Development of Electrolarynx with Hands-Free Prosody Control

Kenji Matsui[1], *Kenta Kimura*[1], *Yoshihisa Nakatoh*[2], *Yumiko O. Kato*[3]

[1] Osaka Institute of Technology, Osaka, Japan
[2] Kyushu Institute of Technology, Kitakyushu, Japan
[3] St. Marianna University School of Medicine, Kawasaki, Japan

matsui@elc.oit.ac.jp

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 pitch frequency control. The vibration device was placed against the throat using support bandage. Very small battery operated ARM-based control unit was developed and placed on the wrist. The control unit has a function to convert the tilt angle into the pitch frequency, as well as the device enable/disable function and pitch range adjustment function. As for the forearm tilt angle to pitch frequency conversion, two different conversion methods, linear mapping method and F0 template-based method, were investigated. A perceptual evaluation, with two well-trained normal speakers and ten subjects, was performed. Results of the evaluation study showed that both methods were able to produce better speech quality in terms of the naturalness.

Index Terms: prosody, electrolarynx, hands-free

1. Introduction and Related work

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 hand-held 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 speakers to insufflate, or inject air into the esophagus, and limits the pitch range and intensity. Both esophageal speech and tracheo-esophageal speech are characterized by low average pitch frequency, large cycle-to-cycle perturbations in pitch frequencies, and low average intensity. As for utilizing esophageal speech, it was found that age was the important factor. 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.

Pitch frequency control is one of the important mechanisms for EL users to be able to generate naturally sounding speech. There are some commercially available EL devices using a single push button with pressure sensor to produce F0-contours [1], [2]. There are also similar studies of pitch controlling methods [3], [4]. However, none are hands-free. The approaches for generating F0-contour without manual interaction have been proposed. Saikachi et al. use the amplitude variation of EL speech [5]. Another approach is to generate a F0-contour using an air-pressure sensor that is put on the stoma [6], [7]. Also, recently, a machine learning F0-contour generation from the EL speech has been proposed [8]. Although most of those studies are in an early stage of research, the results show substantial improvement of EL speech quality.

An EL system that has a hands-free user interface could be useful for enhancing communication by alaryngeal talkers. Also, the appearance can be almost normal because users do not need to hold the transducer by hand against the neck. 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. Furthermore, gesture control has a lot of potential to handle not only just on/off function, but also many other functions because hands can generate various types of motion. However, if users can not even use hand gestures, for example, controlling something, we need to consider other part of body movement, or completely different technique, such as EMG based hands-free EL control [9].

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 questionnaire 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.
The noisy environment issue is well known problem and people usually use portable amplifier, however, we have been investigating smaller, lighter, and low profile speech enhancement system for both esophageal speech [10] and EL.
Other 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 those survey results, present study was conducted to meet the essential user needs.

3. Hands Free UI 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 three-axis MEMS accelerometer was used in this study. MEMS sensors are very small, low cost, and fit the system requirements well [12].

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 [11]. The system needs to have a strategy to generate both the phrase component and the accent component from the MEMS output. The F0 template-based method is easier to generate relatively stable pitch contour, however, it may lose some flexibility to generate various pitch patterns.

In this study, both the simple linear mapping method and the F0 template-based method were prototyped and examined to evaluate the pitch control performance. Also, the comparison study was performed between conventional EL, the linear mapping method and the F0 template-based method.

4. System Implementation

4.1. Hardware System Design

The pitch control algorithms described above were implemented on a small CPU board. A block diagram of the hardware architecture is shown in Fig. 1. A pair of very small EL transducer with neck-bandage has also been prepared to place it to the optimal location on the neck. We introduced the small hardware 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 10 bit PWM with 10kHz sampling rate, a USB interface, 32kB FLASH memory, and three 1.5V batteries. Picture 1 shows the ARM unit and the EL transducers.

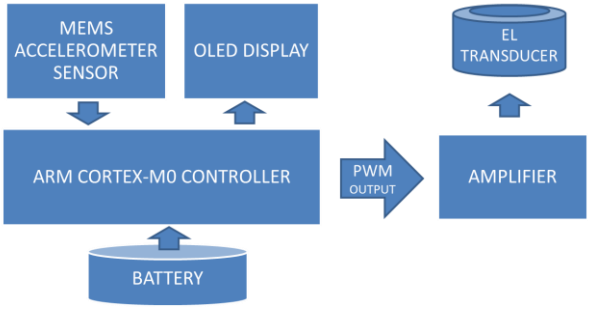
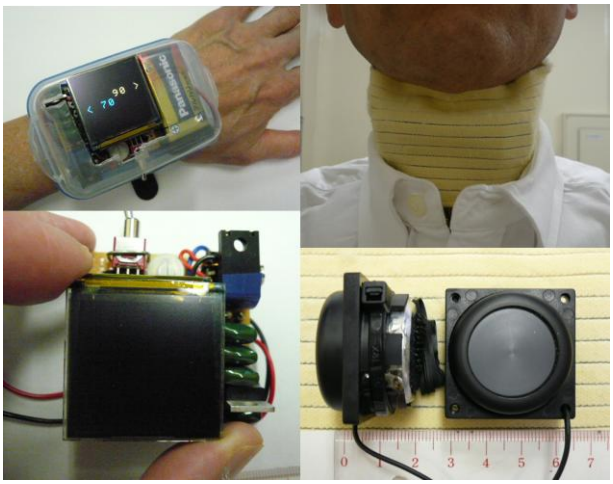


Figure 1: *Block diagram of the Hardware Architecture*



Picture 1: *EL Controller and Transducer unit (upper left: EL controller on the wrist, lower left: ARM-PC board, upper right: transducer with neck bandage, lower right: transducer unit)*

4.2. Pitch Control (linear mapping method)

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. Fig.2 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. Fig.3 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.

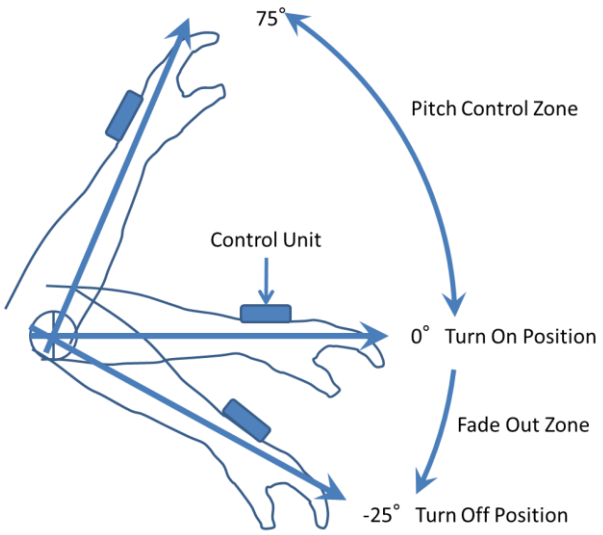


Figure 2: *Forearm Tilt and Pitch Control*

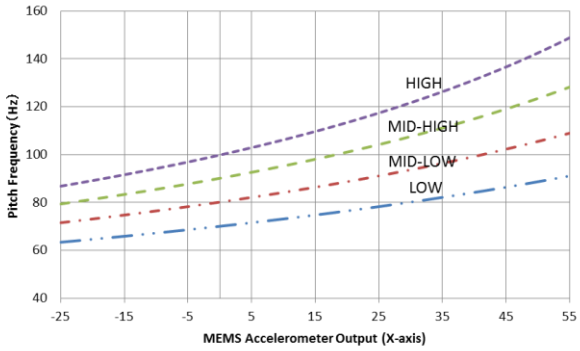


Figure 3: *Relation between MEMS Output and Pitch (linear mapping method)*

4.3. Pitch Control (F0 tmplate-based method)

The linear mapping method is straight forward approach, however, it requires precise sensor control to avoid unnatural pitch behavior. The F0 template-based method applies a basic F0 template to the fine F0 contour generation. The phrase component of Fujisaki's model was used to generate the F0 template *F0*(*t*). While the system is intended to generate both phrase control and accent control, as the first step of testing the template, we utilized only the phrase component.

$$\ln F0(t) = \ln F\min + Ap \cdot Gp(t) \quad (1)$$

where

$$Gp(t) = \alpha^2 t \exp(-\alpha t) \quad (2)$$

The symbols in equations(1) and (2) indicate:

- *F*min is the minimum value of speaker's F0.
- *Ap* is the magnitude of phrase command.
- α is natural angular frequency of the phrase control mechanism.

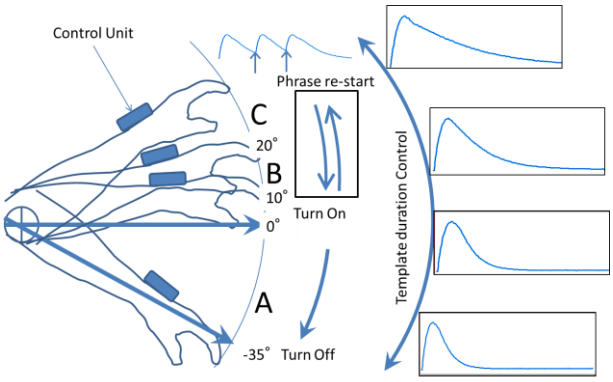


Figure 4: *Relation between Forearm tilt and F0 template generation (F0 model-based Method) A-zone: -35°~0°, B-zone: 0°~20°, and C-zone: 20°~.*

In this study, those values are; *F*min = 80Hz, α =1.5, and *Ap*=0.75. The calculated F0 template data is stored in the controller software. Fig.4 shows the F0 contour generation mechanism using the MEMS sensor output and the F0 template. The oscillation starts at 10° upward from the horizontal position. The template duration is controlled based on the forearm tilt angle as shown in the Figure 4. Also, in that figure, how to re-start the F0 template is shown. Basically, the forearm movement (C-zone → B-zone → C-zone) is required. A-zone is -35°~0°, B-zone is 0°~20°, and C-zone is 20°~, respectively.

4.4. ON/OFF Control

Reliable EL ON/OFF control is very important for users to talk comfortably. As you can see in Fig.2, 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. In case of the F0 template-based approach, in order to turn on the EL device, the tilt angle needs to be 10° or higher, and the turning off position is at -35°.

4.5. 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. Perceptual Evaluation

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 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), and prototype-B (F0 template-based method). Those 60 speech stimuli (2speakers ∗3devices∗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.

Table 1: *Phonetically balanced Japanese test sentences*

1	Papa mo mama mo minna de mamemaki o shita.
2	takai takai tokoro e nobotte iku tokoro da.
3	Achirakara mo kochirakara mo dochirakara mo ikukoto ga dekiru.
4	Aoi o ueru.
5	Anohito wa bunkajin to yobareru no ga fusawashî.
6	Shichi gatsu kara hanshin densha de tûkin shite ima su.
7	Ginkô mo gakkô mo aruite ikeru kyori ni ari masu.
8	Kinkô ga tore te iru no de kakkô ga yoi.
9	shijû gamu o kamuno ga syûkan ni natte iru.
10	Hana o ottari ana o hottari sanzanna meni atta.

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 Figure 5-(a), 5-(b) and 5-(c), the subjective evaluation indicated that both prototype-A(LM) and B(FU) obtained higher naturalness scores than the EL device(EL). On the other hand, intelligibility(clarity) and stability shows almost no difference among those devices.

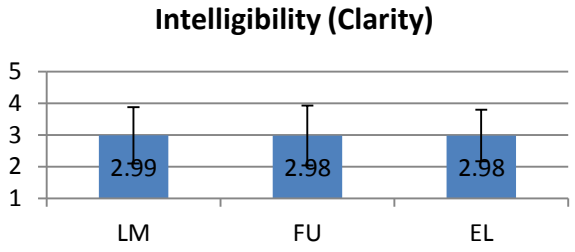


Figure 2-(a): average evaluation scores of intelligibility

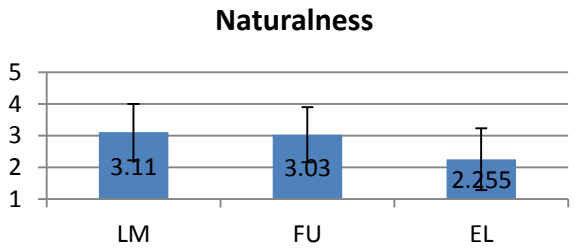


Figure 5-(b): average evaluation scores of naturalness

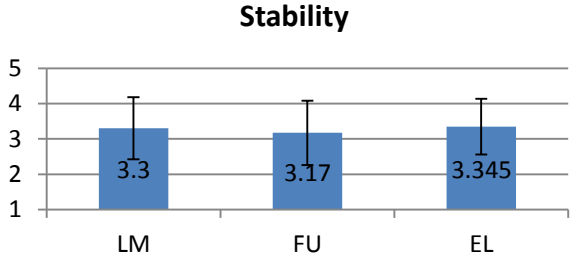


Figure 5-(c): average evaluation scores of "Stability"

6. Discussion

Without losing intelligibility(clarity) and stability of the prosody, both prototype-A and B 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 and F0 template-based method implies that hand gesture control may be adequate for implementation of the hands free user interface for EL device. Our assumption about the performance difference between the two proposed methods is that the F0 template-based method may be easier to learn and easier to stabilize the pitch contour, however, there was almost no difference between those two methods. 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. A hand gesture control unit was designed and prototyped. Two types of pitch contour generation methods were proposed and tested together with conventional EL device. Results of the evaluation indicated that the proposed methods have a potential to make the EL output prosody more natural, easy to use, and less distinct appearance. However, a more detailed and precise study across the talkers, sentences, and learning curve has to be performed.

8. Acknowledgements

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

9. References

[1] SECOM company Ltd., Electrolarynx "MY VOICE", (http://www.secom.co.jp/personal/medical/myvoice.html)
[2] Griffin laboratories, TruTone users guide.
[3] Y. Kikuchi, and H. Kasuya, : "Development and evaluation of pitch adjustable electrolarynx", In SP-2004, 761-764, 2004
[4] H. Takahashi, M. Nakao, T. Ohkusa, Y. Hatamura, Y. Kikuchi, and K. Kaga, 2001. Pitch control with finger pressure for electrolaryngial or intra-mouth vibrating speech.Jp. J. Logopedics and Phoniatrics, 42(1),1-8.
[5] Y. Saikachi, "Development and Perceptual Evaluation of Amplitude-Based F0 Control in Electrolarynx Speech", Journal of Speech, Language, and Hearing Research Vol.52 1360-1369 October 2009
[6] N. Uemi, T. Ifukube, M. Takahashi and J. Matsushima, "Design of a new electrolarynx having a pitch control function", In Proceedings of 3rd IEEE International Workshop on Robot and Human Communication, RO-MAN p.198-203, Nagoya, Japan, July 18-20, 1994.
[7] K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, "The use of air-pressure sensor in electrolaryngeal speech enhancement",

INTERSPEECH, p.1628-1631, Makuhari, Japan, Sept 26-30, 2010.

[8] A. K. Fuchs and M. Hagmüller, "Learning an Artificial F0-Contour for ALT Speech", INTERSPEECH, Portland, Oregon, Sept. 9-13, 2012.

[9] H.L. Kubert, "Electromyographic control of a hands-free electrolarynx using neck strap muscles", J Commun Disord. 2009 May-Jun;42(3):211-25

[10] K. Matsui, et al., "Enhancement of Esophageal Speech using Formant Synthesis", Journal of Acoustical Society of Japan (E) 23,2 pp.66-79, 2002

[11] H. Fujisaki, In Vocal Physiology: Voice Production, Mechanisms and Functions, Raven Press, 1988

[12] K. Matsui, et al., "A preliminary user interface study of speech enhancement system", Proc. of the 1st International Conference on Industrial Application Engineering 2013, 53-56