Effectiveness of A Direct Speech Transform Method Using Inductive Learning from Laryngectomee Speech to Normal Speech

Koji MURAKAMI[1], Kenji ARAKI[1],
Makoto HIROSHIGE[1], and Koji TOCHINAI[2]

[1] Hokkaido University, Graduate school of Engineering,
North 13, West 8, Kita-ku, Sapporo, 060-8628, Japan
{mura, araki, hiro}@media.eng.hokudai.ac.jp
http://sig.media.eng.hokudai.ac.jp/~mura
[2] Hokkai Gakuen University, Graduate school of Business Administration
Asahimachi 4-1-40, Toyohira-ku, Sapporo, 062-8605, Japan
tochinaik@econ.hokkai-s-u.ac.jp

Abstract. This paper proposes and evaluates a new direct speech transform method with waveforms from laryngectomee speech to normal speech. Almost all conventional speech recognition systems and other speech processing systems are not able to treat laryngectomee speech with satisfactory results. One of the major causes is difficulty preparing corpora. It is very hard to record a large amount of clear and intelligible utterance data because the acoustical quality depends strongly on the individual status of such people.

We focus on acoustic characteristics of speech waveform by laryngectomee people and transform them directly into normal speech. The proposed method is able to deal with esophageal and alaryngeal speech in the same algorithm. The method is realized by learning transform rules that have acoustic correspondences between laryngectomee and normal speech. Results of several fundamental experiments indicate a promising performance for real transform.

Keywords : Esophageal speech, Alaryngeal speech, Speech transform, Transform rule, Acoustic characteristics of speech

1 Introduction

Speech is a perfect medium and most common for human-to-human information exchange because it is able to be used without hands or other tools, being a fundamental contribution for ergonomic multi-modality[1]. Much research has also been developed to realize such advantages for human-machine interaction. Many applications have been produced and they are contributing to human life.

On the other hand, many people who are unable to use their larynxes are not able to benefit from such advances in technology although such assistance is hoped for. Both esophageal and alaryngeal speech, which laryngectomee people

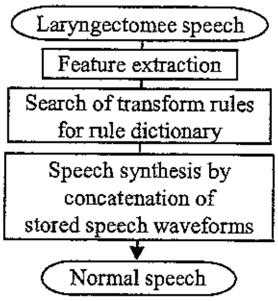

Laryngectomee speech
Feature extraction
Search of transform rules for rule dictionary
Speech synthesis by concatenation of stored speech waveforms
Normal speech

Fig. 1. Processing of the proposed method.

practice to enable conversation, are understandable and enables adequate communication. However, conventional speech processing systems are not able to accept them as inputs because almost all acoustic models in the current systems are trained by intelligible utterances spoken by normal people. It is easy to find a lot of corpora high in both quality and quantity in many languages. However, there are not many resources of laryngectomee or other disordered speech because it is very difficult to sample a number of intelligible and clear utterances. One of the major causes is dependence on individual status of speech.

We only focus on laryngectomee speech waveforms themselves to transform them into normal speech. Many studies have attempted to transform laryngectomee speech to normal speech, for example: re-synthesizing fundamental frequency or formant of normal speech[2], or by utilizing a codebook[3]. We propose a radically different speech transform approach which handles only acoustic characteristics. The concepts of the method have been applied to realize a speech translation method and provided promising effectiveness[4, 5]. Fig.1 shows the processing stages of our method. The proposed method is realized by dealing with only correspondences of acoustic characteristics between both speech waveforms. Our basic conception is based on belief that even laryngectomee utterances have certain contents although these are inarticulate and quite different from normal speech waveforms. At first, acoustic common and different parts are extracted by comparing two utterances within the same speech side. These parts should have correspondences of meaning between two different types of speech. Then we generate transform rules and register them in a transform dictionary. The rules also have the location information of acquired parts for speech synthesis on time-domain. The transform rules are obtained by comparing not only speech samples but also acquired transform rules themselves using Inductive Learning Method[6], still keeping acoustic information within the rules. Deciding the correspondence of meaning between two speech sides is the unique condition necessary to realize our method.

In a transform phase, when an unknown utterance of laryngectomee speech is applied to be transformed, the system compares this sentence with the acoustic information of all rules within the speech side. Then several matched rules are utilized and referred to their corresponding parts of the normal speech side.

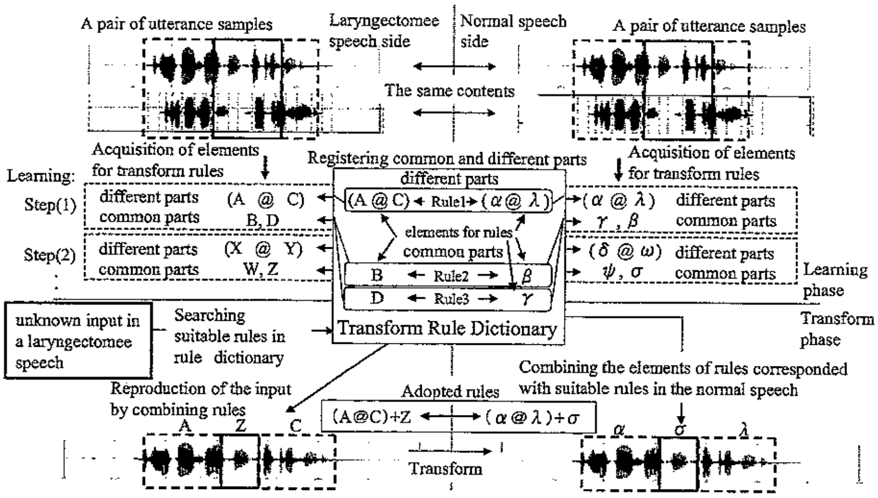

A pair of utterance samples
Laryngectomee speech side
Normal speech side
A pair of utterance samples
The same contents
Acquisition of elements for transform rules
Registering common and different parts
Acquisition of elements for transform rules
Learning:
Step(1)
different parts (A @ C)
common parts B, D
different parts
(A @ C) ← Rule1 → (α @ λ)
(α @ λ) different parts
γ , β common parts
elements for rules
common parts
Step(2)
different parts (X @ Y)
common parts W, Z
B ← Rule2 → β
D ← Rule3 → γ
(δ @ ω) different parts
ψ, σ common parts
Learning phase
Transform phase
unknown input in a laryngectomee speech
Searching suitable rules in rule dictionary
Transform Rule Dictionary
Combining the elements of rules corresponded with suitable rules in the normal speech
Reproduction of the input by combining rules
Adopted rules
(A@C)+Z ⟷ (α @ λ)+σ
A Z C
α σ λ
Transform

Fig. 2. Processing structure.

Finally, we obtain roughly synthesized normal speech utterance by simply concatenating several suitable parts of rules in the normal speech side according to the information of location. Fig.2 explains an overview of the processing structure of our method.

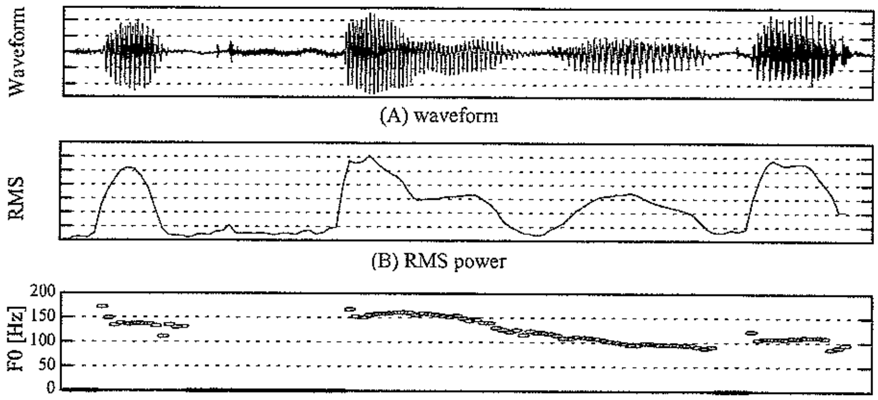
Although the method deals with speech waveforms, the boundaries of word, syllable, or phoneme are not important for our method because the proposed method does not prepare any acoustically correct interpretation, and transform rules are designed by acoustic characteristics of speech from utterances. Therefore, these rules will be able to adapt the speaker's characteristics and habits by recursive learning. We evaluate effectiveness of the transform rules through fundamental experiments and offer discussion on behaviors of the system.

2 Laryngectomee speech

Laryngectomee people try to acquire esophageal or alaryngeal speech as second speech to enable them to once again communicate effectively in society. The characteristics of these types of speech are explained in this section. Fig. 3, 4, and 5 show normal, alaryngeal and esophageal speech, respectively. Each Figure contains (A)waveform itself, (B)RMS power and (C)fundamental frequency.

2.1 Alaryngeal speech

Alaryngeal speech has an unnatural quality and is significantly less intelligible than normal speech. The utterances spoken using an artificial larynx, are not able to contain any intonation, accent and emotion despite the speakers intention. The cause is that this device is only able to vibrate fixed impulse source. Some

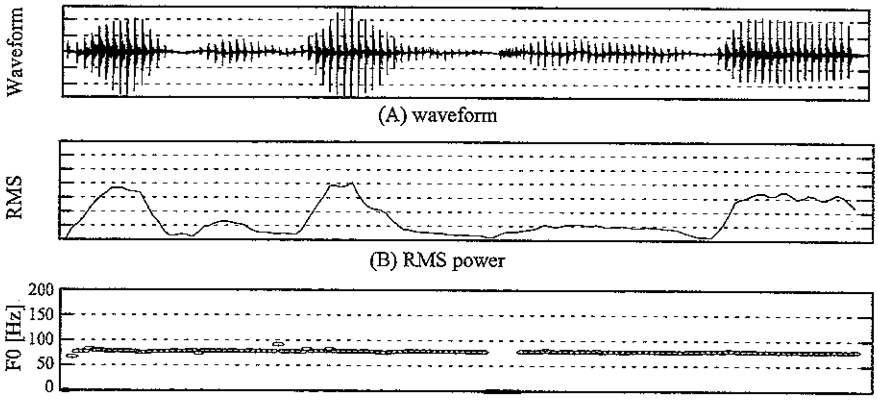


(A) waveform

(B) RMS power

(C) Fundamental frequency

Fig. 3. Normal speech

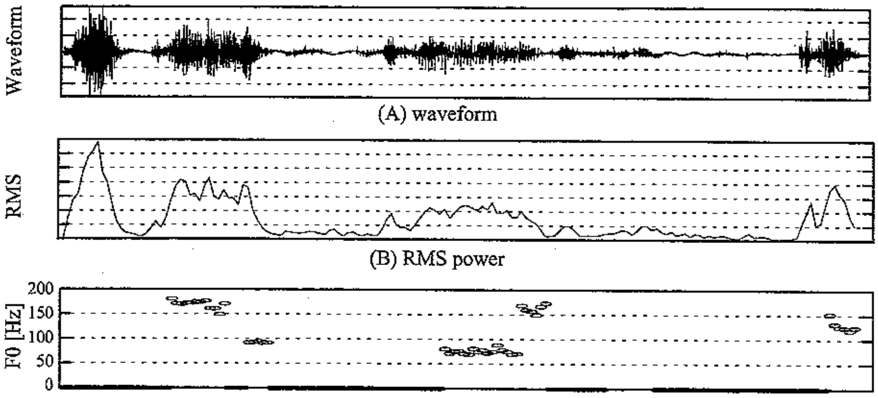


(A) waveform

(B) RMS power

(C) Fundamental frequency

Fig. 4. Alaryngeal speech



(A) waveform

(B) RMS power

(C) Fundamental frequency

Fig. 5. Esophageal speech

Table 1. Result of speech recognition.

Type of speech	Number of utterance	Accuracy of correct words
Normal Speech	80	65.82%
Alaryngeal Speech	119	29.61%
Esophageal Speech	107	24.32%

research has improved the performance and quality of speech[9]. Fig. 4 shows a sample of alaryngeal speech.

2.2 Esophageal speech

Characteristics of esophageal speech mainly depend on difference of sound source mechanism as shown Fig. 5. Several remarkable features are as follows: lower fundamental frequency than normal speech, including a lot of noise and lower volume[7]. Moreover, differences on prosody and spectral characteristics of speech are also reported[8].

2.3 Speech recognition for laryngectomee speech

We need to reveal the actual performance of conventional speech recognition for laryngectomee speech. We utilized Julius[10] as a speech recognition tool. The acoustic and language models in the system were constructed by the learning of normal speech utterances. Table 1 explains the result of recognition performance. It is obvious that the system is not able to treat laryngectomee speech without rebuilding the acoustic model for esophageal and alaryngeal speech utterances.

3 Speech processing

3.1 Speech data and Spectral characteristics

Various acoustic parameters specific to disordered speech have been developed and applied to many studies[11]. Our study has succeeded to show acoustic

Table 2. Experimental conditions of speech processing.

Size of frame	30msec
Frame cycle	15msec
Speech window	Hamming Window
AR Order	14
Cepstrum Order	20

Table 3. Information of speakers.

Type of speech	Age/Gender	Speaker's feature
Normal Speech	24/Male	student
Alaryngeal Speech	70/Male	operation in 1990
Esophageal Speech	65/Male	operation in 1994

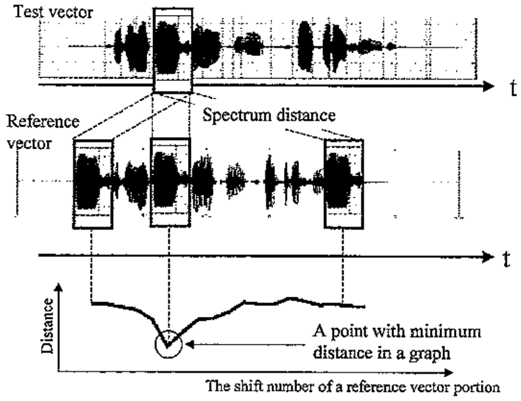

Test vector
t
Reference vector
Spectrum distance
t
Distance
A point with minimum distance in a graph
The shift number of a reference vector portion

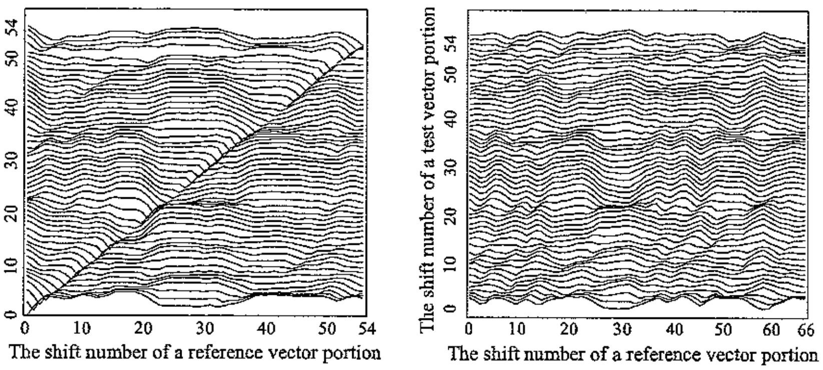
Fig. 6. Comparison of vector sequences.

differences by a clustering method using these values between normal and disordered female voices[12]. However, we have focused on results of comparison experiments using only spectral analysis[8].

We recorded utterance data with 16bit and 48kHz sampling rate, and down-sampled to 16kHz. These data were spoken by three people whose usual speech is normal, esophageal and alaryngeal, respectively. Table 2 shows parameters adopted for speech processing, and Table 3 shows these speaker's characteristics. In this report, LPC Cepstrum coefficients were chosen as spectral parameter, because we could obtain better results than the other representations of speech characteristics[4].

3.2 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[13] 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. We also adopted the Least-Squares Distance Method 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 "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 Fig. 6. Next, the new "test vector" is picked out by the



(A) "*KOKODE HITOWO MIRU* (I watch a man.)."

(B) "*KOKODE HITOWO MIRU* (I watch a man.)."-"*WATASHIWA KARATEWO SURU* (I play Karate.)."

Fig. 7. Difference between utterances.

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.

Fig. 7 shows two examples of the difference between two utterances. Italic characteristics express Japanese. These applied speech samples are spoken by the same esophageal speaker. The contents of the compared utterances are the same in Fig. 7(A), and are quite different in Fig. 7(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 Fig. 7(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 Fig. 7(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" by evaluating the point by a decision method in our previous research[4]. 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.

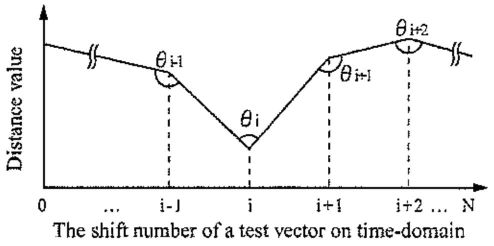

θ i-1
θ i
θ i+1
θ i+2
Distance value
0 ... i-1 i i+1 i+2 ... N
The shift number of a test vector on time-domain

Fig. 8. Angle calculation.

3.3 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 chose the angles formed where two lines cross each other. These lines are linked by two distance points. Figure 8 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. 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 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.

4 Inductive Learning Method

Inductive Learning Method[6] 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 transform stage, respectively. This approach has been applied to many areas of natural language processing. The effectiveness of the method was also confirmed in machine translation. Therefore, we applied this method to acoustic characteristics of speech instead of character strings for realizing a direct speech transform approach.

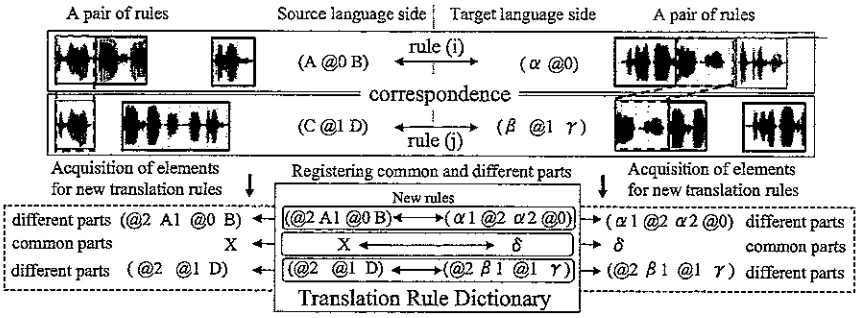

A pair of rules
Source language side
Target language side
A pair of rules
(A @0 B)
rule (i)
(α @0)
correspondence
(C @1 D)
rule (j)
(β @1 γ)
Acquisition of elements for new translation rules
Registering common and different parts
Acquisition of elements for new translation rules
New rules
different parts (@2 A1 @0 B)
(@2 A1 @0 B)
(α1 @2 α2 @0)
(α1 @2 α2 @0) different parts
common parts X
X
δ
δ common parts
different parts (@2 @1 D)
(@2 @1 D)
(@2 β1 @1 γ)
(@2 β1 @1 γ) different parts
Translation Rule Dictionary

Fig. 9. Rule acquisition using Inductive Learning Method.

Figure 9 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 similar to comparisons between speech utterances. 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 transform.

If several rules are not useful for transform, they will be eliminated by generalizing the rule dictionary optimally to keep a designed size of memory. 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.

5 Generation and application of transform rule

5.1 Acquisition of transform rules

Acquired common and different parts are applied to determine the rule elements needed to generate transform rules. At the time, there are three cases of sentence structure as the "rule types". If two compared utterances were almost matching or did not match at all, several common or different parts are acquired, respectively. And the other case is that these utterances have both parts at the time. Combining sets of common parts of both normal and laryngectomee speech become elements of the transform rules for rule generation. The set of common parts extracted from the laryngectomee speech, which have a correspondence of meaning with a set of common parts in normal speech, are kept. The sets of different parts become elements of the transform rules as well. Finally, these transform 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 speech sides are the same. When the types are different, it is impossible to obtain the transform rules and register them in the rule dictionary because we are not able to decide the correspondence between two speech sides uniquely. Information that a transform rule has are as follows:

Table 4. Conditions for experiments.

Frame length of test vector	120msec
Frame rate of both vectors	60msec
Margin of time delay	±180ms(±120ms)
The rate of agreement for adopting rules	95%

- rule types as mentioned above
- index number of an utterance in both speech sides
- sets of start and end points of each common and different part

5.2 Transform and speech synthesis

When an unknown laryngectomee speech is applied to be transformed, acoustic information of acquired parts in the transform rules are compared in turn with the unknown speech, and several matched rules become the candidates to transform. The inputted utterance should be reproduced by a combination of several candidates of rules. Then, the corresponding parts of the normal speech in candidate rules are referred to obtain transformed speech. Although the final synthesized normal speech may be produced roughly, speech can directly be concatenated by several suitable parts of rules in the normal speech side using the location information on time domain in the rules.

6 Evaluation Experiments

All data in experiments were achieved through several speech processes as explained in 3.1. We applied 80 utterances of each speaker. The contents of input speech were 54 three-digit numbers and 26 simple sentence included 8 of *"WATASHIWA * WO SURU."*(I play * .) and 5 of *"SOKODE * WO MIRU."*(I watch a *.). Italic fonts express Japanese. The system was prepared with the same parameters throughout the experiments for both between esophageal or alaryngeal and normal speech to evaluate the generality of the system. The conditions shown in Table 4 were also adopted in these experiments. The rule dictionary had no rule or initial information at the beginning of learning.

We evaluated that the system could obtain a number of useful transform rules created by only the calculation of acoustic similarity between both esophageal and normal speech, and alaryngeal and normal speech. Any other criterion was adopted to limit to acquire transform rules throughout the experiments. When several common parts were found in the calculation result in comparing utterances, the one with the longest match was acquired as the transform rule. Moreover, location of parts on time-domain was also evaluated because this characteristic expressed the accuracy of correspondence of parts to those in another speech side. We allowed a margin for parts appearing in time domain for difference in elasticity of individual utterances. Two margins, ±180ms and ±120ms were applied because they corresponded with 1 and 1.5 mora in the Japanese

Table 5. Comparison of correspondences of acquired rules.

Speech data	Number of Data	Number of acquired rules	±180ms	±120ms
Alaryngeal-Normal	80	2,284	1,665(73.9%)	1,315(57,6%)
Esophageal-Normal	80	1,378	1,055(76.6%)	646(61.4%)

speech rate, respectively. When corresponding parts between two speech sides in a rule appeared in appropriate location on time-domain with suitable length, the rule included these parts was regarded as a correct rule because the correspondences were able to be decided uniquely.

Table 5 shows a number of transform rules acquired by only acoustic similarity. The system could also obtain many rules that have appropriate correspondences without any limitation. Percentages in parentheses show the ratio of total acquired rules to appropriate rules. These rules imply that it is possible to acquire correspondences between both speech sides by only calculating of acoustic similarity.

7 Discussion

Many appropriate rules are obtained in both experiments through the same parameters. The results shows common and different parts appear approximately close location on time-domain independent of speech type. They also indicate that calculation of acoustic similarity is able to be a criterion to partition laryngectomee utterances although these are not clear and intelligible and are not able to be deal with in conventional speech recognition. So, these rules show promising possibilities for transform experiments. The number of appropriate rules from esophageal speech is lower than from alaryngeal speech. Noises accrued from injecting volumes of air into the esophagus are one of the major causes. The table also show corresponding parts were adequately acquired in close location on time-domain between normal and laryngectomee speech. However, the injecting volumes causes the other problem that esophageal utterances have a tendency to be longer than other types of speech. Therefore, longer margin should be needed to deal with esophageal speech.

In both experiments, the system is able to obtain more than 50% of suitable rules with our unique criterion, location of time-domain. This limitation is indispensable to manage the number of rules. We need other criteria to keep a small number of rules in the dictionary, for example, stricter limitation on time-domain, or checking length of.extracted common or different parts, and so on.

We need to increase the number of speech utterances to obtain more suitable transform rules, and it is also necessary to consider the contents of utterances for more effective rule acquisition and application.

8 Conclusion and future works

In this paper, we have described the proposed method and have evaluated rule acquisition without being parameter tuning specific for esophageal and alaryngeal speech. We have confirmed that appropriate acoustic information is able to be extracted by calculation of acoustic similarity and that rules have been generated.

We will consider adopting DP matching method to decrease calculation cost because the method described in 3.2 needs a large amount of calculation.

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

References

1. J. Müller and H. Stahl. 1999. *Speech understanding and speech transform by maximum a-posteriori semantic decoding. Proceedings of Artificial Intelligence in Engineering*, pages 373–384.
2. Wen Ding and N. Higuchi. *A voice conversion method based on complex RBF network. Proceedings of the 1997 autumn meeting of ASJ(Japanese)*, pp.335–336. 1997.
3. Oyton Turk and Levent M.Arslan. *Subband based Voice Conversation. Proceedings of ICSLP2002, pp.289–292.* 2002.
4. K. Murakami, M. Hiroshige, K. Araki and K. Tochinai. 2002. *Evaluation of direct speech transform method using Inductive Learning for conversations in the travel domain. Proceedings of ACL-02 Workshop on Speech-to-Speech Translation*, pages .45–52.
5. K. Murakami, K. Araki, M. Hiroshige, and K. Tochinai. 2003. *Evaluation of the rule acquisition on a direct speech translation method with waveforms using Inductive Learning for nouns and noun phrases. Proceedings of Pacific Association for Computational Linguistics(PACLING)'03*, pp.121–130.
6. K. Araki and K. Tochinai. 2001. *Effectiveness of natural language processing method using inductive learning. Proceedings of Artificial Intelligence and Soft Computing(ASC)'01*, pages 295–300.
7. K. Matsui and E. Noguchi. *Enhancement of esophageal speech. proceedings the 1996 autumn meeting of the ASJ(Japanese)*, pp.423–424. 1996.
8. J. Lu, Y. Doi, S. Nakamura and K. Shikano. *Acoustical Characteristics of Vowels of Esophageal Speech. Technical report of IEICE, SP96-126*, pp.233–240. 1997.
9. Carol Y. Espy-Wilson, Venkatesh R. Chari and Caroline B. Huang. *Enhancement of Alaryngeal Speech by Adaptive Filter. Proceedings of ICSLP '96*, pp764–767. 1996.
10. A. Lee, T. Kawahara and K. Shikano. *Julius – a Open Source Real-Time Large Vocabulary Recognition Engine. Proceedings of EUROSPEECH '01*, pp.1691–1693. 2001.
11. Y. Katoh. *Acoustic characteristics of speech in voice disorders. The 2000 spring meeting of the ASJ(Japanese)*, pp.309–310. 2002.
12. Daniel Callan, Ray D. Kent, Nelson Roy and Stephen M. Tasko. *Self-organizing Map for the Classification of Normal and Disordered Female Voices. Journal of Speech, Language, and Hearing Research*, Vol.43, pp.355–366. 1999.
13. H. F. Silverman and D. P. Morgan. 1990. *The application of dynamic programming to connected speech recognition. IEEE, ASSP Magazine*, pp.6–25.