SPECTRAL ANALYSIS OF ESOPHAGEAL SPEECH

Jinlin LU, Satoshi NAKAMURA and Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-chou, Ikoma, Nara, 630-0101, Japan

Summary: Esophageal speech is widely used by laryngectomized people. The purpose of this paper is to show the spectral characteristics of esophageal speech. The analysis results from 5 sustained Japanese vowels, which produces by 7 esophageal and 7 normal speakers, indicate that the energy in frequency of 4000 - 4500Hz is systematically larger than normal one, and the spectral tilt is less than normal one 0.98dB/Oct. From a preliminary perceptual evaluation, we find that attenuation of the energy in higher frequency range make the reconstructed speech more preferable to listeners.

INTRODUCTION

There are approximately 20,000 patients with laryngeal cancer in Japan and the number has been increasing in recent years. Treatment of larynx cancer may necessitate total removal of the larynx. Total laryngectomy permanently alters the speech production mechanism. Although laryngectomees have lost their vocal cords, the source of the speech, they can communicate with others after mastering a special speech method called esophageal speech. Whereas normal speakers utter speech by vibrating the vocal cords with air stored in the lungs, laryngectomees produce speech sounds by vibrating the upper part of esophagus with air swallowed into esophagus. Since esophageal speakers are able to improve articulation through training, they can communicate to some extent, even though their speech sounds different from normal speech.

However, esophageal speech can not be loud and lasts only for a short period of time because the amount of air available for vocalization is small. In addition, mainly because the vibration of upper part of esophagus is irregular than the vibration of the vocal cords for normal speakers, esophageal speech is more noisy, and strongly perturbated. Therefore, their uttered speech is not so clear, not easy to understand, then their communication is limited. Some studies have been made to enhance the esophageal speech\(^\text{[1-3]}\). However, they are far from complete. In this paper, in order to improve esophageal speech, we investigate the spectral characteristics of esophageal speech in comparison with one of normal speech, and then to try to improve the quality of reconstructed esophageal speech.

SPECTRAL ANALYSIS

Esophageal speech is generated by instead source, the difference of source between esophageal speech and normal one is very large\(^\text{[4]}\). Because the speech generation method of esophageal speech is differ from normal one, the effective to spectra will be existed. To improve the quality of esophageal speech, some spectral characteristics of it must be investigated..
There are many methods of spectral analysis. It is a good method, that is, first extract formant frequency and bandwidth, then compare the differences between these two types of speech. However, extract formant accurately from esophageal speech is very difficult.

Considering the perturbation of period and amplitude, LPC (autocorrelation method) spectra is used to obtain smooth spectra. The average operation is performed for frames and speakers. For this averaged spectra, next two points were discussed.

1. Spectral tilt
   check on the spectral tilt of esophageal and normal speech. Assume the spectra of esophageal speech is \( S_e(k) \), while \( S_n(k) \) for normal speech, the difference of spectral tilt between them can be presented by next equation.
   \[
   20 \cdot \log(S_e(k)) = 20 \cdot \log(S_n(k)) + \alpha \cdot 20 \cdot \log(k)
   \] (1)
   Here, \( k \) is the magnification of frequency resolution. \( k \)’s range \( 0 - 512 \) is corresponding to frequency \( 0 - 5000Hz \) for analysis condition (Table 1). \( \alpha \) can be determined by minimizing the least square error of difference of two sides in equation (1).
   \[
   \alpha = \frac{\sum_{k=1}^{N} \log(S_e(k)/S_n(k)) \cdot \log(k)}{\sum_{k=1}^{N} (\log(k))^2}
   \] (2)
   Because \( 20 \cdot \log(k) \) means \( 6dB/oct \) of spectra tilt, the difference of spectral tilt will be \( \alpha \cdot 6dB/oct \).

2. Energy in higher frequency range
   The spectra of vowel is affected greatly by formant. In lower frequency range (lower than first formant frequency), for example, 0-300 Hz, the spectra will be affected by fundamental frequency. Considering the frequency range of over 4000 Hz which influence from formant became fainter, but low-pass cut-off frequency is 4700 Hz, we select finally frequency range of 4000 - 4500 Hz as our analysis target range to measure the average energy.

**SPEECH MATERIALS**

Fourteen adult men who include 7 esophageal speakers and 7 normal speakers, produced the speech samples for this study. Each subject produced the 5 sustained Japanese vowels ( /a/, /i/, /u/, /e/, /o/). The recording was made in a quiet room. The recorded vowels were passed through a low-pass filter with a cutoff frequency of 4.7 kHz, and digitized into a computer at a sampling frequency of 10 kHz. The relative steady intervals were selected using speech waveform around 0.5 second to 1.5 second for analysis.

Table 1 shows the analysis condition.

**RESULTS**
TABLE 1: Analysis conditions

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame length</td>
<td>35ms (350 points)</td>
</tr>
<tr>
<td>Frame shift</td>
<td>5ms (50 points)</td>
</tr>
<tr>
<td>LPC ord</td>
<td>14</td>
</tr>
<tr>
<td>FFT points</td>
<td>1024</td>
</tr>
</tbody>
</table>

TABLE 2: The difference of spectral tilt

<table>
<thead>
<tr>
<th>Vowels</th>
<th>/a/</th>
<th>/i/</th>
<th>/u/</th>
<th>/e/</th>
<th>/o/</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>dB/oct</td>
<td>1.08</td>
<td>0.99</td>
<td>0.86</td>
<td>0.95</td>
<td>1.04</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Average amplitude spectra of 5 sustained Japanese vowels

Figure 1 shows the average spectra of 5 Japanese vowels. Solid lines indicate the average spectra of 7 esophageal speakers, and dashed lines indicate ones of the normal speakers. There are clearly difference in higher frequency range. In lower frequency range, less difference was existed except the formant frequency shifted to higher a little.

Spectral tilt

Table 2 shows the difference of spectral tilt between esophageal speech and normal one for each vowel. Normal speech have a slope about 0.98dB/oct than esophageal speech. But the difference is little during each vowel.

FIGURE 1 Average spectra of 5 sustained Japanese vowels
Average energy in frequency range of 4000 - 4500Hz

The difference of average spectral energy in frequency range of 4000 - 4500Hz for esophageal speech and normal speech is shown in figure 2. Horizontal axis is indicated to 5 Japanese vowels, vertical axis is to averaged energy in frequency range of 4000 - 4500Hz. "m" is to esophageal speech, and "o" is to normal speech. In this frequency range, the energy of esophageal speech is systematically and significantly large than that of normal speech.

A preliminary perceptual evaluation was performed, from these examination, if the energy in high frequency range(3000 - 4700Hz) are attenuated, then the reconstructed speech is more preferable to listeners.

CONCLUSION

In this paper, we described the spectral characteristics of esophageal speech. From the study of sustained Japanese vowel samples of several speakers, we conclude that (1) the energy of frequency range of 4000 - 4500Hz of esophageal speech is larger than normal speech's, (2) the differences between them are systematical, and (3) from a preliminary perceptual evaluation, if the energy in high frequency range are attenuated, then the reconstructed speech is more preferable to listeners.

References