Motivation: To convert Tracheoesophageal speech into a more natural sounding neutral speech.

Tracheoesophageal speech (TE):
- Laryngectomy: A surgery (involving Tracheoesophageal puncture (TEP)) to remove partially or totally the voice-box of a cancer patient. 
- TE speech is produced by laryngectomies. 
- TEP holds in place a voice prosthesis that functions as a one-way valve to allow the air from the lungs to enter the esophagus.

Voice in TE speech: slow vibrations of cervical esophagus (~60 Hz), which manifests as impulses-like structure in TE speech.

Proposed TE2N scheme

2 major steps:
1) Whisperization: Removal of the impulse-like noise to whisperize TE speech.
2) Neutral speech synthesis: Whispered speech to neutral speech conversion.

Model of the impulses in the residual signal: \[ y[k] = a_i \cdot h_k \]
- where \( h_k \) is a binary random variable
- \( a_i \) is a Gaussian random variable

\[ p(\Theta) \propto p(\theta) p(\Theta|\theta) = p(r|\theta) \prod_{n=1}^{N-1} p(a_n|h_n)p(h_n)p(c_r^2). \quad (2) \]

- \( \theta = [a, b, c_r^2] \) are the parameters to be estimated. \( N_i \) is a frame length.

Estimation of parameters:
- Using Gibbs sampling [2], iteratively estimated the each parameter given the other parameters.

Whisperized speech (\( \hat{s} \)):
\[ s[k] = \hat{r}[k] + \sum_{j=1}^{K} \hat{r}[k-j] e_n[k] \quad (4) \]

where \( \hat{r}[k] = r[k] - y[k] \) and \( \hat{y}[k] \) is the estimate of impulses.

Introduction

STEP 1: Whisperization

Pitch prediction:
\[ f_0[n] = \begin{cases} c_0[n] + f, \quad &\text{if } \hat{r}[n] > \epsilon \\ 0, \quad &\text{otherwise} \end{cases} \quad (5) \]

The final pitch contour \( f_0[n] \):
\[ f_0[n] = \begin{cases} c_0[n] + f, \quad &\text{if } \hat{r}[n] > \epsilon \\ 0, \quad &\text{otherwise} \end{cases} \quad (5) \]

Synthesis: Given the \( f_0[n] \), we generate the excitation signal (\( c \)) similar to [3].

The neutral speech (\( x \)) in each frame given by:
\[ x[k] = c[k] + \sum_{j=1}^{K} x[k-j] e_n[k] \quad (6) \]

The final neutral speech is generated using overlap add method.

References


Experiments & Results

Data set:
- Subject: 70 year old female laryngectomee.
- Stimuli: 21 phonetically balanced questions and 29 all-voiced sentences.

Baseline schemes:
- B1: Reconstructed neutral speech from TE without whisperization
- B2: Same as the B1 with residual replaced by white Gaussian noise
- B3: Reconstructed neutral speech using method proposed by [4]
- B4: Considers whispered speech as input to method [4]

Subjective evaluation: Two listening tests (LT) with 10 listeners.
- LT-1: To understand the significance of whisperization using 2 baseline schemes B1 and B2.
- LT-2: To evaluate the performance of the proposed TE2N method using 2 baseline schemes B3 and B4.
- Listener’s consistency: LT-1 - 70%, LT-2 - 80%.
- Average time taken: LT-1 - 19.05(±3.58), LT-2 - 9.57(±3.08) minutes.

Subjective evaluation results:

<table>
<thead>
<tr>
<th>Methods</th>
<th>Te2N</th>
<th>B1</th>
<th>B2</th>
<th>B1,B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preference (in %)</td>
<td>62.4</td>
<td>16.4</td>
<td>15.6</td>
<td>3.6</td>
</tr>
</tbody>
</table>

LT-1 shows whisperization of TE speech improve naturalness.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Te2N</th>
<th>B3</th>
<th>B4</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preference (in %)</td>
<td>94.40</td>
<td>0.8</td>
<td>0</td>
<td>4.80</td>
</tr>
</tbody>
</table>

LT-2 shows the proposed method outperforms baseline schemes.

Conclusion

Using listening tests, we find that the ‘whisperization’ step is vital to improve the naturalness of the TE speech.

Future work: Extending the proposed framework to data collected from several laryngectomies.

Acknowledgement: Authors thank the subject, Dr. Manjula B.V and Sahana Monhandoss for the data collection and Pratiksha Trust for their support.