A STUDY OF A DIRECT SPEECH TRANSFORM METHOD ON LARYNGECTOMEE SPEECH

Koji Murakami, Kenji Araki, Makoto Hiroshige¹, and Koji Tochinai²
¹Graduate School of engineering, Hokkaido University, Japan
²Graduate School of business administration, Hokkai-gakuen University, Japan

Abstract: This paper proposes and evaluates a new direct speech transform method with waveforms from laryngectomee speech to normal speech. Most conventional speech recognition systems and 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. Our proposed method focuses on the acoustic characteristics of speech waveform of laryngectomee people and transforms such characteristics 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

I INTRODUCTION

Speech is a perfect medium and the most common for human-to-human information exchange because it is able to be used without hands or other tools, being a fundamental contributor to ergonomic multi-modality. Much research have been developed to realize such advantages for human-machine interaction. Many applications are produced and they are recently 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 expected. Both esophageal and alaryngeal speech, which laryngectomee people practice to enable conversation, are understandable and enable adequate communication. However, conventional speech processing systems are not able to accept them as inputs because almost all current systems deal with only normal speech. Many intelligible utterances spoken by normal people have to be prepared as learning data to construct useful acoustic models for the systems. It is easy to find a lot of corpora valuable 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. Thus it is not easy to obtain a high acoustic quality of corpora.

We 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 the fundamental frequency or formant of normal speech[1], or by utilizing a codebook[2]. We propose a radically different speech transform approach which handles only acoustic characteristics. Fig.1 shows the processing stages of our method. The proposed method is realized by dealing with only the correspondence in acoustic characteristics of speech waveforms. Our basic conception is based on our belief that laryngectomee utterances contain acoustic characteristics although these are inarticulate and quite different from normal speech waveforms. Thus acoustic common and different parts extracted by comparing with two utterances within the same speech side have correspondences of meaning between two different types of speech. We generate transform rules and register them in a translation dictionary. The rules also have the location information of acquired parts for speech synthesis on time-domain. 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. 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.

The boundaries of word, syllable, or phoneme are not important for our method because we acquire only acoustic common and different parts as transform knowledge by comparing speech utterances. We evaluate effectiveness of the transform rules through fundamental experiments and offer discussion on behaviors of the system.

II. LARYNGECTOMEE SPEECH

Laryngectomy 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.

2.1 Esophageal speech

Characteristics of esophageal speech mainly depend on difference of sound source mechanism. Several remarkable features are as follows: lower fundamental frequency than normal speech, including a lot of noise and lower volume[3]. Moreover, differences on prosody and spectral characteristics of speech are also reported[4].

2.2 Aralyngeal speech

Aralyngeal speech has an unnatural quality and is significantly less intelligible than normal speech. The utterances spoken using artificial larynx, are not able to contain any accent and intonation despite the speaker’s intention. The cause is that this device is only able to vibrate fixed impulse source. Therefore, it is impossible to express their emotion or intention with speech.

2.3 Speech recognition for laryngectomee speech

We need to reveal the actual performance of usual speech recognition for laryngectomee speech. We utilized Julius[5] 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 very clear that the system is not able to treat laryngectomee speech without rebuilding the acoustic model of many esophageal or alaryngeal speech utterances.

III. SPEECH PROCESSING

3.1 Speech data and spectral characteristics

Various acoustic parameters specific to disordered speech have been developed and applied to many studies[6]. One such study has succeeded to show acoustic differences by a clustering method using these values between normal and disordered female voices[7]. However, we have focused on results of comparison experiments using only spectral analysis [4].

We recorded utterance data with 16bit and 48kHz sampling rate, and downsampled to 16kHz. These data were spoken by three people whose 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 focused on frequency characteristics of speech and could obtain better results than other representations of speech characteristics[8].

<table>
<thead>
<tr>
<th>Size of analysis frame</th>
<th>30ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame cycle</td>
<td>15ms</td>
</tr>
<tr>
<td>Speech window</td>
<td>Hamming Window</td>
</tr>
<tr>
<td>AR Order</td>
<td>14</td>
</tr>
<tr>
<td>Cepstrum Order</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 3 Information of speakers.

<table>
<thead>
<tr>
<th>Type of Speech</th>
<th>Age/Gender</th>
<th>Speaker’s feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal Speech</td>
<td>24/male</td>
<td>Student</td>
</tr>
<tr>
<td>Alaryngeal</td>
<td>70/male</td>
<td>Operation in 1990</td>
</tr>
<tr>
<td>Esophageal</td>
<td>65/male</td>
<td>Operation in 1994</td>
</tr>
</tbody>
</table>

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[9] 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.2. 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 can get several distance curves results between two speech samples.

Fig.2 Comparison of vector sequences.

Fig. 3 shows an example of the difference between two utterances. This applied speech sample is spoken by the same normal speaker and the contents of the utterances are the same. 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.

Fig.3 Difference of utterances: “airmail.”

As shown in Fig.3, 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[8]. 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.

IV. GENERATION AND APPLICATION OF TRANSFORM RULES

4.1 Acquisition of transform rules

Acquired common and different parts are applied to determine the rule elements needed to generate translation 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 same 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:

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

4.2 Transform and speech synthesis

When an unknown utterance of a laryngectomee person 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.

Table 4 Condition for experiments.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame length of test vector</td>
<td>120msec</td>
</tr>
<tr>
<td>Frame rate of both vectors</td>
<td>60msec</td>
</tr>
<tr>
<td>Margin of time delay</td>
<td>±180ms, ±120ms</td>
</tr>
</tbody>
</table>

V. RULE ACQUISITION EXPERIMENTS

All data in experiments are achieved through several speech processes as explained in 3.1. We applied 80 utterances of each speaker. The system is prepared with the same parameters throughout the experiments between both esophageal or alaryngeal and normal speech to evaluate the generality of the system. The conditions shown in Table 4 are also adopted in these experiments. The rule dictionary has no rule or initial information at the beginning of learning.

We evaluate that the system could obtain a number of useful transform rules created by only the calculation of acoustic similarity. Moreover, location of parts on time-domain is also evaluated because this characteristic expresses the accuracy of correspondence of parts to those in another speech side. We allow a margin for parts appearing in time domain, ±180ms and ±120ms to consider for individual uttering differences. When corresponding parts between two speech sides in a rule appear in appropriate location on time-domain with corresponding parts between two speech sides in a rule, the rule included these parts is regarded as a correct rule because the correspondences are able to be decided uniquely. Table 5 shows a number of acquired rules and those that have appropriate correspondence.

Table 5 Comparison of correspondences of acquired rules.

<table>
<thead>
<tr>
<th>Speech Data</th>
<th>Num. of Data</th>
<th>Num. of acquired rules</th>
<th>±180ms</th>
<th>±120ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaryngeal &lt;&gt; normal</td>
<td>80</td>
<td>2,284</td>
<td>1,665  [73.9%]</td>
<td>1,315  [57.6%]</td>
</tr>
<tr>
<td>Esophageal &lt;&gt; normal</td>
<td>80</td>
<td>1,378</td>
<td>1,055  [76.6%]</td>
<td>846    [61.4%]</td>
</tr>
</tbody>
</table>

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

REFERENCES