Development of Hand Gesture Controlled Electrolarynx

KENJI MATSUI† Member, YOSHIHISA NAKATO† Non-member
YUMIKO O. KATO‡ Non-member

(Received June 30, 2013, revised September 19, 2013)

Abstract: The feasibility of using a motion sensor to replace a conventional electrolarynx (EL) user interface was explored. Forearm motion signals from MEMS accelerometer was used to provide on/off and prosody control. The small transducer was placed against the throat using support bandage. A battery operated ARM-based small control unit was developed and placed on the wrist. The control unit can convert the tilt angle into the pitch frequency. Speech generation was tested with various forearm movements, and then a simple and small action was chosen to control the device. A simple comparison study has been made with well-trained normal speakers. Results of the study showed that the prototype system was able to produce the pitch patterns similar to those in natural utterances.

Keywords: Electrolarynx, MEMS accelerometer, hands free

1. Introduction

People who have had laryngectomies have several options for the restoration of speech, but no currently available device is satisfactory. The artificial larynx, typically a handheld device which introduces a source vibration into the vocal tract by vibrating the external walls, is the easiest for patients to master, but does not produce airflow, so the intelligibility of consonants is diminished and the speech is uttered at a monotone frequency. Alternatively, esophageal speech does not require any special equipment, but requires utterance at a monotone frequency. Speech generation was tested with various forearm movements, and then a simple and small action was chosen to control the device. A simple comparison study has been made with well-trained normal speakers. Results of the study showed that the prototype system was able to produce the pitch patterns similar to those in natural utterances.

Keywords: Electrolarynx, MEMS accelerometer, hands free

For that reason, the electrolarynx is an important device even for the people who use esophageal speech. As for the advantages of EL, firstly, one can speak in long sentences even for the people who use esophageal speech because of the waning strength. When laryngectomized patients get older, they face difficulty in mastering the esophageal speech or keep using esophageal speech because of the waning strength. For that reason, the electrolarynx is an important device even for the people who use esophageal speech. As for the advantages of EL, firstly, one can speak in long sentences that are easily understood. Secondly, no special care requirements are needed; the EL only has to be placed up against the neck and turned on. Thirdly, the EL can be used by almost everybody, regardless of the post-operative changes in the neck. In those few cases where scarring prevents proper placement of the EL, an intraoral version can be used. On the other hand, there are a couple of disadvantages. Firstly, the EL has a very mechanical tone that does not sound natural. There usually is little change in pitch or modulation. Secondly, one must use their hand to control the EL all the time, and its appearance is far from normal. An EL system that has a hands-free user interface could be useful for enhancing communication by alaryngeal talkers, especially in hands busy environments. Also, the appearance can be almost normal even though the system requires slight hand movement. Almost all people frequently use gestures when they talk. It would be quite convenient if the EL users could utilize gestures to control the device because of its hands-free features. Furthermore, gesture control has a lot of potential to control not only just on/off function, but also many other various functions. Pitch frequency control is one of the important mechanisms for EL users to be able to generate naturally sounding speech. There are many studies of pitch controlling methods[1]-[3]. In fact, a couple of EL devices with pitch control mechanism are commercially available now[4][5]. However, none are hands free. The present study was undertaken to explore the feasibility of using gesture control method to replace the conventional EL user interface in terms of both on/off function and pitch control. Also, a wrist-watch type EL control device was designed and evaluated in order to determine the actual speech generation performance in a real environment. The specific goals were: 1) to determine the practical hands free user interface method for EL system, and 2) to determine whether the generated speech has high intelligibility and naturalness.

2. Determination of Needs - Survey Results

2.1 User Profile A set of techniques - including user observations, interviews, and questionnaires - were used to understand implicit user needs. As for the question-

Paper
naire survey, the total number of laryngectomized participants was 121 (87% male, 13% female), including 65% esophageal talkers, 12% EL users, 7% both, and 21% used writing messages to communicate.

2.2 Survey Results Almost all of the participants claimed that most public areas are difficult for oral communication due to the noisy environment. Typical public areas include train stations, inside of train cars, inside of vehicles, restaurants/pubs, and conventions/gatherings. Some of the needs confirmed from the survey are:

- Naturally sounding voice, not like mechanical tone
- Light weight device
- Smaller device, low profile
- Hands-free, easy to use
- Low cost

Based on the survey results, the present study was conducted to meet the essential user needs.

3. Hands Free User Interface Design

3.1 Gesture Control Gesture control UI can be developed through the use of a system based on photo detector, camera, or accelerometer. Based on the survey results, a 3-axis MEMS accelerometer was used in this study. MEMS sensors are very small, low cost, and a fit the system requirements well.

3.2 Pitch Control A MEMS accelerometer accurately measures acceleration, tilt, shock and vibration in applications. The challenge in designing the pitch control algorithm that use a MEMS accelerometer output to control pitch contour is to reconcile the numerical ranges between two types of data. MEMS output bytes are integers in the range -128 to 127 for a range of ±2G. Often this issue can be easily reconciled by linear mapping of one range of values (such as MEMS data values -128 to 127) into another range (such as 67 to 205 expected as the typical male pitch range).

Another possible pitch control method is to utilize a pitch contour generation model, such as Fujisaki’s model [6]. The system needs to have a strategy to extract both the phrase component and the accent component from the MEMS output. The model based method is easier to generate relatively stable pitch contour, however, it may lose some flexibility to generate various pitch patterns.

In this study, the simple linear mapping method was used to evaluate the pitch control performance. However, we plan to run the comparison study between the linear mapping method and the model-based method as one of our next phase tasks.
4. System Implementation

4.1 Hardware System Design  The pitch control algorithm described above was implemented on a small ARM cortex-M0 CPU board. A block diagram of the hardware architecture is shown in Figure 1. EL transducer with neckbandage has also been prepared to place it to the optimal location on the neck. We introduced the ARM board in order to meet the user requirements, i.e. small, comfortable weight, and low cost. The ARM-based hardware unit consists of a small board (34mm×34mm) with a 48MHz C1114, a 32 bit ARM cortex-M0, a ten bit PWM with 10kHz sampling rate, a USB interface, 32kB FLASH memory, and three 1.5V batteries. Figure 2 shows the ARM unit and the EL transducer.

4.2 Pitch Control  Hand gestures are a very important part of language. A preliminary UI study using forearm movement was conducted in order to evaluate feasibility of the pitch control mechanism. Figure 3 shows the forearm tilt and the MEMS output (x-axis) when the controller was placed on the wrist. From the horizontal position (0°) to the 75° upward position is the normal pitch control zone. From the horizontal position to the -25° downward position is the fading out zone, where phrase ending pitch pattern is adjusted based on the forearm moving speed. As for the conversion from the MEMS output to the pitch frequency, there are four pitch ranges. Figure 4 shows the relation between the MEMS output and the four ranges of pitch frequency, i.e. high, mid-high, mid-low, and low. Users can select one of the four ranges.

4.3 ON/OFF Control  Reliable EL ON/OFF control is very important for users to talk comfortably. As you can see in Fig.3, EL vibrates at the normal pitch control zone, i.e. from 0° position and higher. EL stops the vibration at the -25° position or lower. The hysteresis is necessary to avoid unstable behavior near the on/off threshold. If the phrase does not have an accent, the pitch rises from a low starting point on the first mora, and then levels out. Such pitch contour is generated by moving the forearm downward very quickly. However, most of the accented phrases are generated by gradual movement.

4.4 LOCK Mechanism  It is very important to enable/disable the controller easily and quickly while users are wearing the device. Y-axis output of the MEMS accelerometer was used to implement such a lock mechanism. By twisting the wrist quickly and generating 2G acceleration, the user can enable/disable the EL.

5. Evaluation  A preliminary usability test was performed. Three normal speakers participated in this study. After 30 minutes of practice, 10 test sentences (Table 1) were pronounced using their own voice and using the prototype. Then, the pitch contours were analyzed. Figure 5 shows one of the comparison results. The comparison study showed that the overall pitch trends were similar, however, more practice is necessary to control the pitch precisely.

Table 1: Phonetically balanced Japanese test sentences.

<table>
<thead>
<tr>
<th>No.</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Papa mo mama mo minna de mamemaki o shita.</td>
</tr>
<tr>
<td>2</td>
<td>takai takai tokoro e nobette ikuto koroda.</td>
</tr>
<tr>
<td>3</td>
<td>Achirakara mo kochirakara mo dohirakara mo iukotoko ga dekiru.</td>
</tr>
<tr>
<td>4</td>
<td>Aoi o ueru.</td>
</tr>
<tr>
<td>5</td>
<td>Anohito wa bunkajin to yobareru no ga fusawashi.</td>
</tr>
<tr>
<td>6</td>
<td>Shichi gatsu kara hanzhin densha de tukin shite ima su.</td>
</tr>
<tr>
<td>7</td>
<td>Ginko mo gakkyo aruite ikeru kyori ni ari masu.</td>
</tr>
<tr>
<td>8</td>
<td>Kinko ga tore te iru no de kakko ga yoi.</td>
</tr>
<tr>
<td>9</td>
<td>shiu gamu o kamuno ga syukan ni natte iru.</td>
</tr>
<tr>
<td>10</td>
<td>Hana o ottari ana o hortari sanzanna meni atta.</td>
</tr>
</tbody>
</table>

Then, subjective evaluation tests (by rating scale method) have been made with 2 male well-trained normal speakers, and 10 (one female and nine male) subjects. Each speaker read also the phonetically balanced test materials as shown in table 1. We used one commercially available EL device (SECOM EL-X0010), prototype-A (linear mapping method, with 70Hz mode). Those 40 speech stimuli (2speakers*2devices*10sentences) were recorded, and two sets of differently randomized stimuli were prepared. 5 subjects evaluated the one set of stimuli, and other 5 subjects rated the other set of stimuli. Each speech stimuli were presented two times. The subjects rated the speech stimuli in terms of “intelligibility (Clarity)”, “naturalness of the prosody” and “stability of the prosody” using five level scaling. As shown in Fig. 6, the subjective evaluation indicated that the prototype-A(LM) obtained higher naturalness score than the EL device(EL). On the other hand, intelligibility(clarity) and stability shows almost no difference among those devices.
6. Discussion

Without losing intelligibility (clarity) and stability of the prosody, the prototype-A showed substantial improvement in terms of the naturalness of the prosody. The results of this study indicate that both usability and speech quality of EL speakers could be improved by MEMS accelerometer based hands-free UI controller. The ability to control the pitch contour of EL speech with the proposed linear mapping method implies that hand gesture control may be adequate for implementation of the hands free user interface for EL device. We plan to run the same evaluation with actual EL-users, and confirm if the proposed methods show similar performance. Also, a more detailed and precise study across the talkers, sentences, and learning curve has to be performed. As for the gesture control, we tested only the forearm movement, however, it is necessary to test other body locations where users might be able to control the EL device more easily and naturally. According to the user requirements, the evaluation of appearance also needs to be considered. In the study, we set a relatively narrow pitch range in order to avoid wild swings in pitch. A better pitch control range needs to be investigated.

7. Conclusion

MEMS accelerometer based hands free UI for EL device was proposed, and a hand gesture control unit was designed. Results of the preliminary evaluation indicated that the proposed method has a potential to make the EL output prosody more natural, easy to use, and less distinct appearance. However, users may be required to undergo some training before utilizing the device comfortably. The device also needs further evaluation in terms of how to generate the pitch, where to wear the device, and how to make the device less conspicuous.

Acknowledgment

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) Grant Number 24500664.

References

[5] Griffin laboratories, “All Products by Griffin Laboratories”, http://www.griffinlab.com/Products/default.html, access date: 2013.06.09

Kenji Matsui (Member) received a Ph.D. degree in electronics and information engineering from Osaka City University in 2002, and is presently a professor at Osaka Institute of Technology. He has worked on speech signal processing. He is a member of IIAE, IEICE, and ASJ.

Yoshihisa Nakatoh (Non-member) received his Ph.D degree from Shinshu University in 2007. He worked for Sharp Corporation from 1986 to 1989 and worked for Matsushita Electric Industrial Co. Ltd from 1991 to 2008. He is currently a professor at Kyushu Institute of Technology. His research interests include speech recognition, audio coding, hearing aid and accessibility technology. He is a member of ASJ, IEICE, IPSJ.

Yumiko O. Kato (Non-member) received a Ph.D. degree in medical sciences from St. Marianna University School of Medicine in 2006, and is presently a researcher at St. Marianna University School of Medicine. She has worked on otology and location perception.