Evaluation of Direct Speech Translation Method Using Inductive Learning for Conversations in the Travel Domain

Koji MURAKAMI
Makoto HIROSHIGE
Kenji ARAKI
Graduate school of Engineering
Hokkaido University, Japan
{mura, hiro, araki}@media.eng.hokudai.ac.jp

Koji TOCHINAI
Graduate school of Business Administration
Hokkai Gakuen University, Japan
tochinai@econ.hokkai-s-u.ac.jp

Abstract

This paper evaluates a direct speech translation Method with waveforms using the Inductive Learning method for short conversation. The method is able to work without conventional speech recognition and speech synthesis because syntactic expressions are not needed for translation in the proposed method. We focus only on acoustic characteristics of speech waveforms of source and target languages without obtaining character strings from utterances. This speech translation method can be utilized for any language because the system has no processing dependent on an individual character of a specific language. Therefore, we can utilize the speech of a handicapped person who is not able to be treated by conventional speech recognition systems, because we do not need to segment the speech into phonemes, syllables, or words to realize speech translation. Our method is realized by learning translation rules that have acoustic correspondence between two languages inductively. In this paper, we deal with a translation between Japanese and English.

1 Introduction

Speech is the most common means of communication for us because the information contained in speech is sufficient to play a fundamental role in conversation. Thus, it is much better that the processing deals with speech directly. However, conventional approaches of speech translation need a text result, obtained by speech recognition, for machine translation although several errors or unrecognized portions may be included in the result.

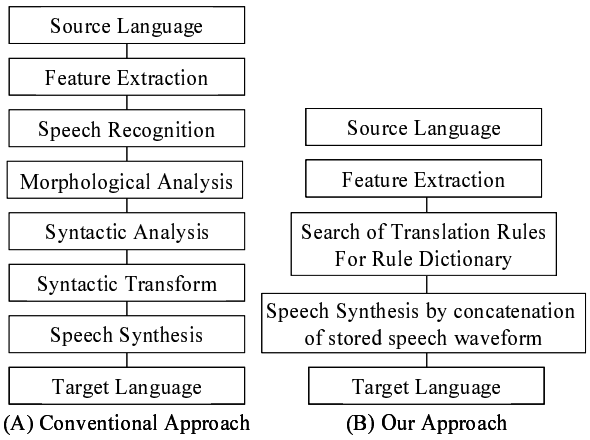

Figure 1: Comparison of conventional and our approach.

A text is translated through morphological analysis, syntactic analysis, and parsing of the sentence of the target language. Finally, the speech synthesis stage produces speech output of the target language. Figure 1(A) shows the whole procedure of a traditional speech translation approach.

The procedure has several complicated processes that do not give satisfying results. Therefore, the lack of accuracy in each stage culminates into a poor final result. For example, character strings obtained by speech recognition may represent different infor-

Proceedings of the Workshop on Speech-to-Speech Translation:
Algorithms and Systems, Philadelphia, July 2002, pp. 45-52.
Association for Computational Linguistics.

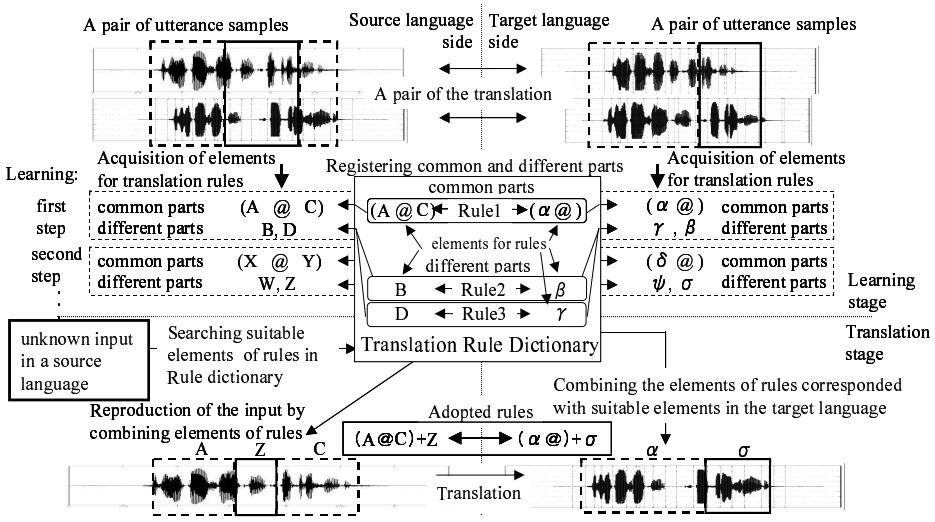

Figure 2: Processing structure.

mation than the original speech.

Murakami et al.(1997) attempted to recognize several vowels and consonants using Neural Networks that had different structures with TDNN (ATR Lab., 1995), however, they could not obtain a high accuracy of recognition. They confirmed that distinguishing the boundaries of words, syllables, or phonemes is a task of great difficulty. Then, they only focused on speech waveform itself, not character strings obtained by speech recognition to realize speech translation. Murakami et.al decided on dealing with the correspondence of acoustic characteristics of speech waveform instead of character strings between two utterances.

Our approach handles the acoustic characteristics of speech without lexical expression through a much simpler structure than the reports of Takizawa et al.(1998) , Müller et al.(1999) or Lavie et al.(1997) because we believe that simplification of the system would prevent inaccuracies in the translation. Figure 1(B) shows the processing stages of our approach. If speech translation can be realized by analyzing the correspondence in character strings obtained by speech recognition, we can also build up speech translation by dealing with the correspondence in acoustic characteristics. In our method, we extract acoustic common parts and different parts by comparing two examples of acoustic characteristics of speech between two translation pairs within the same language. Then we generate translation rules and register them in a translation dictionary. The rules also have the location information of acquired parts for speech synthesis on time-domain. The translation rules are acquired not only by comparing speech utterances but also using the Inductive Learning Method (K. Araki et al., 2001), still keeping acoustic information within the rules. Deciding the correspondence of meaning between two languages is a unique condition to realize our method. In a translation phase, when an unknown utterance of a source language is applied to be translated, the system compares this sentence with all acoustic information of all rules within the source language. Then several matched rules are utilized and referred to their corresponding parts of the target language. Finally, we obtain roughly synthesized target speech by simply concatenating several suitable parts of rules in the target language according to the information of location. Figure 2 shows an overview of the processing structure of our method.

Our method has several advantages over other approaches. First, the performance of the translation is not affected by the lack of accuracy in speech recognition because we do not need the segmentation of speech into words, syllables, or phonemes. Therefore, our method can be applied for all languages without having to make processing changes in the machine translation stage because there is no processing dependent on any specific language. With conventional methods, several processes in the machine translation stage must be altered if the target language is to be changed because morpholog-

ical analysis and syntactic analysis are dependent on each individual character of language completely.

Any difference in language has no affect on the ability of the proposed method, fundamentally because we focus on the acoustic characteristics of speech, not on the character strings of languages. It is very important to approach speech translation with a new methodology that is independent of individual characters of any language.

We also expect our approach can be utilized in speech recuperation systems for people with a speech impediment because our method is able to deal with various types of speech that is not able to be treated by conventional speech recognition systems for normal voice.

Murakami et al.(2002) have successfully obtained several samples of translation by applying our method using local recorded speech data and spontaneous conversation speech.

In this paper, we adopt speech data of travel conversations to the proposed method. We evaluate the performance of the method through experiments and offer discussion on behaviors of the system.

2 Speech processing

2.1 Speech data

It is necessary to extract time-varying spectral characteristics in utterances and apply them to the system. We used several conversation sets from an English conversation book (GEOS Publishing Inc., 1999). The Japanese speech data was recorded with a 48kHz sampling rate on DAT, and downsampled to 8kHz. All speech data in the source language was spoken by Japanese male students of our laboratory. The speech data was spoken by 2 people in the source and target languages, respectively.

The content of the data sets consists of conversations between a client and the front desk at a hotel and conversations between a client and train station staff.

Table 1: Experimental conditions of speech processing.

Size of frame	30msec
Frame cycle	10msec
Speech window	Hamming Window
AR Order	14

2.2 Spectral characteristics of speech

In our approach, the acoustic characteristics of speech are very important because we must find common and different acoustic parts by comparing them. It is assumed that acoustic characteristics are not dependent on any language. Table 1 shows the conditions for speech analysis. The same conditions and the same kind of characteristic parameters of speech are used throughout the experiments.

In this report, the LPC coefficients are applied as spectral parameters because Murakami et al.(2002) could obtain better results by using these parameters than other representations of speech characteristics.

2.3 Searching for the start point of parts between utterances

When speech samples were being compared, we had to consider how to normalize the elasticity on time-domain. Many methods were investigated to resolve this problem. We tried meditating a method that is able to obtain a result similar to dynamic programming (H. Sakoe et al., 1978; H. F. Silverman et al., 1990) to execute time-domain normalization. We adopted a method to investigate the difference between two characteristic vectors of speech samples for determining common and different acoustic parts. The Least-Squares Distance Method was adopted for the calculation of the similarity between these vectors.

Two sequences of characteristic vectors named "test vector" and "reference vector" are prepared. The "test vector" is picked out from the test speech by a window that has definite length. At the time, the

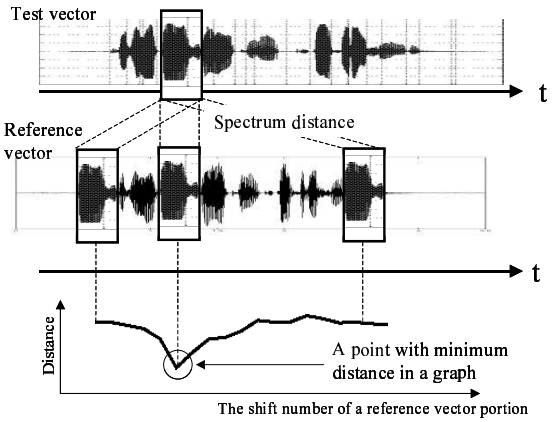

Figure 3: Comparison of vector sequences.

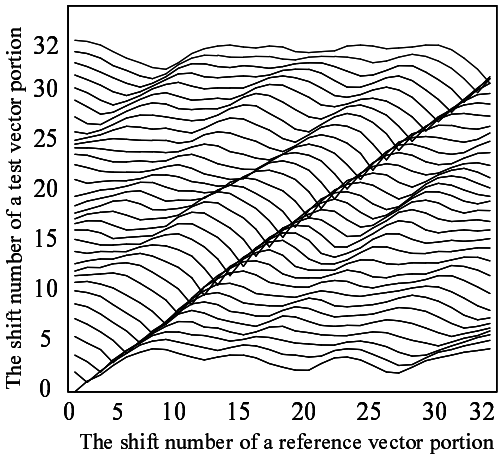

Figure 4: Difference between utterances(1): "All right, Mr. Brown."

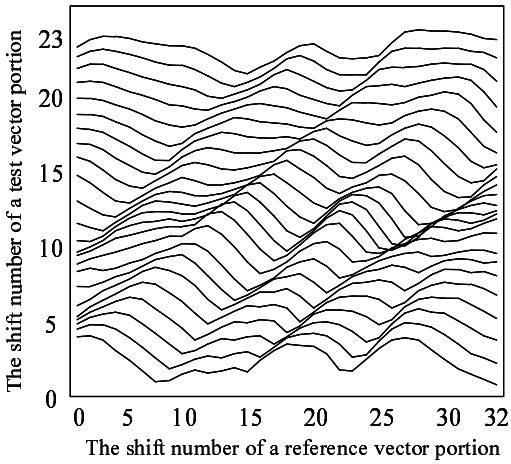

Figure 5: Difference between utterances(2):"All right, Mr. Brown." - "Good afternoon."

"reference vector" is also prepared from the reference speech. A distance value is calculated by comparing the present "test vector" and a portion of the "reference vector". Then, we repeat the calculation between the current "test vector" and all portions of the "reference vector" picked out and shifted in each moment with constant interval on time-domain. When a portion of the "reference vector" reaches the end of the whole reference vector, a sequence of distance values is obtained as a result. The procedure of comparing two vectors is shown as Figure 3. Next, the new "test vector" is picked out by the constant interval, then the calculation mentioned above is repeated until the end of the "test vector". Finally, we should get several distance curves as the result between two speech samples.

Figure 4 and Figure 5 show examples of the difference between two utterances. These applied speech samples are spoken by the same speaker. The contents of the compared utterances are the same in Figure 4, and are quite different in Figure 5. The horizontal axis shows the shift number of reference vector on time-domain and the vertical axis shows the shift number of test vector, i.e., the portion of test speech. In the figures, a curve in the lowest location has been drawn by comparing the top of the test speech and whole reference speech. If a distance value in a distance curve is obviously lowest than other distance values, it means that the two vectors have much acoustic similarity.

As shown in Figure 5, the obvious local minimum distance point is not discovered even if there is the lowest point in each distance curve. On the other hand, as shown in Figure 4, when the test and reference speech have the same content, the minimum distance values are found sequentially in distance curves. According to these results, if there is a position of the obviously smallest distance point in a distance curve, that portion should be regarded as a "common part". Moreover, if these points sequentially appear among several distance curves, they will be considered a common part. At the time, there is a possibility that the part corresponds to several semantic segments, longer than a phoneme and a syllable.

2.4 Evaluation of the obvious minimal distance value

To determine that the obviously lowest distance value in the distance curve is a common part, we adopt a threshold calculated by statistical information. We calculate the variance of distance values shown as σ and the mean value within the curve. The threshold is conducted as $\theta = 4\sigma^2$ from the equation of the Gaussian distribution and the standardized normal distribution.

A point of the smallest distance value within a curve is represented by x and a parameter m shows the mean value of distances. A common part is detected if $(x - m)^2 > \theta$, because the portion of reference speech has much similarity with the "test vector" of the distance curve in a point, and that common part is represented by "0". Otherwise the speech portion for "test vector" is regarded as a dif-

ferent part and represented by "1". If several common parts are decided continuously, we deal with them as one common part, and the first point in that part will be the start point finally. In our method, the acoustic similarities evaluated by several calculations are only the factor for judgment in classifying common or different parts in the speech samples.

3 Generation and application of translation rule

3.1 Correction of acquired parts

The two reference speech samples are divided into several common and different parts by comparison. However, there is a possibility that these parts include several errors of elasticity normalization because the distance calculation is not perfect to resolve this problem on time-domain. We attempt to correct incomplete common and different parts using heuristic techniques when a common part is divided by a discrete different part, or a different part is divided by a discrete common part.

3.2 Acquisition of translation rules

Common and different parts corrected in 3.1 are applied to determine the rule elements needed to generate translation rules. Figure 6 and 7 show the results of comparing utterances. In the first case, a part containing continuous values of "0" represents a common part. In the second case, a part consisting of only "1" is regarded as a different part. In Figure 6, two utterances are calculated as a long common part. On the contrary, two utterances are calculated as a long different part in Figure 7. These results are comparable with lexical contents because the syntactic sentence structures are the same in both cases.

Moreover, when a sentence structure includes common and different parts at the same time, we can treat this structure as a third case. We deal with these three cases of sentence structure as rule types. In all the above-mentioned cases, several sets of common and different parts are acquired if those utterances were almost matching or did not match at all. Combining sets of common parts of the source and target languages become elements of the translation rules for its generation. At this time, the set of common parts extracted from the source language, that have

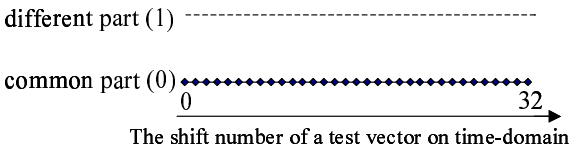

Figure 6: Common and different parts(1):"All right, Mr. Brown."

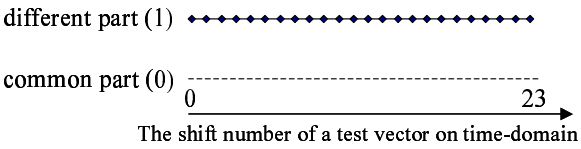

Figure 7: Common and different parts(2). "All right, Mr. Brown." - "Good afternoon."

a correspondence of meaning with a set of common parts in target language, are kept. The sets of different parts become elements of the translation rules as well.

Finally, these translation rules are generated by completing all elements as below. It is very important the rules are acquired if the types of sentences in both languages are the same. When the types of sentence structures are different, it is impossible that translation rules are obtained and registered in the rule dictionary because we can not decide the correspondence between two languages samples uniquely. Acquired rules are categorized in the following types:

Rule type 1: those with a very high sentence similarity

Rule type 2: those with sentences including common and different parts

Rule type 3: those with very low sentence similarity

When a new rule containing the information of several common parts is generated, the rule should keep the sentence form so that different parts in the speech sample are replaced as variables. Information that a translation rule has are as follows:

- rule types as mentioned above
- index number of a source language's utterance
- sets of start and end points of each common and different part

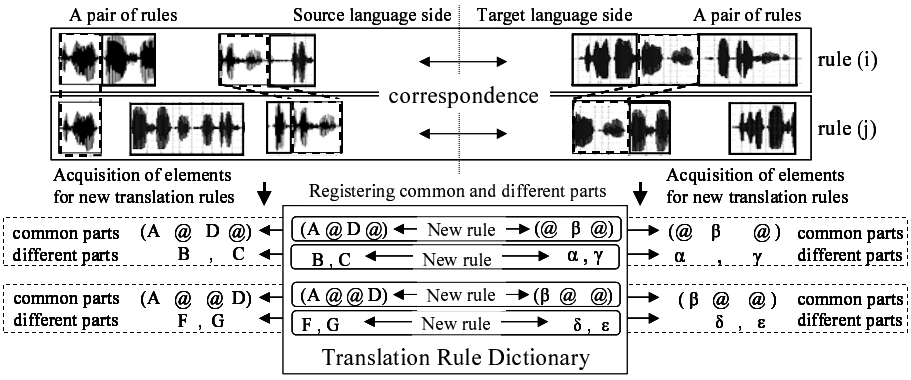

Figure 8: Rule aquisition using the Inductive Learning Method

- index number of an utterance in the target language

3.3 Translation and speech synthesis

When an unknown speech utterance of a source language is adapted to get the result of translation, acoustic information of acquired parts in the translation rules are compared in turn with the unknown speech, and several matched rules become the candidates to translate. The inputted utterance should be reproduced by a combination of several candidates of rules. Then, the corresponding parts of the target language in candidate rules are referred to obtain translated speech. Although the final synthesized target speech may be produced roughly, speech can directly be concatenated by several suitable parts of rules in the target language using the location information on time-domain in rules.

4 The Inductive Learning Method

The Inductive Learning that Araki et al.(2001) proposed acquires rules by extracting common and different parts through the comparison between two samples. This method is designed from an assumption that a human being is able to find out common and different parts between two samples although these are unknown. The method is also able to obtain rules by repetition of the acquired rules registered in the rule dictionary.

Figure 8 shows an overview of recursive rule acquisition by this learning method. Two rules acquired as rule(i) and rule(j) are prepared and compared to extract common and different acoustic parts as well as comparisons between speech samples. Then, these obtained parts are designed as new rules. If the compared rules consist of several common or different parts, the calculation is repeated within each part. It is assumed that these new rules are much more reliable for translation.

If several rules are not useful for translation, they will be eliminated by generalizing the rule dictionary optimally to keep a designed size of memory. The ability of optimal generalization in the Inductive Learning Method is an advantage, as less examples have to be prepared beforehand. Much sample data is needed to acquire many suitable rules with conventional approaches.

5 Evaluation Experiments

5.1 Experiments of rule acquisition

All data in experiments are achieved through several speech processes explained in 2.1. Table 2 shows the conditions for experiments. The parameters concerning frame settings have been decided from the results of several preliminary experiments for rule acquisition.

Table 2: Conditions for experiments.

Frame length of test vector	400msec
Frame rate of both vectors	50msec
The rate of agreement for adopting rules	95%

Table 3: Translation rules.

Set of data	Utterances	Registed rules
Hotel	50	8,500
Station	32	22,846

Table 4: Appropriately acquired parts with correspondence.

Sentence ID	Rule Type	Corresponded Part/Length	Speech
ja110g	common	(22-40)/41	*SOREDEWA, BRAUN-SAMA.*
ja110t	common	(106-124)/128	*KOCHIRANI GOKICYOWO ONEGAIITASHIMASU, BRAUN-SAMA.*
en110g	common	(17-32)/33	All right, Mr. Brown.
en110t	common	(57-69)/71	Please fill out this form, Mr.Brown.

Many sets of common and different parts were extracted by comparing acoustic characteristics of speech in each language, and translation rules were registered in the translation rule dictionary. Table 3 shows the number of speech utterances and registered translation rules between two languages.

5.2 Experimental results of translation

If an unknown speech utterance of a source language can be replaced with acoustic information from rules in the dictionary, the speech will be translated and synthesized roughly without losing it's meaning. Each matched rule includes certain equivalent correspondence parts of the target language. The system needs to decide the most suitable candidates of rules from the rule dictionary for each translation. If the level of similarity between the whole applied unknown speech and all parts of the rules is higher than a rate of agreement as in Table 2, the rules that include appropriate parts can become candidates for current translation.

82 utterances of limited domain have been applied to the system for translation. Regretfully, we could not obtain any complete translated utterances, although several samples have been incompletely translated by adapting translation rules.

5.3 Discussion

We have to investigate several sources of the experimental results. The first cause of the failure in the translation can be found in speech data utilized in these experiments. The contents of these utterances do not exactly include the same expression because

Table 5: Failures of rule acquisition.

	whole rule acquisition	the case of the same content
The number of failure	527	22

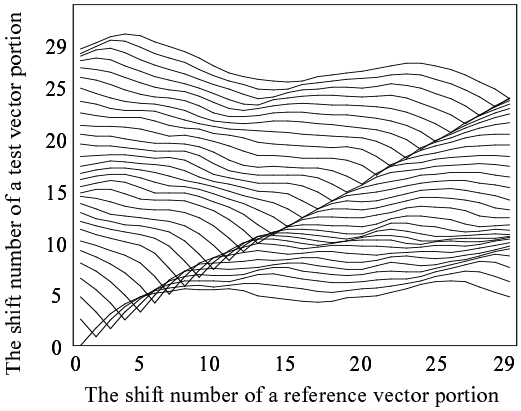

Figure 9: Difference between utterances: "Good afternoon."

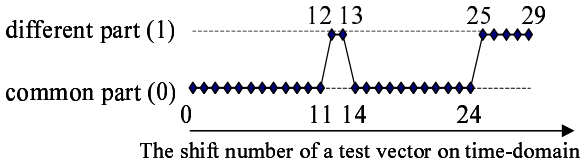

Figure 10: A failed result of parts extraction:"Good afternoon."

contents of speech samples are prepared with various ways of speaking even if the semantic information is the same among them.

Moreover, a small amount of speech data also is another factor because more translation rules should be acquired and adapted for translation.

The system has performed the task because many suitable rules are registered in the rule dictionary. A sample of parts acquired properly is shown as Table 4. In this table, Japanese words are expressed with an italic font. These parts are successfully acquired through the learning stage, so that many suitable rules can be applied to other unknown speech utterances.

Therefore, we need to increase the number of speech samples to obtain more translation rules, and it is also necessary to consider the contents of utter-

ances for more effective rule acquisition and application.

In addition, we have paid attention to the parts themselves acquired as translation rules. We have to consider several causes where the same type of sentences is not determined correctly even when the contents are the same. Table 5 shows the number of failures in whole rule acquisition and in the case of comparisons of the same utterances. The types of sentences are determined by the results of the parts extraction stage. In this stage, thresholds have a much important role for deciding common and different parts. Figure 9 shows the distance curves of the same utterances that were not determined as a common part by a threshold. And Figure 10 shows the result of the extraction of common and different parts. Several minimum points of distance curves have been determined as different parts by threshold although two portions of utterances also have the highest similarity in these points. This kind of failure means that the definition of the threshold has a problem. Therefore, the definition of the threshold needs to be reconsidered for extracting common and different parts much more correctly.

6 Conclusion and future works

In this paper, we have described the proposed method and have evaluated the translation performance for conversations on travel English. We have confirmed that much appropriate acoustic information is extracted by comparing speech, and rules have been generated even if no target speech was obtained through the system.

Many rules have been decided as candidates for each translation by calculating all registered rules with a high calculation cost. Therefore, we will need to apply a method for selecting most suitable rules from candidates and a clustering algorithm to decrease the number of registered rules and the calculation cost.

We will consider adopting a new approach for realizing a more effective threshold without statistical information.

We will also consider a possibility of the direct speech translation system from speech by a person with a handicap in the speech production organ to normal speech because conventional speech recognition methods are not able to assist those with a speech impediment.

Acknowledgement This work is partially supported by the Grants from the Government subsidy for aiding scientific researches (No.14658097) of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates and M. Gavaldà. 1997. *Janus-iii: Speech-to-speech translation in multiple languages*. In *Proceedings of ICASSP '94*, pages 99–102.

ATR Lab. 1995. *Application of Neural Network*.

GEOS Publishing Inc., 1999. *English for Salespeople*.

H. F. Silverman and D. P. Morgan. 1990. *The application of dynamic programming to connected speech recognition*. In *IEEE, ASSP Magazine*, pages 6–25.

H. Sakoe and S. Chiba. 1978. *Dynamic programming algorithm optimization for spoken word recognition*. In *IEEE, Trans. on ASSP*, pages 43–49.

J. Müller and H. Stahl. 1999. *Speech understanding and speech translation by maximum a-posteriori semantic decoding*. In *Proceedings of Artificial Intelligence in Engineering*, pages 373–384.

K. Araki and K. Tochinai. 2001. *Effectiveness of natural language processing method using inductive learning*. In *Artificial Intelligence and Soft Computing(ASC)'01*, pages 295–300.

K. Murakami, M. Hiroshige, K. Araki and K. Tochinai. 2002. *Evaluation of rule acquisition for a new speech translation method with waveforms using inductive learning*. In *Proceedings of Applied Informatics '02*, pages 288–293.

K. Murakami, M. Hiroshige, K. Araki and K. Tochinai. 2002. *Behaviors and problem of the speech machine translation system for various speechdata*. In *Proceedings of the 2002 spring meeting of the ASJ*, pages 385–386.

K. Murakami, M. Hiroshige, Y. Miyanaga and K. Tochinai. 1997. *A prototype system for continuous speech recognition using group training based on neural network*. In *Proc. ITC-CSCC '97*, pages 1013–1023.

T. Takizawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo and S. Yamamoto. 1998. *A Japanese-to-English speech translation system:atr-matrix*. In *Proc. of ICSLP '98*, pages 2779–2782.