EVALUATION OF THE RULE ACQUISITION ON A DIRECT SPEECH TRANSLATION METHOD WITH WAVEFORMS USING INDUCTIVE LEARNING FOR NOUNS AND NOUN PHRASES

KOJI MURAKAMI, KENJI ARAKI, MAKOTO HIROSHIGE AND KOJI TOCHINAI

Graduate School of Engineering, Hokkaido University, JAPAN

This paper evaluates a direct speech translation method with waveforms using Inductive Learning Method for nouns and noun phrases extracted from real conversations. The proposed method treats only acoustic characteristics of speech waveforms of source and target languages without obtaining character strings from speech utterances. And our method is realized by learning translation rules that have acoustic correspondence between two languages recursively. Therefore, we are able to avoid some serious problems of conventional speech recognition and speech synthesis because syntactic expressions are not needed for translation. 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 to realize speech translation. In this paper, we deal with a translation between Japanese and English.

Key words: Speech translation, Inductive learning, Acoustic characteristics of speech waveforms, Translation rules

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 [Müller and Stahl1999]. Thus, it is much better that the processing deals with speech to understand speaker's intension directly. However, as conventional speech translation approaches require a text result, obtained by speech recognition for machine translation stage, several errors or unrecognized portions may be included in the result.

A text is usually 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 procedure of a conventional speech translation approach. The procedure has several complicated processes that do not provide 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 information than the original speech.

We confirmed that distinguishing the boundaries of words, syllables, or phonemes is a task of great difficulty by the results of speech recognition [Murakami et al.1997]. Then, we only focused on speech waveform itself, not character strings obtained by speech recognition to realize speech translation, and 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 that of other approaches [Takezawa et al.1998, Metze et al.2002] Figure 1(B) shows the processing stages of our approaches. 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 trans-

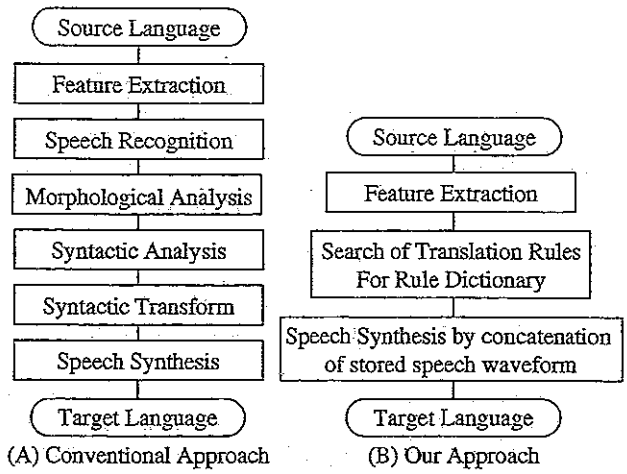

FIGURE 1. Comparison of a conventional and our proposed approach.

lation 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 Inductive Learning Method [Araki and Tochinai2001], still keeping acoustic information within the rules. Deciding the correspondence of meaning between two languages is the unique condition necessary 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 the 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. In conventional speech recognition, language and acoustic models have to be prepared for high accuracy even if these models are dependent on individual language. However, our proposed method does not need these models to be realized. Therefore, our method can be applied for all languages without changing any processing because there is no processing dependent on any specific language. We have successfully obtained several samples of translation by applying our method using local recorded speech data [Murakami et al.2002b]. However, the difficulty of translation without lexical information becomes apparent when applying our system to complicated speech [Murakami et al.2002a].

In this paper, we adopt the speech data of nouns and noun phrases, extracted from spontaneous conversations, to the proposed method. We chose nouns and noun phrases because not only do they appear frequently in speech but their actual number is increasing with the growth of language. We evaluate the effectiveness of the translation rules through experiments and offer discussion on behaviors of the system.

2. SPEECH PROCESSING

2.1. Speech data and Spectral characteristics

It is necessary to extract spectral characteristics in utterances and apply them to the system. The nouns and noun phrases, used throughout the experiments, were manually

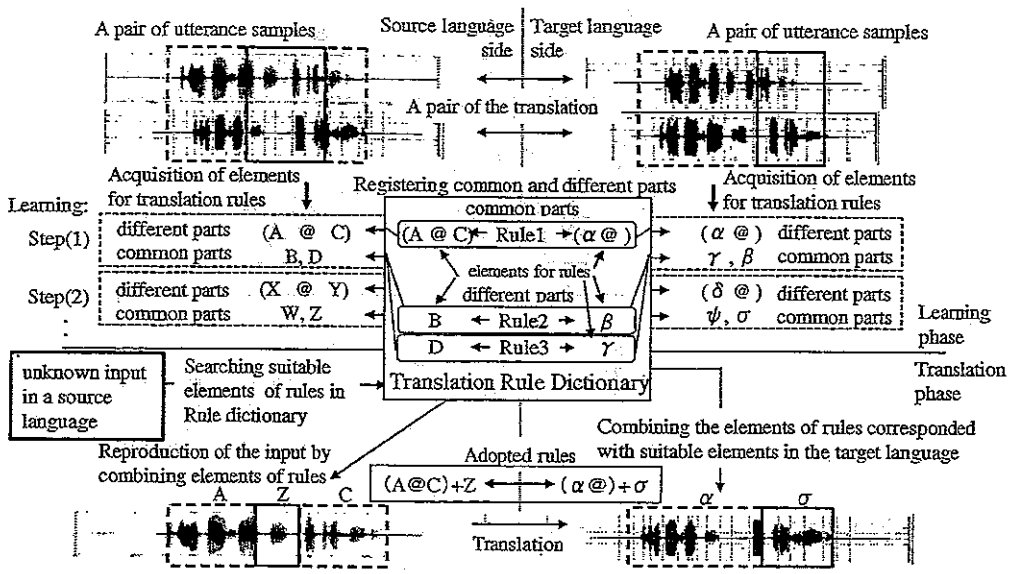

FIGURE 2. Processing structure.

extracted from conversations in a speech and language database (SLDB) [Morimoto et al.] for evaluating the experiments of translation. The content of original data sets consists of conversations between a client and the front desk of a hotel.

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 condition and the same kind of characteristic parameters of speech are used throughout the experiment. In this report, the LPC Cepstrum coefficients are applied as spectral parameters because we could obtain better results by using these parameters than other representations of speech characteristics [Murakami et al.2002a].

2.2. 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. We meditated upon suitable methods that would be able to give a result similar to dynamic programming [Silverman and Morgan1990] to execute time-domain normalization. The method computes the similarity between two characteristic vectors of speech samples by Least-Squares Distance Method for determining common and different

TABLE 1. Experimental conditions of speech processing.

Sampling rate	16kHz
Speaker	2 males and 2 females
Size of frame	30msec
Frame cycle	15msec
Speech window	Hamming Window
AR Order	12
Cepstrum Order	20

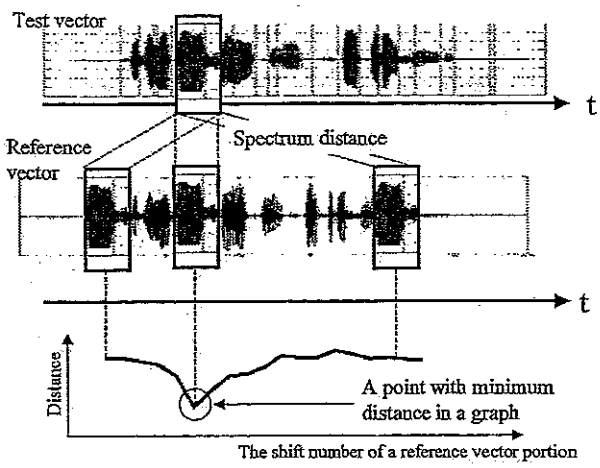

FIGURE 3. Comparison of vector sequences.

acoustic parts.

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 "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" that are 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 in 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 shows 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(A), and are quite different in Figure 4(B). 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 head of the test speech and whole reference speech. If a distance value in a distance curve is obviously lower than other distance values, it means that the two vectors have much acoustic similarity.

As shown in Figure 4(B), 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(A), 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 point should be regarded as a frame in the "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.

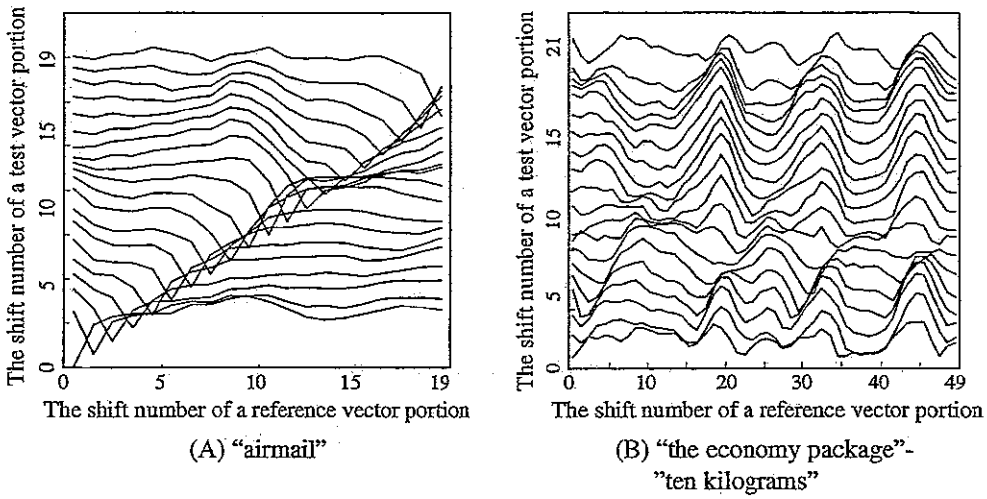

FIGURE 4. Difference between utterances

2.3. Evaluation of the obvious minimal distance value

To determine that the obviously lowest distance value in the distance curve is a common part, we need to employ a method based on a criterion. We have adopted thresholds calculated by statistical information [Murakami et al.2002b] for evaluating the experiments of the proposed method and several translated utterances have been satisfyingly obtained as results. However, a problem in the definition of the thresholds was found in that the thresholds might be inadequately calculated by distance values in each curve when the contents of utterances were the same [Murakami et al.2002a]. Of the 22 calculations 7 appeared to be incorrect because when the threshold was determined, the variance of distance values within the curve was also very influential on the calculation.

Therefore, another method was considered for partitioning acoustic common and different parts more correctly without the thresholds based on statistical information to avoid this problem. We chose the angles formed where two lines cross each other. These lines are linked by two distance points. Figure 5 shows angle calculation within the distance curve. The angles θ_i are determined by a focused point(i) and two neighboring distance points($i-1, i+1$) excluding the start and end points of the distance curve. If these two points have the minimal distance in a distance curve, we evaluate them with heuristic techniques that will be mentioned in 4.1. A common part is detected if the lowest degree of an angle θ_i is formed by a minimal distance point(i) and its neighboring points($i-1, i+1$), because the portion of

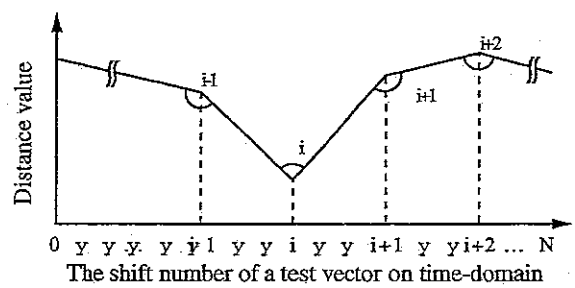

FIGURE 5. Angle calculation.

reference speech has much similarity with the "test vector" of the distance curve at a point.

If several common parts are decided continuously, we deal with them as one common part because we want to partition the utterance into few parts, and the first point in this part will be the start point finally. In our method, the acoustic similarities evaluated by several calculations based on Least-Square Distance Method are the only factor for judgment in classifying common or different parts in the speech samples.

3. INDUCTIVE LEARNING METHOD

Inductive Learning Method [Araki and Tochinai2001] acquires rules by extracting common and different parts through the comparison between two examples. This method is designed from an assumption that a human being is able to find out common and different parts between two examples although those contents are unknown. The method is also able to obtain rules by repetition of the acquired rules registered in the rule dictionary. At the time, a sentence form rule is also acquired with a different part rule by each comparison of utterances, and registered in the dictionary. This type of rule consists of different parts and a common part which is replaced with the variable. Therefore, the rule should represent the abstracted sentence without losing its meaning and be able to restore its context structure in the learning stage and the translation stage, respectively.

If a few rules are not useful to translate, they will be eliminated by generalizing the rule dictionary optimally to keep a designed size. The ability of optimal generalization in 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.

4. GENERATION AND APPLICATION OF TRANSLATION RULE

4.1. Correction of acquired parts

The two reference speech samples are divided into a common part and one or two 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 correct incomplete common and different parts using heuristic techniques when a common part is fragmented by a small different part, or a different part is fragmented by a small common part.

4.2. Acquisition of translation rules

Common and different parts corrected in 4.1 are applied to determine the rule elements needed to generate translation rules. Figure 6 shows the results of comparing utterances. In Figure 6(A), a part containing continuous values of "0" represents a calculated common part where two utterances have much acoustic similarity. On the contrary, a part consisting of only "1" is regarded as a different part where two utterances are calculated as a long different part.

When a sentence structure includes common and different parts at the same time, we can also treat this structure as a third case in Figure 6(B). We deal with these three cases of sentence structure as the "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

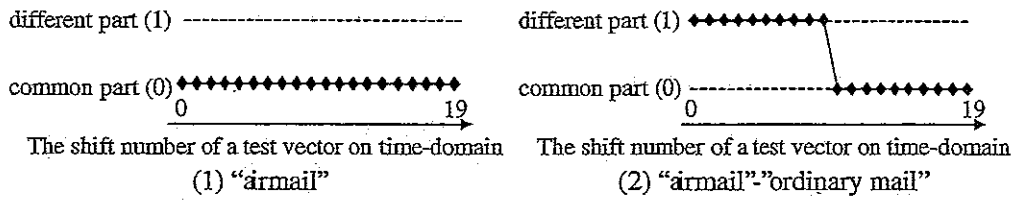

FIGURE 6. Common and different parts

elements of the translation rules for the rule generation. The set of common parts extracted from the source language, that have 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 that 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 as the following types:

Rule type 1: those with a very high acoustic sentence similarity
Rule type 2: those with sentences including common and different parts
Rule type 3: those with very low acoustic 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 an utterance in source language
- sets of start and end points of each common and different part
- index number of an utterance in the target language

4.3. Translation and speech synthesis

When an unknown speech of a source language is applied to be translated, 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 the rules.

5. EVALUATION EXPERIMENTS

5.1. Data and rule acquisition

All data in experiments are achieved through several speech processes as explained in 2.1. The conditions shown in Table 2 are also adopted in these experiments. The system needs to decide the most suitable candidates of rules from the rule dictionary, and to combine them for each translation. If the level of calculated acoustic 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.

The rule dictionary has no rule or initial information at the beginning of learning. We applied 171 nouns or noun phrases extracted from conversations to the system for translating. The average of data length was 1.65 and 1.965 seconds in source language and target language, respectively. These nouns and noun phrases had corresponded correctly between two languages in preparation of data.

5.2. Experimental results

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 nouns and noun phrases, and the number of registered translation rules between two languages.

We adopted a coverage of rules to evaluate the effectiveness of rules for translation. The coverage is defined as ratio of useful rules to all rules created by learning. This measure provides us with a performance of candidates of rules combined by a sentence form rule as mentioned in 3. If a rule combination produces a high level coverage for an unknown input, the adopted rules are able to show an easy way to translate. Figure 7 shows the results of the experiment.

TABLE 2. Conditions for experiments.

Frame length of test vector	150msec
Frame rate of both vectors	75msec
The rate of agreement for adopting rules	95%

TABLE 3. Translation rules.

Set of data	Nouns and Noun phrases	Registered rules
Conversation 1	30	771
Conversation 2	33	3,369
Conversation 3	52	11,577
Conversation 4	56	26,510

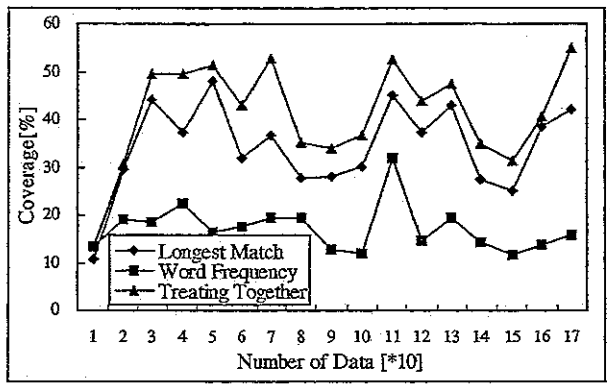

FIGURE 7. Changes in the coverage of rules

TABLE 4. Examples of high coverages

longest match	word frequency	total
0.719	0.14	0.86
0.389	0.222	0.611
0.667	0.242(50% overlap)	0.788

5.3. Discussion

Figure 7 explains that the performance of coverages indicate which type of criteria has a priority to combine rules for translating among rules with the longest match on time-domain, rules with the highest word frequency and rules combining both of these. With an increasing amount of data for learning, the rules applied with the longest match are more dependable than the rules frequently selected as candidates. However, treating with both types is the best for the translation.

Table 4 shows several examples of high coverages conducted by treating both information together. In the first two examples, there is no overlap between a rule with the longest match and another with highest word frequency. Even if the rules used to form a translated utterance overlap with other rules, they are able to give a better rate without affecting the translation.

However, many nouns and noun phrases do not have high coverage. We have to consider several reasons why suitable results could not be obtained through the experiments. A small amount of speech data is regarded as a factor because it is impossible to translate words not registered in the rule dictionary, so more translation rules should be acquired and adapted for translation.

In our method, the rules registering in the rule dictionary are acquired when very similar expressions are needed several times for comparison, so that we need to select the data utilized for translation. The system has performed the task because many suitable rules are registered in the rule dictionary. 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 utterances to obtain more translation rules, and it is also necessary to consider the contents of utterances for more effective rule acquisition and application.

6. CONCLUSION AND FUTURE WORKS

In this paper, we have described the proposed method and have evaluated the effectiveness of translation rules for nouns and noun phrases without acoustic and language models in the method. We have confirmed that appropriate acoustic information can be extracted by comparing speech, and that rules have been generated even if no target speech was obtained through the system.

We will have to implement translation experiments with a large amount of data, and confirm the synthesized speech in target language by listening.

We will consider the possibility of a direct speech translation system from the speech of a person with an esophageal speech impediment to normal speech because conventional speech recognition methods are not able to assist.

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

[Araki and Tochinai2001] K. Araki and K. Tochinai. 2001. Effectiveness of natural language processing method using inductive learning. In *Proceedings of Artificial Intelligence and Soft Computing(ASC)'01*, pages 295–300.

[Metze et al.2002] F. Metze, J. McDonough, H. Soltau, C. Langley, A. Lavie, L. Levin, T. Schultz, A. Waibel, R. Cattoni, G. Lazzari, N. Mana, F. Pianesi, and E. Pianta. 2002. The nespole! speech-to-speech translation system. In *Proceedings of HLT 2002*.

[Morimoto et al.] T. Morimoto, N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi, and Y. Yamazaki. A speech and language database for speech translation research. In *Proceedings of ICSLP'94*, pages 1791–1794.

[Müller and Stahl1999] 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.

[Murakami et al.1997] 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 *Proceedings of ITC-CSCC '97*, pages 1013–1016, Okinawa.

[Murakami et al.2002a] K. Murakami, M. Hiroshige, K. Araki, and K. Tochinai. 2002a. Evaluation of direct speech translation method using inductive learning for conversations in the travel domain. In *Proceedings of ACL-02 Workshop on Speech-to-Speech Translation.*

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

[Silverman and Morgan1990] H. F. Silverman and D. P. Morgan. 1990. The application of dynamic programming to connected speech recognition. IEEE, ASSP Magazine, pages 6–25.

[Takezawa et al.1998] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto. 1998. Japanese-to-english speech translation system:atr-matrix. In *Proceedings of ICSLP '98*, pages 2779–2782.