Real-time clarification of esophageal speech using a comb filter

A Hisada and H Sawada

Department of Intelligent Mechanical Systems Engineering, Faculty of Engineering, Kagawa University
2217-20, Hayashi-cho, Takamatsu-city, Kagawa, 761-0369, Japan

sawada@eng.kagawa-u.ac.jp

http://www.eng.kagawa-u.ac.jp/~sawada/index_e.html

ABSTRACT

The only treatment for the terminal symptoms of the laryngeal cancer is to remove the larynx including vocal cords, which mean that the patient loses his voice. Esophageal speech is a method of speech production using an esophagus. An air inhaled in the upper esophagus generates the esophagus vibration to produce a belch-like sound that can be shaped into speech. Although this method has difficulties to be mastered, the voice is able to keep the individuality since the speech is generated by his own vocal organs. This paper introduces a software filtering algorithm which clarifies esophageal speech with the individuality preserved, together with user's evaluations by questionnaires.

1. INTRODUCTION

Although almost all the animals have voices, only human beings are able to employ speech to make verbal communication. The vocal sound is produced by the relevant operations of the vocal organs such as a lung, trachea, vocal cords, vocal tract, tongue and muscles. The airflow from the lung causes a vocal cord vibration to generate a source sound, then the glottal wave is led to the vocal tract, which works as a sound filter as to form the spectrum envelope of a specific voice. If any part of the vocal organs is injured or disabled, we may be involved in the impediment in the vocalization and, in case of the worst, we may lose our voices.

Over 20,000 patients are currently suffered from laryngeal cancer in Japan, and the only treatment for the terminal symptoms is to remove the larynx including vocal cords, which means that the patient loses his voice. Losing voice causes various difficulties in the communication with other people, since the employment of a voice is essentially important for us to make verbal communication.

Incidentally there are mainly two ways to recover voice. One is to use an artificial larynx, which is also called an electrolarynx and is a hand-held device with a pulse generator that produces a vocal cord-like vibration. The electrolarynx has a vibrating plastic diaphragm, which is placed against the neck during the speech. The vibration of the diaphragm generates a source sound in the throat, and the speaker then articulates with the tongue, palate, throat and lips as usual. The other is to learn esophageal speech, which is a method of speech production using an esophagus (Sato, 1993; Max *et al,* 1996). In the speech air is inhaled and caught in the upper esophagus instead of being swallowed, and then the released air generates the esophagus vibration to produce a "belch-like" sound that can be shaped into speech.

The former has an advantage to be used by just being held to the neck and to be easily mastered, but the sound quality is rather electronic and artificial. Furthermore one hand is occupied to hold the device during the speech, which disturbs the gestural communication. On the other hand, the latter method has difficulties to be mastered (several years are ordinary required for the practice), but the voice is able to keep the speaker's individuality since the speech is generated by his own vocal organs, although several distinctive characters exist. Moreover the body parts such as hands and facial expressions are freely and actively used for the communication to assist the speech. Mastering the esophageal speech requires repetitious training, and the progress is slow and is hard to be recognized by the patient himself.

This paper introduces a real-time software filtering algorithm which clarifies esophageal speech with the individuality preserved, and presents the results of a listening experiment to evaluate the filtering ability.

2. BACKGROUND OF THE STUDIES CONCERNING THE ESOPHAGEAL SPEECH

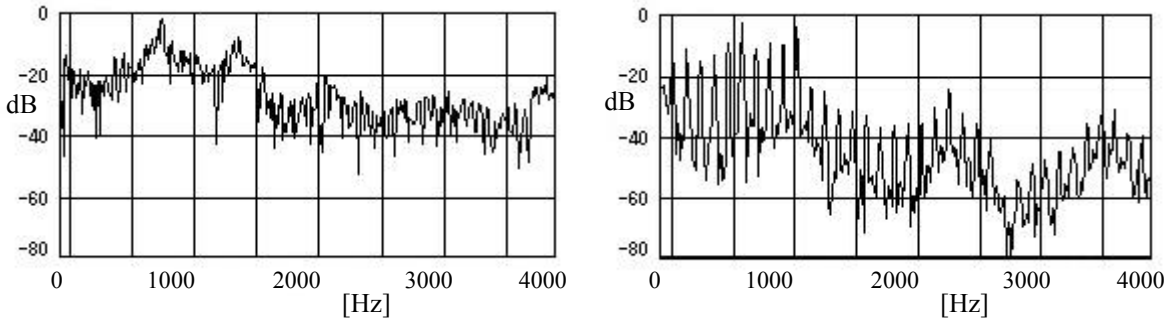
Several researches which analyze the characteristics of the esophageal speech have been reported so far (Noguchi & Matsui, 1996; Doi *et al*, 1996), and a device to improve the esophageal voice is now commercially available (e.g. VivaVoice by Ashida Sound Co., Ltd.). The device is a package of a small circuit board equipped with a compact microphone and speaker, and transforms an unclear voice into clear by using the formant synthesis techniques, which has the disadvantage in keeping the individuality in the voices. Acoustic characteristics of the esophageal voice have been also studied (Lu *et al*, 1997;Bellandese *et al*, 2001; Robbins *et al*, 1984), and have accounted for the lack of the phonetic clarity. There still remain unknown characteristics and features in the esophageal voices, however, the clarification theory and algorithm are not yet established.

3. PURPOSE OF THE STUDY

An example of the spectrum of esophageal voice is shown in Figure 1 with the comparison of the normal laryngeal vocalization. Typical esophageal speech is characterized as below.

1) The voice contains noises caused by the irregular vibrations of esophagus.
2) Fundamental frequency is rather low.
3) Fundamental frequency and its overtone structure are not clear.
4) Fluctuation exists, which causes the instability of the voice.
5) Spectrum envelope is flatter than the laryngeal voice.
6) Volume is low because of the shortage of the expiration. (Average is 150 ml; laryngeal voice is 2000ml)

By considering the features mentioned above, we constructed a software algorithm by controlling a digital comb filter, which was considered to be applied effectively to control the noises and overtone structures for the clarification of the esophageal speech in real time.



(a) Esophageal voice *(b) Laryngeal voice*

Figure 1. *Spectrum of Japanese vowel "a".*

4. FILTERING ALGORITHM

The filtering system is realized by using Max/Msp programming environment (Roads, 1996) on Macintosh Power Book (Power PC, 400 MHz) equipped with a microphone and a speaker. Sampling frequency was set to 44.1kHz with 16bits input. We paid attention to the characteristics 1), 3), 4) and 5), and applied the comb filter for the clarification.

Figure 2 shows the flow diagram of the filtering process. First the input signal is judged to distinguish vowels and consonants, and is divided into two procedures. For the vowel sound, fundamental frequency is extracted to form a comb filter as to enhance pitch components. The consonant sound, on the other hand, is enhanced in the time domain by multiplying its amplitude by a suitable coefficient. The calculations above are processed in realtime for every input to generate clarified speech outputs.

4.1 Distinction between vowels and consonants

Comb filter is able to enhance pitch components. It means the filter affects opposite effect for the consonants

which can often be approximated as noise components. Since the consonant sound basically doesn't have the overtone structure, the comb filtering performs inadequately. Because of the different characteristics of sound phonemes to be clarified, vowels and consonants were required to be classified in advance to employ the different filtering techniques. We paid attention to the differences of the spectrum envelope observed in the frequency domain, and have proposed a filter to distinguish with each other.

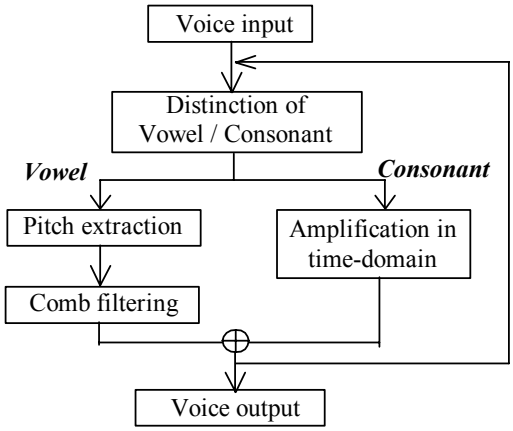
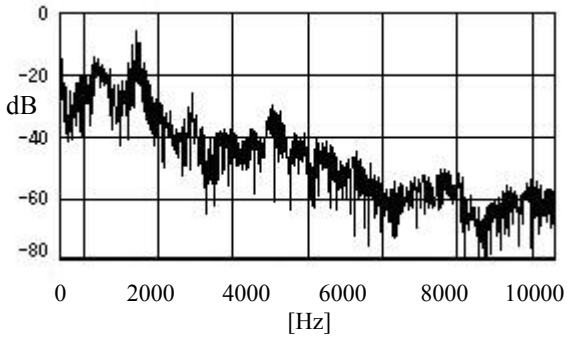
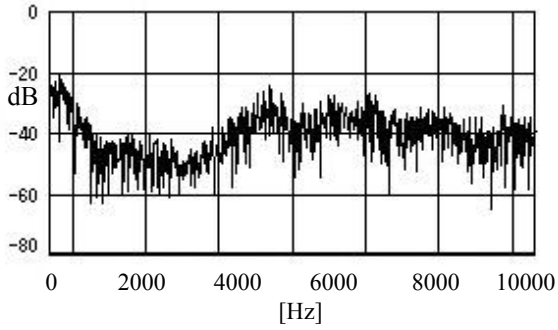


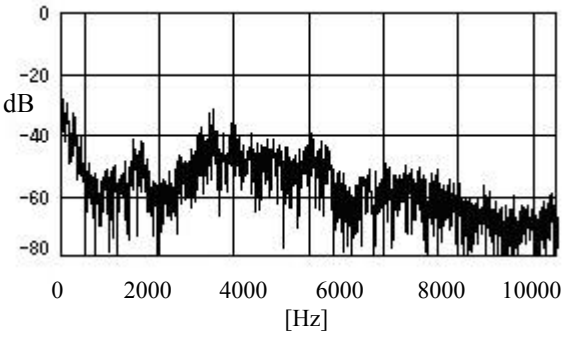
Figure 2. *Flow diagram of filtering*



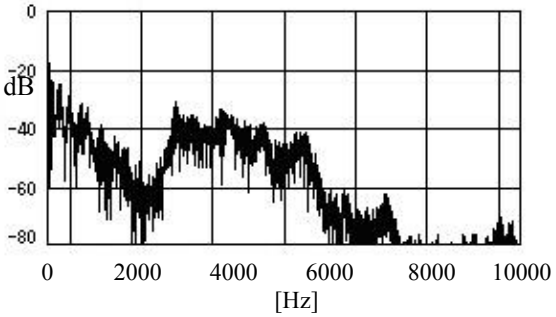
(a) Esophageal speech "a"



(b) Esophageal speech "s"



(a) Esophageal speech "k"



(b) Esophageal speech "n"

Figure 3. *Comparison of vowel and consonant.*

Figure 3 shows the power spectra of a Japanese vowel "a" and consonants "s", "k" and "n" for the comparison of each characteristics. Since the spectrum of the consonant is broadly distributed in the frequency range, the power spectra are presented by 10kHz in the figures.

Notable differences were found between the two spectra. The spectrum envelope of the vowel sound shows a couple of peaks in the lower frequency range, which are characterized as the formants, and the power gradually decreases as the frequency rises. The consonant sounds, on the other hand, are characterized as noises, and their spectra are uniformly distributed along the frequency range without a specific peak. The above observations justify the assumption that vowels and consonants can be separated by comparing the power balance between the low frequency range and the high frequency range. We defined an index as

$$B_{index} = \frac{\text{Average power of low frequency range}}{\text{Average power of high frequency range}}$$

where the low frequency range was set to 200Hz ~ 4kHz and the high frequency range was 6kHz ~ 10kHz. For vowel sounds, the index value becomes larger because of the concentration of the power in the lower frequency range, on the other hand, the value becomes around 1 for consonant sounds. The threshold for the distinction was set to 1.5 by the experiment, which was not influenced by the amplitude of the input signal.

4.2 Experiment of distinction between vowels and consonants

We conducted an experiment of the separation of vowels from consonants by using the method mentioned above. Table 1 shows the result, in which 1 and 0 represent the recognition results of the vowels and the consonants respectively.

In the experiment, the algorithm could output correct answers for all the fricative and plosive consonants such as “k”, “s”, “p” and “t”, which are regarded as the voiceless consonants. The other consonants like “n”, “m” and “r” were, however, sometimes mis-recognized as vowels, since these sounds are categorized as voiced consonants which attend the vocal cord vibration. Fundamental frequency exists, though it is almost indistinct, and the negative effects of the comb filter were not found in the filtering of the voiced consonants. In this experiment, the mis-recognitions for the voiced consonants were overlooked.

Table 1. *Result of vowel/consonant separation (1: vowel, 0: consonant)*

s	a	s	i
0000000	11111111	000000	11111111
k	a	k	i
00000	111111111	00000	111111111
t	a	t	i
00000	11111111	000000	111111111

n	a	n	i
010111	11111111	11111	111111111
m	a	m	i
011111	11111111	001111	1111111111
r	a	r	i
0011111	11111111	111111	111111111

4.3 Comb filtering

As obviously comprehended by Figure 4 which shows the impulse response of the comb filter, the overtone structure of a comb filter enhances a fundamental frequency and its harmony components. By suitably operating comb filters, esophageal voices can be clarified with their overtone structure enhanced and their noise components controlled. The importance is to extract the fundamental frequency precisely. In this system we adopted a short time zero-cross method for the pitch calculation. The extracted pitch works to form a comb filter with its pitch span and enhancement ratio suitably defined. The individuality of a speaker can be also preserved, since the comb filter enhances the overtones without changing the formants and the spectrum envelope.

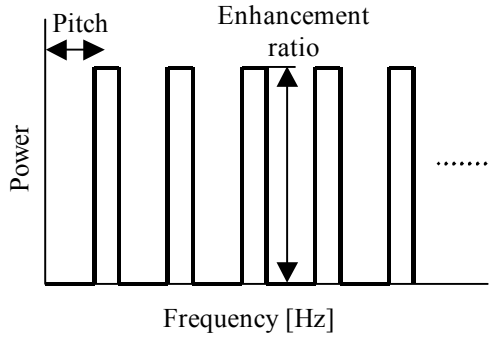
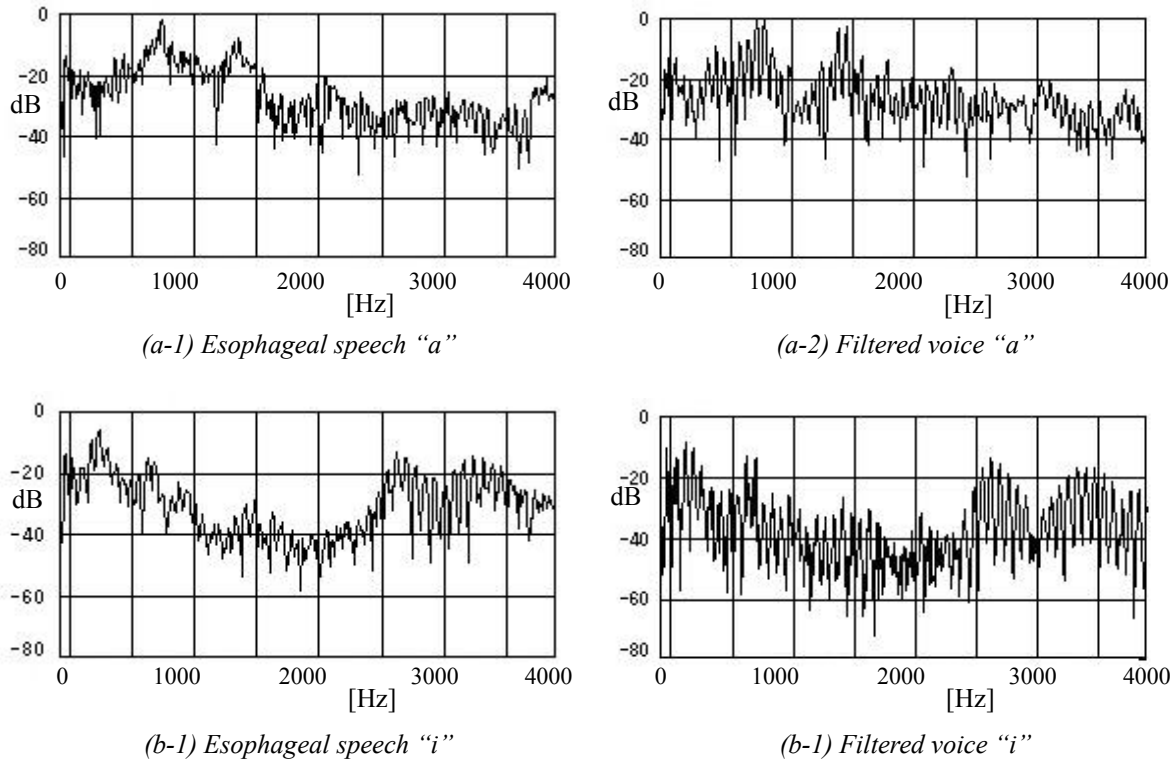


Figure 4. *Impulse response of a comb filter*

4.4 Experiment of comb filtering for vowels

We conducted an experiment of the filtering for esophageal vowels. Figure 5 shows the power spectra of esophageal speech “a” and “i” before and after the filtering.

The proposed filter could form the clear overtone structures of the fundamental frequencies. The spectrum envelope was preserved throughout the frequency range, and the peaks and valleys of the overtones were enhanced, which contributed to the noise reduction. The individuality of the speaker was well preserved and the improvement of the clarification was also recognized.



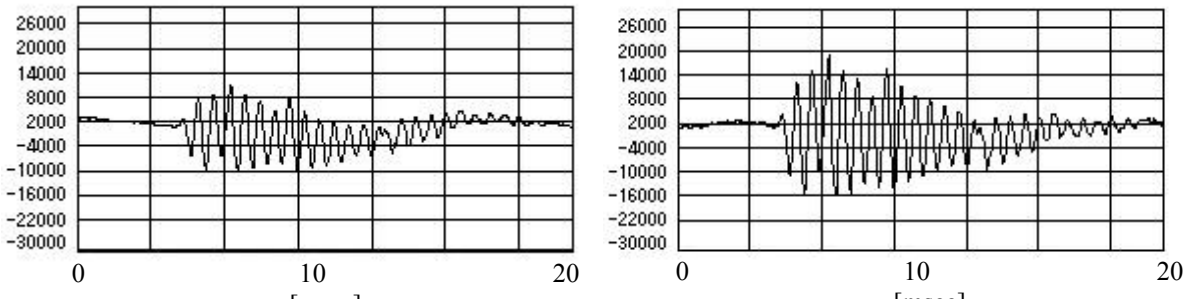
(a-1) Esophageal speech "a" *(a-2) Filtered voice "a"*

(b-1) Esophageal speech "i" *(b-1) Filtered voice "i"*

Figure 5. *Results of the comb filtering.*

4.5 Processing for the consonants

The volume of the esophageal speech is generally low because of the lack of the expiration pressure, and the consonant pronunciations are insufficient and difficult to be distinguished. The consonant sound is enhanced in the time domain by multiplying its amplitude by a suitable coefficient. Figure 6 shows the waveform of the enhanced consonant "k" with the original sound wave for the comparison.



Original consonant "k" Enhanced consonant "k"

Figure 6. *Waveforms of consonant "k" before and after the filtering*

By considering the characteristics of the esophageal speech described in the sections above, we constructed a real-time clarification filter in the Max/Msp programming environment as shown in Figure 7. The box (or block) in the program is called a graphical object, in which algorithms, computations or processes are described and stored by the programming language. By connecting graphical objects with lines, the filtering program can be built and executed.

To compare the filtered voice with the input esophageal speech, recorded voices are used for the listening experiment. The Max/Msp program shown in the Figure 7 opens a sound file at the beginning, and executes the filtering to output a voice from the speaker.

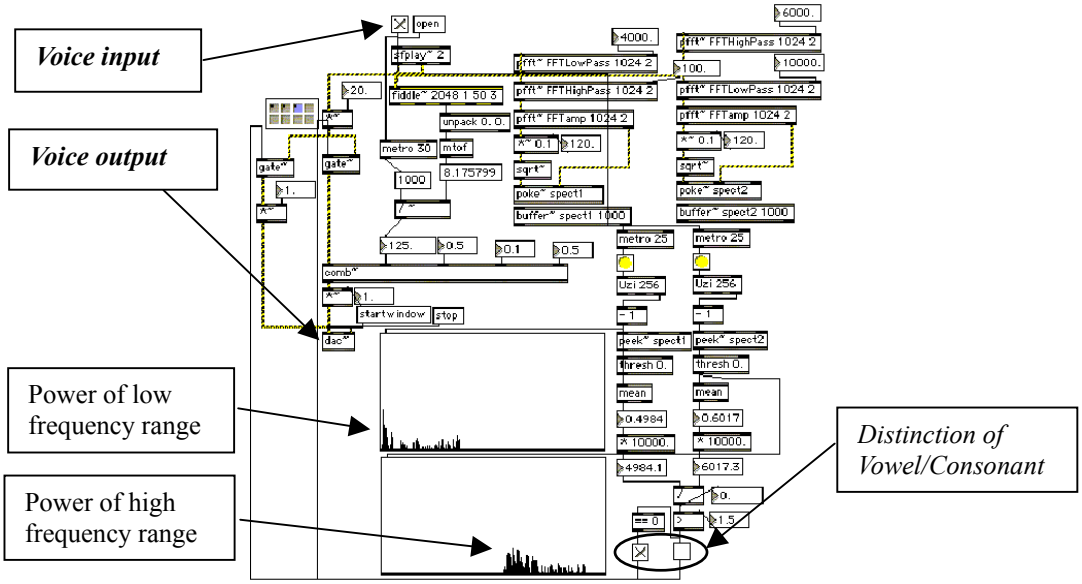


Figure 7. *Clarification program of the esophageal speech*

5. LISTENING EXPERIMENT

We conducted a listening experiment to evaluate the filtering ability by questionnaires. 11 able-bodied subjects listened to 12 pairs of unfiltered and filtered esophageal speech in random orders, and evaluated them in 5 levels from 11 points of view, which are

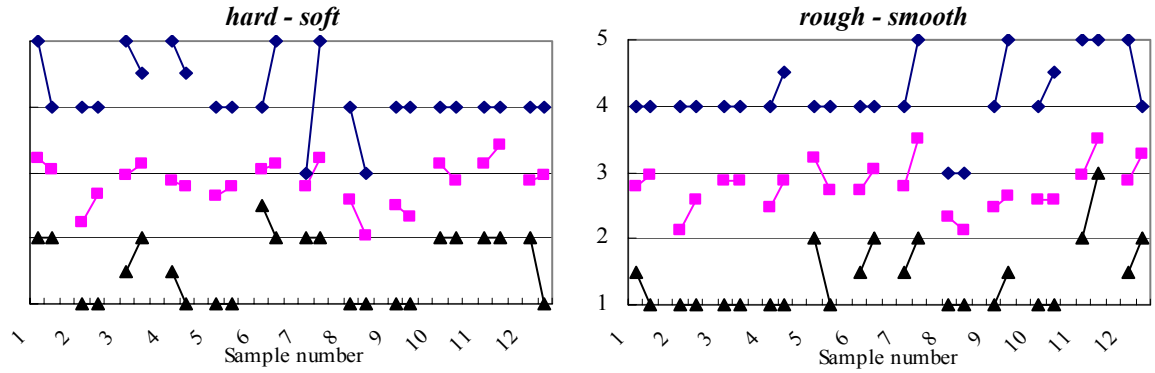
hard-soft, rough-smooth, unclear-clear, blur-sharp, unpleasant-pleasant, electronic-human, noise level, echo level, clarification level of consonants, clarification level of vowels, and total evaluation.

In the adjective pairs, the former was assigned as level 1 and the latter was rated as level 5, and the levels between 1 and 5 were rated according to the subject's evaluation. The larger the points are, the higher the filtering results were evaluated for the other evaluation points such as noise level and clarification level.

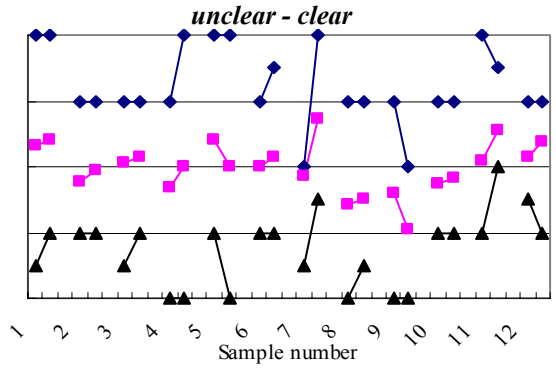
Figure 8 shows the results of the listening experiment, in which the maximum points are plotted with the mark ◆, the average points with ■, and minimum points with ▲. In the pair of each plot, the left one indicates the evaluation of the original speech and the right one gives the results of the filtering. If the right plot of a pair is higher, the filtering is considered to be effective.

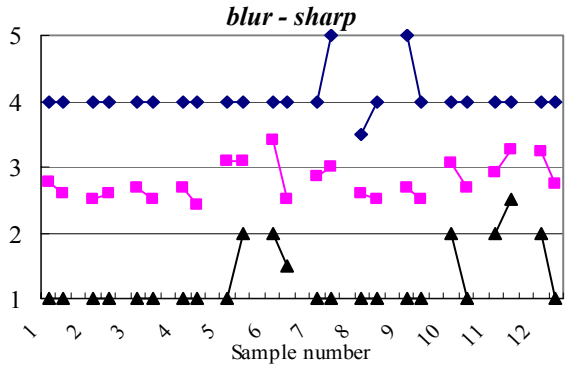
In almost all points, the filtered voices achieved higher evaluations, especially in the points of smoothness, clearness and sharpness. In the evaluations of ***noise level***, ***clarification level of consonants***, ***clarification level of vowels***, ***and total evaluation***, better impression were obtained as expected by the consideration of effectiveness of the comb filters.

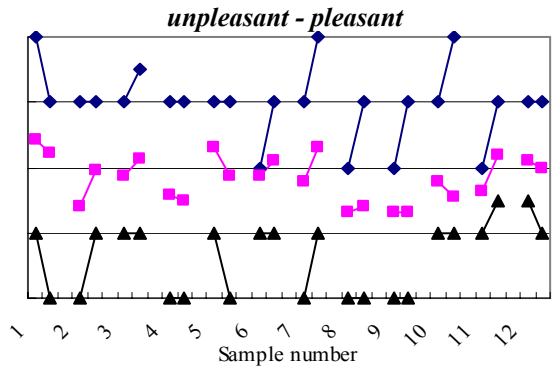
In the points of ***hard-soft*** and ***unpleasant-pleasant***, on the other hand, obvious differences could not be found. A few subjects pointed out the unexpected echo effects and the electronic impressions, which could have been caused by the mis-detection of the pitch.

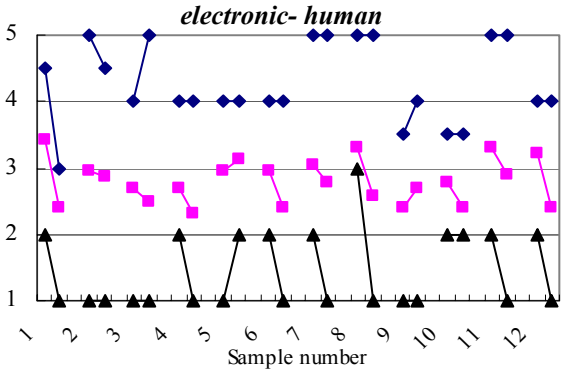


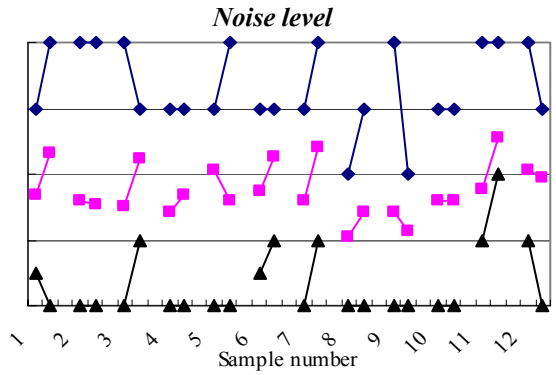
Proc. 4th Intl Conf. Disability, Virtual Reality & Assoc. Tech., Veszprém, Hungary, 2002

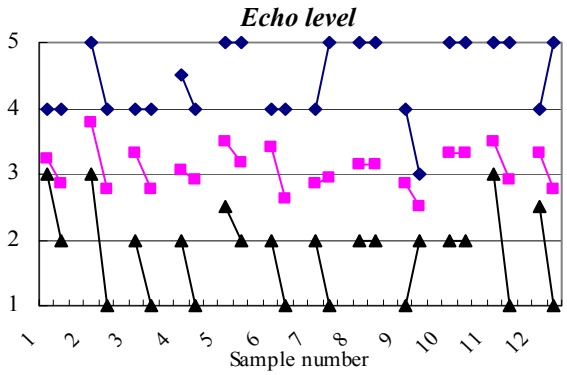


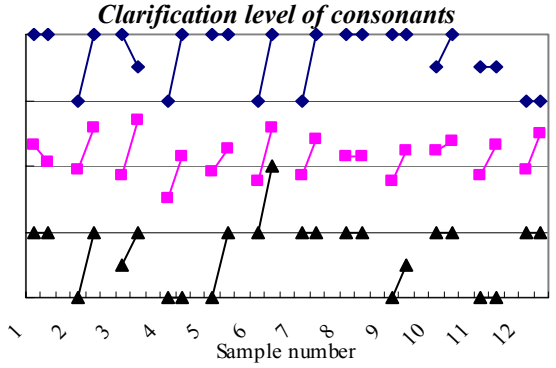


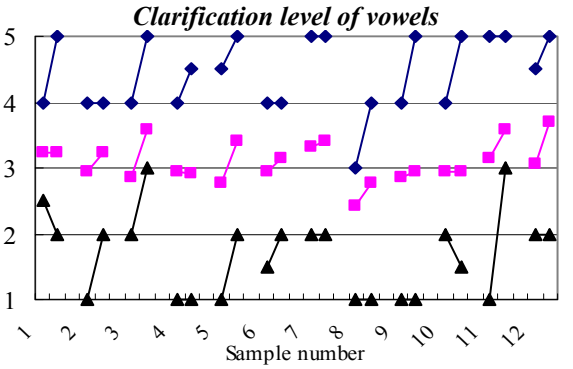












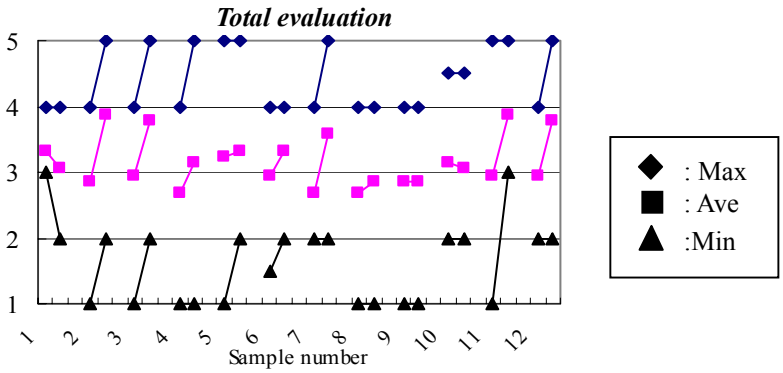


Figure 8. *Results of listening experiment*

6. CONCLUSIONS

This paper introduced a real-time software filtering algorithm which clarifies esophageal speech with the speaker's individuality preserved, and presented the results of the listening experiment to evaluate the filtering ability. The filtering results were preferably evaluated by the listeners in the points of the noise level and the clarification of vowels and consonants, as we had designed to complement the unclear factors and to enhance the clarity by considering the effectiveness of the comb filter. There still remain the problems such as the echo effects and electronic impressions caused by the mis-detection of the pitch. We continue to construct the precise filtering system to clarify the esophageal speech with the speakers individuality preserved, and examine the possibility to provide a portable device to generate the natural vocalization at any time and place.

We are currently working to expand the filtering algorithm to apply to the voice with the cerebral palsy. Since the voice has distinct characteristics which seem to be different from the esophageal speech, the different filtering techniques have to be examined. The software algorithm can be installed in a portable device to contribute to the use in mobile, and the study will support the verbal communication without being aware of the handicaps in speech.

Acknowledgements: This work was supported by the Grants-in-Aid for Scientific Research, the Japan Society for the Promotion of Science (No. 13780288). The authors would like to thank the members of Kagawa Kouyukai for their helpful supports for the experiment and the useful advice.

7. REFERENCES

T Sato, "Esophageal Speech and Rehabilitaion of the Laryngectomized", Kanehara & Co., Ltd., Tokyo, 1993

L. Max, W. Steurs, and W. De Bruyn, "Vocal capacities in esophageal and tracheoesophageal speakers", Laryngoscope, 106, 93-96, 1996

E. Noguchi and K. Matsui, "An evaluation of esophageal speech enhancement", *The Acoustical Society of Japan, Autumn Meeting* 2-6-13, pp. 421-422, 1996

T. Doi, S. Nakamura, J.L. Lu and K. Shikano, "Improvement in esophageal speech by replacing excitation components in cepstrum domain", *The Acoustical Society of Japan, Autumn Meeting* 2-4-17, pp. 253-254, 1996

J.L Lu, S. Nakamura and K. Shikano, "Study on Pitch Characteristics of Esophageal Speech", *The Acoustical Society of Japan, Spring Meeting* 2-7-19, pp. 253-254, 1997

M. Bellandese, J. Lerman, and J. Gilbert, "An Acoustic Analysis of Excellent Female Esophageal, Tracheoesophageal and Laryngeal Speakers", Journal of Speech, Language and Hearing Research, 44, pp. 1315-1320, 2001

J. Robbins, H. Fisher, E. Blom, and M. Singer, "A comparative acoustic study of normal, esophageal and tracheoesophageal speech production". Journal of Speech and Hearing Disorders, 49, pp. 202-210, 1984

C Roads, "The Computer Music Tutorial", The MIT Press, 1996.