transcription style. Sentences outside the travel domain or containing very special meanings were removed. Currently, we have about 420 000 Japanese–English sentence (utterance) pairs in the travel domain. Parts of them have been translated into Chinese, French, Italian, Korean, and Spanish by C-Star partners.

The MT-assisted dialogs (MAD) is a small-scale corpus intended for collecting representative utterances that people will input to S2ST systems. For this purpose, we carried out simulated (i.e., role play) dialogs between two native speakers of different mother tongues with a Japanese/English bidirectional S2ST system, instead of using human interpreters. In order to concentrate on the effects of MT by circumventing communication problems caused by speech recognition errors, we replaced the speech recognition modules with human typists. The resulting system is, thus, considered equivalent to using an S2ST system whose speech recognition part is almost perfect. We employed a combined version of the example-based machine translation system developed in our previous project, TDMT [39] and

D-cube [27] for the MT module (J-to-E and E-to-J) and CHATR [40] for the speech synthesizer.

This environment is somewhere between the "Wizard-of-Oz" (WOZ) approach in Verbmobil [1], which replaced the entire S2ST process with humans, and an approach that relies only on an S2ST system [2]. We have carried out three sets of simulated dialogs so far. The first set (MAD1) is to see whether this approach is feasible with rather simple tasks such as "asking an unknown foreigner where a bus stop is." The second set (MAD2) focused on task achievement with slightly complicated tasks, such as planning a guided tour with travel agents. The third set contains carefully recorded speech data (MAD3).

VIII. Evaluation of the S2ST System

A. Evaluation Data

For evaluation of the speech recognition and translation modules of our S2ST system, we randomly selected 510 Japanese–English sentence pairs from the BTEC corpus and kept them unused for training. These sentences, as well as the corresponding Chinese sentences, were used to collect acoustic speech data. For each language, we recorded data from 20 male and 20 female native speakers. Each speaker's data consists of about 300 read style utterances. The text material was designed in such a way that each of the 510 sentences was read by at least 10 male and 10 female speakers.

For the overall system evaluation, to assess the translation performance of the actual dialogs, an additional test corpus comprising 502 pairs of Japanese and English sentences was used, in which the pairs were randomly selected from the MAD Dialogs.

B. Japanese Speech Recognition

1) Acoustic Model: For acoustic model training, we used the Japanese travel dialogs in "The Travel Arrangement Task" (TRA) of the ATR spontaneous speech database [41]. This corpus consists of role-playing pseudodialogs between a hotel clerk and a customer about room reservations, cancellation, trouble-shooting, etc. We also used 503 phonetically balanced sentences (BLA) read by the same 407 speakers of the TRA. TRA includes about 5 h of speech and BLA includes about 25 h of speech. We do not have actual BTEC speech data for Japanese, but the TRA corpus includes many similar expressions to the BTEC, and the BLA corpus is helpful for creating Japanese standard phoneme models.

The speech analysis conditions were as follows: The frame length was 20 ms and the frame shift was 10 ms; 12-order MFCC, 12-order ΔMFCC, and Δ log power were used as feature parameters. The cepstrum mean subtraction was applied to each utterance. Table I shows the phoneme units for our Japanese ASR. A silence model with three states was built separately from the phoneme models. Three states were used as the initial model for each phoneme. The scaling factors $C_c = 2$, $C_t = 20$ were used for the MDL-SSS. After a topology was obtained by each topology training method, mixture components were increased, and a five Gaussian mixture model was created.

2) Language Model and Decoding: The BTEC corpus and ITLDB database collected during our previous project [41] were used to create language models. In total, there are about 6.2 M words. A word bi-gram model, a word tri-gram model, and a multiclass composite bi-gram model were created. The multiclass composite bi-gram included 4000 classes for each direction. The size of the lexicon was 54 K words, and the number of extracted composite words was 24 K words. For recognition, the gender-dependent acoustic model and the multiclass 2-gram model were used in the first pass, and the word tri-gram model was used to rescore word lattices in the second pass.

3) Performance: For the test data, a subset of the Japanese BTEC test data was used consisting of about 4000 utterances (100 utterances from 20 males and 20 female speakers).

Table II shows the perplexity for each model. The multiclass composite bi-gram obtained the middle performance between the word bi-gram model and the word tri-gram model.

Table III shows the recognition performance represented by word accuracy rates. The model with 2086 states created by the MDL-SSS using $C_c = 2$, $C_t = 20$ obtained almost the same performance as that with 2100 states created by the ML-SSS.

Table IV shows the word accuracy rates by two combinations of language models. For the first-pass search, one used the word bi-gram model, and the other used the multiclass composite (MCC) bi-gram model. Furthermore, both of them used the word tri-gram model for rescoring in the second-pass search. The acoustic model was the same as the MDL-SSS' model with $C_c = 2$, $C_t = 20$ in Table III. The MCC bi-gram model obtained a 17.3% error reduction rate compared to the word bi-gram model, and the combination of the MCC bi-gram model and the word tri-gram model obtained a 9.25% error reduction rate compared to the combination of the word bi-gram model and tri-gram model.

TABLE I
Phoneme Units for Japanese ASR

Vowels	a,i,u,e,o
Consonants	b, ch, d, g, f, h, j, k, m, n, ng, p, q, r, s, sh, t, ts, w, z, zh

TABLE II
Perplexity for Japanese BTEC Test

word 2-gram	word 3-gram	MCC 2-gram
30.64	17.45	24.81

TABLE III
Recognition Performance for Japanese BTEC Test

	#states	WA[%]
ML-SSS	2,100 (max state length = 4)	94.51
MDL-SSS	2,086 ($C_c = 2; C_t = 20$)	94.41

TABLE IV
Word Accuracy Rates [Percent] by Two Combinations of Japanese Language Models

Model	Word 2-gram		MCC 2-gram	
3-gram Rescore	No	Yes	No	Yes
Acc.(%)	91.98	93.84	93.37	94.41

C. English Speech Recognition

1) Acoustic Model: In contrast to the Japanese language system, even similar domain acoustic training data were not available to us at this time. However, as Lefevre *et al.* demonstrated in [42], out-of-domain speech training data do not cause significant degradation of the system performance. In fact, it was found to be more sensitive to the language model domain mismatch. Thus, we choose the Wall Street Journal (WSJ) corpus [43], since we needed a speech database that is large enough and contains clean speech from many speakers. About 37 500 utterances recommended for speaker-independent training (WSJ-284) were selected as the training set for our acoustic model. The total number of speakers is 284 (143 male and 141 female). Feature extraction parameters were the same as for the Japanese language system: 25 dimensional vectors (12 MFCC + 12 Delta MFCC + Delta pow) extracted from 20-ms-long windows with 10-ms shift. First, we trained a model with 1400 states and five mixture components per state using the ML-SSS algorithm. This was a rather small model compared to the other models that have been built on the same data [44], so it was not expected to have high performance. Nevertheless, we regarded it as a starting point for further model development and optimization. Next, we trained several models using the MDL-SSS algorithm where the temporal splitting constant C_t is set to 20 and the contextual splitting constant C_c takes values from 2 to 10. In this way, we obtained models with state numbers ranging from about 1500 to about 7000. Initially, they all had five mixture components per state. The preliminary tests showed that the model with 2009 states was the best and was, therefore, selected for further experiments. Two more versions of this model—with 10 and 15 mixture components per state—were trained as well.

2) Language Model: For the language model training, the BTEC English data was used. Standard bi-gram and tri-gram models were trained as well as one multi-class composite word bi-gram model. The number of classes is 8000, while the number of composite words is about 4000.

3) Pronunciation Dictionary: Although the BTEC task domain is quite broad, there are many travel-oriented words that are not included in publicly available pronunciation dictionaries. Also, there are many specific proper names of sightseeing places, restaurants, travel-related companies, and brand names. A large portion of the task word list represents Japanese words including Japanese first and family names. In total, there were about 2500 such words (≈10% of the 27 K-word dictionary) and to develop good pronunciation variants for them was quite a challenge for us. Especially difficult were the Japanese words because there is no principled way to predict how a native English speaker would pronounce a given Japanese word. This will depend heavily on the speaker's Japanese proficiency with the two extremes of being fluent in Japanese and speaking just couple of widely known words. Therefore, we decided to cover at least these two cases by taking one pronunciation from the Japanese dictionary and converting it to the English phone labels, and generating one pronunciation according to the English phonetic rules. The latter was done by using TTS software "Festival" [45] followed by a manual correction of some of the pronunciations judged as "making no sense."

TABLE V
ENGLISH ACOUSTIC MODEL'S PERFORMANCE COMPARISON

Model	ML-SSS	MDL-SSS			
State #	1400	1578	2009		3028
Mix. #	5	5	5	15	5
Acc.(%)	87.5	88.1	88.5	89.4	88.2

TABLE VI
ENGLISH LANGUAGE MODEL'S PERFORMANCE COMPARISON

Model	Word 2-gram		MCC 2-gram	
3-gram Rescore	No	Yes	No	Yes
Acc.(%)	89.21	92.35	89.63	93.29

TABLE VII
SUBWORD UNITS FOR CHINESE ASR SYSTEM

Unit	Types
Initials	b,p,m,f,d,t,n,l,g,k,h,j,q,x,z, c,s,zh,ch,sh,r
Finals	a,ai,an,ang,ao,e,ei,en,eng,er,i1,i2,i3,ia,ian iang,iao,ie,ing,in,iu,iong,o,ou,u,ua,uai,uang uan,ui,un,uo,ong,v,van,ve,vn

Our phoneme set consists of 44 phonemes, including silence. They are the same as those used in the WSJ corpus official evaluations because in this way we could use its dictionary as a source of pronunciation base-forms. In addition, we could run the WSJ task tests with our model to compare performance.

4) Performance: In the first series of experiments, we evaluated the performance of the several acoustic models we have trained. The test data comprised 1200 utterances from 35 speakers included in the BTEC test set. Small conventional bi-gram and tri-gram language models covering about 25% of the entire text training data were used to speed up the evaluation. The recognition results in terms of word accuracy are given in Table V. As can be seen, the MDL-SSS model with 2009 states and 15 mixture components was the best one; thus, it was used for the next experiments involving different types of language models.

Next, we evaluated the language model's performance. In these experiments, we used 204 utterances taken randomly from the larger BTEC test set. The results are summarized in Table VI.

D. Chinese Speech Recognition

1) Acoustic Model: The basic subword units for the Chinese speech recognition front-end used are the traditional 21 Initials and 37 Finals (see Table VII):

The acoustic model was developed using a well-designed speech database: the ATR Putonghua (ATRPTH) speech database of 2003 [46]. The database has a rich coverage of the triplet Initial/Finals phonetic context, and sufficient samples for each triplet with respect to balanced speaker factors including gender and age.

The phonetically rich sentence set of ATRPTH has 792 sentences. An investigation on the *token coverage rates* has been

TABLE VIII
TOKEN COVERAGE RATES OF DIFFERENT CHINESE SUBWORD UNITS

Unit	792 set	Newspaper	Token Coverage
A	974	1,306	98.81%
B	402	408	99.99%
C	10,906	48,392	70.15%
D	4,653	4,598	99.42%

carried out on one month's total of a daily newspaper for different types of phonetic units. Table VIII shows the results, where the following hold.

- Unit *A*: stands for the tonal syllable.
- Unit *B*: stands for the base syllable without tone discrimination.
- Unit *C*: stands for the normal Initial/Final triplets.
- Unit *D*: stands for the context tying Initial/Final triplets, which are tied based on phonetically articulatory configurations. They are assumed to cover the major variants of each triplet phonetic context [47].

The speakers were chosen to have a balanced coverage of different genders and ages. Each unique triplet has at least 46 tokens in the speech database, guaranteeing a sufficient estimation for each triplet HMM.

During the model estimation, accurate pause segmentation and context dependent modeling [48] were done iteratively to guarantee the model's accuracy and robustness. The HMnet structure was derived through a phonetic decision tree-based maximum likelihood state splitting algorithm. The acoustic feature vector consists of 25 dimensions: 12-dimensional MFCCs, their first-order deltas, and the delta of frame power. The baseline gender dependent HMnets have 1200 states, with five Gaussian mixtures at each state.

2) Language Model: The language model for Chinese ASR also uses the composite multiclass N-gram model. The basic lexicon has 19 191 words, while the text BTEC Chinese corpus contains 200 000 sentences for LM training. After they were segmented and POS tagged, word clustering was investigated based on the right- and left-context Markov dependencies. A normal word-based bi-gram model showed a perplexity of 38.4 for the test set with 1500 sentences. With a clustering of 12 000 word classes, the composite multiclass bigram model showed a perplexity of 34.8 for the same test data. The bigram language model was used to generate a word lattice in the first pass, and a trigram language model with a perplexity of 15.7 was used to rescore the word lattice.

3) Performance: The evaluation data here is the BTEC Chinese language-parallel test data. It includes 11.59 h of speech by 20 females and 20 males. The ages of the speakers range from 18 to 55 years old. All the speakers spoke Chinese Putonghua, with little dialect accent. Table IX shows the gender-dependent, Chinese character-based recognition performances. The total performance is 95.1% for Chinese character accuracy with a real-time factor of 26. The performance degraded to 93.4% when the searching beam was narrowed to obtain a real time factor of 6. An urgent task for the near future is to increase the search speed without harming the recognition performance.

TABLE IX
CHINESE CHARACTER-BASED RECOGNITION PERFORMANCE

Group	Character Corr.	Character Acc.
Male	96.1%	95.7%
Female	95.2%	94.4%
Total	95.7%	95.1%

TABLE X
TRANSLATION QUALITY OF FOUR SYSTEMS FOR BTEC

	SAT	HPAT	D-cube	SELECTOR
A	67.2549	42.5490	63.7255	68.2353
AB	74.7059	63.7255	72.1569	75.8824
ABC	82.5490	79.0196	78.8235	83.5294

E. Machine Translation Evaluation

Training corpora for machine translation systems is BTEC explained in Section VII. The target part of test set sentences is paraphrased into as many as 16 multiple reference translations for each source sentence, for which we utilized automatic evaluation programs.

Translations by four machine translation systems, i.e., SAT, HPAT, D-cube, and the SELECTOR based on them, were shown simultaneously to each of multiple Japanese-to-English professional translators, who were native speakers of English, to keep the evaluation results as consistent as possible. The evaluation was done according to ATR's evaluation standard of four grades, A, B, C, and D. Each translation was finally assigned to the median grade from among its grades from multiple evaluators.

Table X shows the translation quality of Japanese-to-English translations for BTEC. The figures are accumulative percentages of four systems for the quality grade. It is fairly high even for the difficult language pair of Japanese-to-English. In addition, we can see in every grade, A, AB, and ABC, the selector outperforms every single-element machine translation.

The Test of English for International Communication (TOEIC), which is the test for measuring English proficiency of nonnative speakers such as Japanese (http://www.ets.org/toeic/). The total score ranges from 10 (lowest) to 990 (highest).

We proposed a method that estimates the TOEIC score of a speech translation system [49]. The translation of a Japanese-to-English speech translation system is compared with that of a native Japanese speaker whose TOEIC score is known. A regression analysis using the pair wise comparison results (wins/loses) shows the translation capability of the speech translation system.

Our Japanese-to-English translation quality is so high that it achieves a TOEIC score of 750. This is 100 points higher than the average score of a Japanese businessperson in the overseas department of Japanese corporations.

F. Text-To-Speech Evaluation

A perception experiment was conducted in which the naturalness of synthetic speech for XIMERA and ten commercial TTS systems were evaluated. A set of 100 Japanese sentences that were evenly taken from ten genres was processed by the

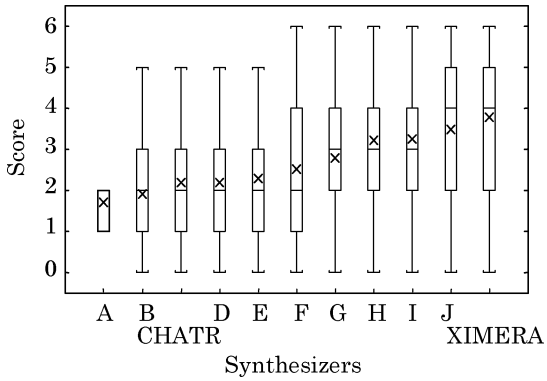


Fig. 6. Result of an evaluation experiment for naturalness between several TTS products.

TABLE XI
TRANSLATION QUALITY OF FOUR SYSTEMS FOR MAD

	SAT	HPAT	D-cube	SELECTOR
A	33.4661	21.5139	31.4741	36.0558
AB	46.8127	40.8367	44.6215	50.1992
ABC	56.7729	60.5578	55.7769	60.9562

TABLE XII
TRANSLATION QUALITY OF FOUR SYSTEMS FOR RECOGNITION RESULTS OF MAD

	SAT	HPAT	D-cube	SELECTOR
A	29.8805	18.1275	27.8884	30.4781
AB	41.4343	35.8566	37.6494	41.8327
ABC	51.7928	52.3904	49.2032	53.9841

11 TTS systems to form a set of stimuli comprising 1100 synthetic speech samples. The stimuli were randomized and presented to 40 listeners through headphones in a quiet meeting room. The listeners rated the naturalness of each stimulus with a seven-point scale, namely, 0 (very bad) to 6 (very good).

Fig. 6 shows the result, in which XIMERA outperforms the other systems. However, the advantage over the second-best system, which is not a corpus-based system, is not substantial, although it is statistically significant.

G. Overall System Evaluation

Tables XI and XII show the quality of the Japanese-to-English machine translation results of MAD with/without SR for four MT systems. The translation quality for MAD is lower than that of BTEC (e.g., selector (ABC): 83.5% for BTEC, 61.0% for MAD without SR). This result reflects the large linguistic difference between BTEC, which is the training corpus for every MT system, and MAD [38]. Regarding the effect of speech recognition, the translation quality for MAD with SR is slightly lower than that for MAD without SR (e.g., selector (ABC): 61.0% for MAD without SR, 54.0% for MAD with SR). The performance of the speech recognition system on MAD data is about 85% word accuracy, which is quite a bit lower than on BTEC test data (see Table IV). This degradation is due to the different speaking style of the MAD speech data. To enhance the translation quality of MAD, we need robust translation techniques designed for actual conversation style sentences, such as paraphrasing from conversation style sentences to BTEC style sentences, and automatic rejection of MT output with problematic translation and/or speech recognition errors.

IX. CONCLUSION

We have developed multilingual corpora and machine-learning algorithms for speech recognition, translation, and speech synthesis. The results have convinced us that our strategy is a viable way to build a high-quality S2ST system.

The current translation system needs improvement in translating longer sentences often found in natural dialogs; therefore, we are studying a method to split a longer sentence into shorter ones and translate them. It is also weak in translating variations often found in natural dialogs; therefore, we are studying a method to normalize a variation in dialog into a stereotypical one found in BTEC by automatic paraphrasing. Finally, a confidence measure for translation is now being pursued and it will be incorporated to reject erroneous translations.

REFERENCES

[1] W. Wahlster, Ed., *Vermobil: Foundations of Speech-to-Speech Translations*. Berlin, Germany: Springer-Verlag, 2000.
[2] E. Costantini, S. Burger, and F. Pianesi, "NESPOLE!'s multi-lingual and multi-modal corpus," in *Proc. LREC*, 2002, pp. 165–170.
[3] A. Lavie, L. Levin, T. Schultz, and A. Waibel. Domain portability in speech-to-speech translation. presented at *Proc. HLT Workshop*. [Online] Available: http://www.is.cs.cmu.edu/papers/speech/HLT2001/HLT_alon.pdf
[4] T. Hirokawa and K. Hakoda, "Segment selection and pitch modification for high quality speech synthesis using waveform segments," in *Proc. Int. Conf. Spoken Language Processing*, 1990, pp. 337–340.
[5] R. Donovan, "Trainable speech synthesis," Ph.D. dissertation, Eng. Dept. Cambridge Univ., Cambridge, U.K., 1996.
[6] A. Breen and P. Jackson, "Nonuniform unit selection and the similarity metric within BT's laureate TTS system," in *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Blue Mountians, Australia, Nov. 1998, p. G.1.
[7] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "ATR ν-talk speech synthesis system," in *Proc. Int. Conf. Spoken Language Processing*, Banff, AB, Canada, Oct. 1992, pp. 483–486.
[8] A. W. Black and P. Taylor, "Chatr: a genetic speech synthesis system," in *Proc. Conf. Computational Linguistics*, Kyoto, Japan, Aug. 1994, pp. 983–986.
[9] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of nonuniform synthesis units," in *Proc. IEEE Int. Conf. Speech, Acoustics, Signal Processing*, New York, Apr. 1988, pp. 679–682.
[10] N. Iwahashi, N. Kaiki, and Y. Sagisaka, "Speech segment selection for concatenative synthesis based on spectral distortion minimization," *Trans. IEICE*, vol. E76-A, no. 11, pp. 1942–1948, Nov. 1993.
[11] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Workshop on Human Language Technology*, 1994, pp. 307–312.
[12] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," in *Proc. ICASSP*, vol. I, 1992, pp. 573–576.
[13] M. Ostendorf and H. Singer, "HMM topology design using maximum likelihood successive state splitting," *Comput. Speech Lang.*, vol. 11, pp. 17–41, 1997.
[14] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of nonuniform context-dependent HMM topologies based on the MDL criterion," in *Proc. Eurospeech*, 2003, pp. 2721–2724.
[15] P. Brown, V. Pietra, P. Souza, J. Lai, and R. Mercer, "Class-based N-gram models of natural language," *Comput. Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.
[16] H. Yamamoto, S. Isogai, and Y. Sagisaka, "Multi-class composite N-gram language model," *Speech Commun.*, vol. 41, pp. 369–379, 2003.

[17] H. Ward, Jr., "Hierarchical grouping to optimize an objective function," *J. Amer. Statist. Assoc.*, vol. 58, pp. 236–244, 1963.
[18] M. Nagao, "A framework of a mechanical translation between Japanese and English by analogy principle," in *Artificial and Human Intelligence*, M. Elithorn and R. Banerji, Eds. Amsterdam, The Netherlands: North-Holland, 1984, pp. 173–180.
[19] H. Somers, "Review article: example-based machine translation," *J. Mach. Translat.*, pp. 113–157, 1999.
[20] P. Brown *et al.*, "A statistical approach to machine translation," *Comput. Linguistics*, vol. 16, pp. 79–85, 1993.
[21] K. Knight, "Automating knowledge acquisition for machine translation," *AI Mag.*, vol. 18, no. 4, pp. 81–96, 1997.
[22] H. Ney, "Stochastic modeling: from pattern classification to language translation," in *Proc. ACL Workshop DDMT*, 2001, pp. 33–37.
[23] H. Alshawi, S. Bangalore, and S. Douglas, "Learning dependency translation models as collections of finite-state head transducers," *Comput. Linguistics*, vol. 26, no. 1, pp. 45–60, 2000.
[24] Y. Wang and A. Waibel, "Fast decoding for statistical machine translation," in *Proc. ICSLP*, 1998, pp. 2775–2778.
[25] F. Och, C. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation," in *Proc. EMNLP/WVLC*, 1999, pp. 20–28.
[26] A. Venugopal, S. Vogel, and A. Waibel, "Effective phrase translation extraction from alignment models," in *Proc. ACL*, 2003, pp. 319–326.
[27] E. Sumita, "Example-based machine translation using DP-matching between word sequences," in *Proc. ACL Workshop DDMT*, 2001, pp. 1–8.
[28] K. Imamura, "Application of translation knowledge acquired by hierarchical phrase alignment," in *Proc. TMI*, 2002, pp. 74–84.
[29] K. Imamura, E. Sumita, and Y. Matsumoto, "Feedback cleaning of machine translation rules using automatic evaluation," in *Proc. 41st Annu. Meeting Assoc. Computational Linguistics*, 2003, pp. 447–454.
[30] T. Watanabe, K. Imamura, and E. Sumita, "Statistical machine translation based on hierarchical phrase alignment," in *Proc. TMI*, 2002, pp. 188–198.
[31] Y. Akiba, T. Watanabe, and E. Sumita, "Using language and translation models to select the best among outputs from multiple MT systems," in *Proc. COLING*, 2002, pp. 8–14.
[32] T. Watanabe and E. Sumita, "Example-based decoding for statistical machine translation," in *Proc. 9th MT Summit*, 2003, pp. 410–417.
[33] C. Hogan and R. Frederking, "An evaluation of the multi-engine MT architecture," in *Proc. AMTA*, 1998, pp. 113–123.
[34] C. Callison-Burch and S. Flournoy, "A program for automatically selecting the best output from multiple machine translation engines," in *Proc. MT-SUMMIT-VIII*, 2001, pp. 63–66.
[35] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *Proc. IEEE Int. Conf. Speech, Acoustics, Signal Processing*, Istanbul, Turkey, Jun. 2000, pp. 1315–1318.
[36] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing subcost functions for segment selection based on perceptual evaluations in concatenative speech synthesis," in *Proc. IEEE Int. Conf. Speech, Acoustics, Signal Processing*, vol. I, Montreal, QC, Canada, Jun. 2004, pp. 657–660.
[37] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bi-lingual corpus for speech translation of travel conversations in the real world," in *Proc. LREC*, 2002, pp. 147–152.
[38] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proc. Eurospeech*, 2003, pp. 381–384.
[39] E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai, "Solutions to problems inherent in spoken language stranslation: the ATR-MATRIX approach," in *Proc. 7th MT Summit*, 1999, pp. 229–235.
[40] N. Campbell, "CHATR: a high-definition speech resequencing system," in *Proc. ASA/JASA Joint Meeting*, 1996, pp. 1223–1228.
[41] T. Takezawa, T. Morimoto, and Y. Sagisaka, "Speech and language databases for speech translation research in ATR," in *Proc. 1st Int. Workshop on East-Asian Language Resources and Evaluation (EALREW)*, 1998, pp. 148–155.
[42] F. Lefevre, J. L. Gauvain, and L. Lamel, "Improving genericity for task-independent speech recognition," in *Proc. Eurospeech*, 2001, pp. 1241–1244.
[43] D. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. DARPA Speech and Natural Language Workshop*, Feb. 1992, pp. 357–362.
[44] *Proc. Spoken Language Technology Workshop*, Plainsboro, NJ, Mar. 1994.
[45] P. Taylor, A. Black, and R. Caley, "The architecture of the Festival speech synthesis system," in *Proc. Third Int. Workshop Speech Synthesis*, Sydney, Australia, Nov. 1998, pp. 147–151.
[46] J. S. Zhang, M. Mizumachi, F. Soong, and S. Nakamura, "An introduction to ATRPTH: a phonetically rich sentence set based Chinese Putonghua speech database developed by ATR," in *Proc. ASJ Meeting*, Fall 2003, pp. 167–168.
[47] J. S. Zhang, S. W. Zhang, Y. Sagisaka, and S. Nakamura, "A hybrid approach to enhance task portability of acoustic models in Chinese speech recognition," in *Proc. Eurospeech*, vol. 3, 2001, pp. 1661–1663.
[48] J. S. Zhang, K. Markov, T. Matsui, and S. Nakamura, "A study on acoustic modeling of pauses for recognizing noisy conversational speech," *Proc. IEICE Trans. Inf. Syst.*, vol. 86-D, no. 3, pp. 489–496, 2003.
[49] F. Sugaya, T. Takezawa, A. Yokoo, Y. Sagisaka, and S. Yamamoto, "Evaluation of the ATR-MATRIX speech translation system with a pair comparison method between the system and humans," in *Proc. ICSLP*, 2000, pp. 1105–1108.
[50] T. Watanabe, E. Sumita, and H. Okuno, "Chunk-based statistical translation," *Proc. ACL*, pp. 303–310, 2003.

Satoshi Nakamura (M'89) was born in Japan on August 4, 1958. He received the B.S. degree in electronic engineering from the Kyoto Institute of Technology, Kyoto, Japan, in 1981 and the Ph.D. degree in information science from Kyoto University, Kyoto, in 1992.

From 1981 to 1993, he was with the Central Research Laboratory, Sharp Corporation, Nara, Japan. From 1986 to 1989, he was with ATR Interpreting Telephony Research Laboratories, Kyoto. From 1994 to 2000, he was an Associate Professor of the Graduate School of Information Science, Nara Institute of Science and Technology. In 1996, he was a Visiting Research Professor of the CAIP Center, Rutgers University, New Brunswick, NJ. He is currently the Head of the Acoustics and Speech Research Department, ATR Spoken Language Translation Laboratories. He also has been an Honorary Professor at the University of Karlsruhe, Karlsruhe, Germany since 2004. His current research interests include speech recognition, speech translation, spoken dialog systems, stochastic modeling of speech, and microphone arrays.

Dr. Nakamura received the Awaya Award from the Acoustical Society of Japan in 1992, and the Interaction 2001 Best Paper Award from the Information Processing Society of Japan in 2001. He served as an Associate Editor for the *Journal of the IEICE Information* from 2000 to 2002 and is currently a member of the Speech Technical Committee of the IEEE Signal Processing Society. He is a member of the Acoustical Society of Japan, Institute of Electrical and Electronics Engineers (IEICE), Iand the nformation Processing Society of Japan.

Konstantin Markov (M'05) was born in Sofia, Bulgaria. received the M.Sc. and Ph.D. degrees in electrical engineering from the Toyohashi University of Technology, Toyohashi City, Japan, in 1996 and 1999, respectively.

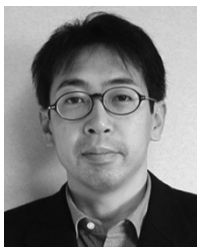
After graduating with honors from the St. Petersburg Technical University, St.Petersburg, Russia, he worked for several years as a Research Engineer at the Communication Industry Research Institute, Sofia. In 1999, he joined the Research Development Department of ATR, Kyoto, Japan, and in 2000 became an Invited Researcher at the ATR Spoken Language Translation (SLT) Research Laboratories. Currently, he is a Senior Research Scientist at the Acoustics and Speech Processing Department, ATR SLT.

Dr. Markov received the Best Student Paper Award from the IEICE Society in 1998. He is a member of ASJ, IEICE, and ISCA. His research interests include signal processing, automatic speech recognition, Bayesian networks, and statistical pattern recognition.

Hiromi Nakaiwa received the M.S. and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1987 and 2002, respectively.

He joined the Nippon Telegraph and Telephone Company (NTT), Tokyo, Japan, in 1987 and Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan, in 2002. He is Head of the Department of Natural Language Processing, ATR Spoken Language Translation Research Laboratories. His research interests include computational linguistics, machine translation, machine learning of NLP rules from corpora, and semantic dictionaries for NLP. From 1995 to 1996, he was a Visiting Researcher at the University of Manchester Institute of Science and Technology (UMIST), Manchester, U.K. He is a coeditor of GoiTaikei—a Japanese Lexicon (in Japanese), an editorial board member of the *Machine Translation Journal*, a board member of the ANLP, and a part-time lecturer at Ibaraki University.

Dr. Nakaiwa was awarded the IPSJ Excellence Award in 1999, the IEEE ICTAI Best Paper Award in 2000, and the TELECOM System Technology Award in 2001. He is a member of the ACL, the IPSJ, the ANLP, and the JSAI.



Genichiro Kikui received the B.E. and M.E. degrees in electrical engineering from Kyoto University, Kyoto, Japan, in 1984 and 1986, respectively.

In 1986, he joined the Machine Translation Research Group in Nippon Telegraph and Telephone Corporation, Kyoto, where he conducted research on machine translation, automatic text revision, and cross-language information retrieval. He made a research stay at the Center for the Study of Language and Information (CSLI), Stanford University, Stanford, CA, from 1997 to 1998. In 2001, he moved to the ATR Spoken Language Translation Research Laboratories, Kyoto, as the Head of Department Two. He is interested in natural language processing, including spoken language modeling, lexicography, and information retrieval.

Mr. Kikui is a member of IPSJ and JSAI.

Hisashi Kawai (M'04) received the B.E, M.E., and D.E. degrees in electronic engineering from the University of Tokyo, Tokyo, Japan, in 1984, 1986, and 1989, respectively.

He joined the Kokusai Denshin Denwa Company, Ltd. in 1989. He has been working for the ATR Spoken Language Translation Research Laboratories since 2000. He has been engaged in the research of speech recognition and speech synthesis. His research interests are in the areas of speech information processing.

Takatoshi Jitsuhiro (M'98) received the B.E. and M.E. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991 and 1993.

In 1993, he joined the Human Interface Laboratories, Nippon Telegraph and Telephone Corporation (NTT), Kanagawa, Japan, and began work on speech recognition. Since 2000, he has been a Researcher at the ATR Spoken Language Translation Research Laboratories, Kyoto, Japan. His research interests include speech recognition and speech signal processing.

Mr. Jitsuhiro is a member of the Acoustical Society of Japan and the IEICE.

Jin-Song Zhang was born in China on October 4, 1968. He received the B.E. degree in electronic engineering from Hefei University of Technology, Hefei, China, in 1989, the M.E. degree from the University of Science and Technology of China (USTC), Hefei, in 1992, and the Ph.D. degree from the University of Tokyo, Yokyo, Japan, in 2000.

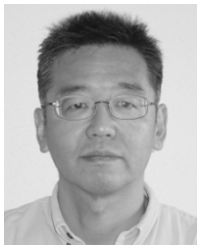
From 1992 to 1996, he was a Teaching Assistant and Lecturer in the Department of electronic Engineering, USTC. Since 2000, he has been at the ATR Spoken Language Translation Research Laboratories, Kyoto, Japan, as a Researcher. His main research interests include speech recognition, prosody information processing, and speech synthesis.

Dr. Zhang is a member of the Acoustic Society of Japan.

Hirofumi Yamamoto received the B.E. and M.E. degrees from Tokyo University, Tokyo, Japan, in 1979 and 1981 respectively, and the Ph.D. degree from Waseda University, Tokyo, in 2004.

In 1981, he joined the CSK Company. In 1996, he began working for ATR Interpreting Telecommunication Laboratories, Kyoto, Japan. Currently, he is a Senior Researcher at the ATR Spoken Language Translation Research Laboratories. His research interests include speech recognition and natural language processing.

Dr. Yamamoto is a member of ASJ, IEICE, and ANLP.



Eiichiro Sumita received the M.S. degree in computer science from the University of Electro-Communications, Tokyo, Japan, in 1982 and the Ph.D. degree in engineering from Kyoto University, Kyoto, Japan, in 1999.

His research interests include natural language processing (machine translation, paraphrasing, and summarization), spoken-language processing, information retrieval, e-Learning, and parallel processing. He is currently leading a project on machine translation, named Corpus-Centered Computation (C3). He serves as an Associate Editor of the *ACM Transactions On Speech and Language Processing*.

Dr. Sumita is a member of the ACL, the IEICE, the IPSJ, the ASJ, and the ANLP.

Seiichi Yamamoto (F'04) graduated from Osaka University in 1972 and received the M.S. and Ph.D. degrees from Osaka University, Osaka, Japan, in 1974 and 1983, respectively.

He joined the Kokusai Denshin Denwa Company, Ltd. in April 1974, and the ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, in May 1997. He was appointed President of ATR-ITL in 1997. He is currently the Director of ATR Spoken Language Translation Research Laboratories. His research interests include digital signal processing, speech recognition, speech synthesis, and natural language processing. He is an Invited Professor of Kobe University, Kobe, Japan.

Dr. Yamamoto received Technology Development Awards from the Acoustical Society of Japan in 1995 and 1997. He is Vice president of the ISS Society of IEICE and a board member of the Acoustical Society of Japan and the Institute of Systems, Control and Information Engineers.