Real-time clarification filter of a dysphonic speech and its evaluation by listening experiments

H Sawada[1], N Takeuchi[2] and A Hisada[3]

[1,2,3] 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

[1]www.eng.kagawa-u.ac.jp/~sawada

ABSTRACT

This paper presents a digital filtering algorithm which clarifies dysphonic speech with the speaker's individuality preserved. The study deals with the clarification of oesophageal speech and the speech of patients with cerebral palsy, and the filtering ability is being evaluated by listening experiments. Over 20,000 patients are currently suffered from laryngeal cancer in Japan, and the only treatment for the terminal symptoms requires the removal of the larynx including vocal cords. The authors are developing a clarification filtering algorithm of oesophageal speech, and the primal algorithm of software clarification and its effectiveness was reported in the previous ICDVRAT. Several algorithms for the clarification have been newly developed and implemented, and are being evaluated by questionnaires. The algorithms were extended and applied for the clarification of the speech by the patients of cerebral palsy.

1. INTRODUCTION

A voice is the most important and effective medium employed not only in the daily communication but also in logical discussions. Only humans are able to use words as means of verbal communication, although almost all animals have voices. Vocal sounds are generated 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 particular voice. If any part of the vocal organs is injured or disabled, we may be involved in the impediment in the vocalization and, in the worst case, 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. The removal of vocal cords means the loss of the voice, and causes various difficulties in the communication with other people, since the employment of a voice is essentially important for humans to make verbal communications.

There are mainly two ways to recover voice. One is to use an artificial larynx, which is a hand-held device with a pulse generator that produces a vocal cord-like vibration. An 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 he does for the usual vocalization. The device 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.

The other way is to train oesophageal speech, which is a method of speech production using an oesophagus (Sato, 1993; Max *et al*, 1996). In the speech, air is inhaled and caught in the upper oesophagus instead of being swallowed, and then the released air generates the oesophagus vibration to produce a "belch-like" sound that can be shaped into speech. A patient has difficulties to master the oesophageal speech (several years are ordinary required for the practice), however the voice is able to keep the speaker's individuality since the speech is generated by his own vocal organs, although several distinctive characteristics exist. Moreover the speaker is able to employ his body parts such as hands and facial expressions freely and actively for the communication to assist the speech.

Cerebral palsy (CP) is a condition caused by an injury to the parts of the brain which controls the ability to use muscles and bodies. The injury may happen before birth, sometimes during delivery, or soon after

©2004 ICDVRAT/University of Reading, UK; ISBN 07 049 11 44 2

being born. Severe CP may affect plural parts of patient's physical abilities, and requires to use a wheelchair and other special equipments to move. However, most of the patients with CP doesn't have mental retardation or disorder. CP doesn't get worse over time, and most children with CP have a normal life span. Patients of CP often have difficulties in moving and controlling vocal apparatus due to the insufficiency of the muscle controls, and the clarity of vocalized sounds is low.

The authors are developing a clarification filtering algorithm of oesophageal speech, and the primal algorithm of software clarification and its effectiveness was reported in Hisada and Sawada (2002). Several algorithms for the clarification have been newly developed and implemented, which were evaluated by questionnaires. The algorithms were extended and applied for the clarification of the speech by the CP patients. The paper presents the recent developments of the clarification filtering for the dysphonic speech, together with the implementation to a PC to be tested by a patient for the practical use.

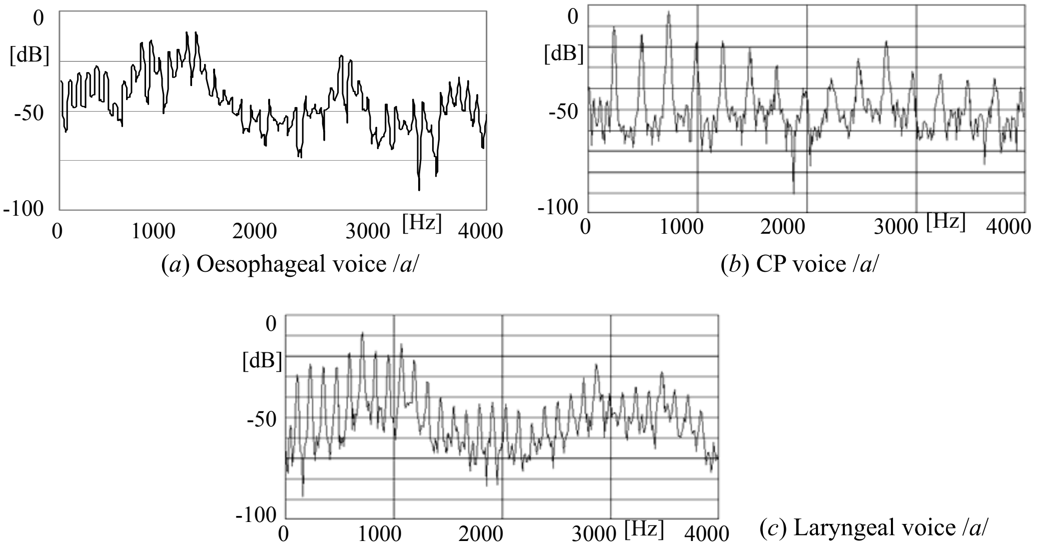
2. BACKGROUND OF STUDIES CONCERNING DYSPHONIC SPEECH

Several researches which analyze the characteristics of the oesophageal speech have been reported so far (Noguchi and Matsui, 1996; Doi *et al*, 1996), and a device to improve the oesophageal 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 speaker's individuality in the voices. Acoustic characteristics of the oesophageal 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 oesophageal voices, and the clarification theory and algorithm are not yet established.

Patients of CP, on the other hand, often have difficulties in moving and controlling vocal apparatus, and the clarity of vocalized sounds is low. To assist the verbal communication, many support tools such as *Talking-aid* and *Kinex* have been commercially available (Koroko, 2003). However, no studies to clarify patient's voice in real-time are not found so far.

3. PURPOSE OF THE STUDY

The purpose of the study is to develop clarification filtering algorithms of oesophageal speech and the speech of CP, for presenting a hand-held device to clarify the speech in real-time with speakers' individuality preserved. Spectra of an oesophageal voice and a CP voice are shown in Figure 1, together with the comparison of a laryngeal vocalization.



(*a*) Oesophageal voice /*a*/

(*b*) CP voice /*a*/

(*c*) Laryngeal voice /*a*/

Figure 1. *Comparison of spectra of three speeches*

Typical oesophageal speech is characterized as follows.

ES1) The voice contains noises caused by the irregular vibrations of oesophagus.

ES2) Fundamental frequency is rather low.
ES3) Fundamental frequency and its overtone structure are not clear.
ES4) Fluctuation exists, which causes the instability of the voice.
ES5) Spectrum envelope is flatter than the laryngeal voice.
ES6) Volume is low because of the shortage of the expiration. (Average is 150 ml; laryngeal voice is 2000ml)
ES7) Mastering the oesophageal speech requires repetitious training, and the progress is slow and is hard to be recognized by the patient in training.

The features of a CP voice, on the other hand, are listed below:

CP1) Spectrum envelope is flatter than laryngeal voice.
CP2) Formant frequency is unstable due to the fluctuation.
CP3) Starting of utterance is indistinct and the consonants are vague, due to the shortage of the expiration.
CP4) Fundamental frequency is rather high, and its overtone structure is indistinct in the higher frequency range.
CP5) Distinctive resonance characteristic is found around 2 kHz.
CP6) Spectrum power around 3 kHz decreases.

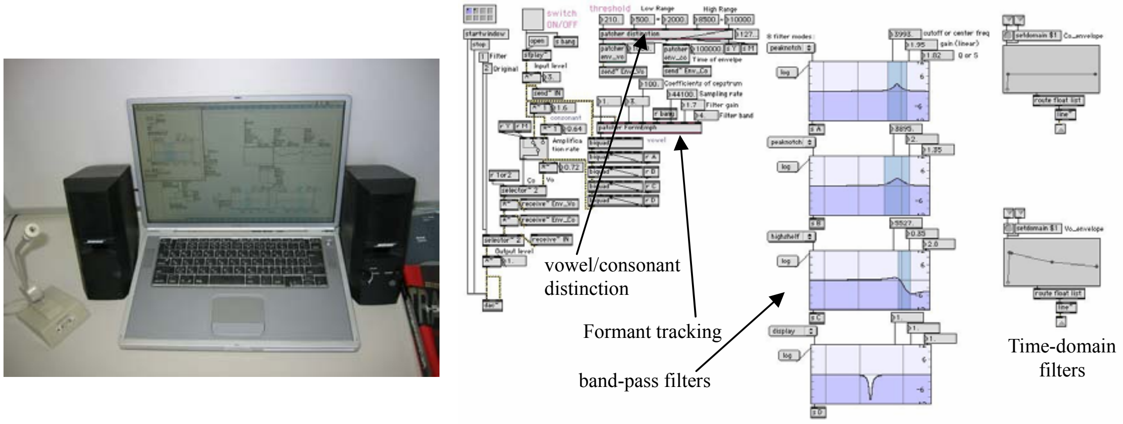
In this study, a real-time software filter for the clarification of the oesophageal speech and cerebral palsy speech has been constructed by taking account of the characteristics listed above. The implemented algorithms were tested by inputting actual dysphonic speeches of oesophagus vocalization and CP patients, and the filtering ability was evaluated by an listening experiment.

4. FILTERING ALGORITHMS

4.1 System Configuration

The algorithm was realized with the Max/MSP programming environment (Roads, 1996) in Macintosh Power Book (Power PC G4/667 MHz), equipped with a microphone and a pair of stereo speakers as shown in Figure 2(*a*). Sampling frequency was set to 44.1 kHz with 16 bits input, and the real-time processing is conducted in the programs. The software filter constructed in the Max/MSP is also shown in Figure 2(*b*).

Figure 3 shows the flow diagram of the filtering. First the input signal is judged to distinguish vowels and consonants, and is divided into two procedures. For a vowel sound, fundamental frequency is extracted to form a comb filter as to enhance pitch components, and at the same time two peaks are extracted from the spectrum envelope to be enhanced. For a CP voice, due to the features CP5) and 6), only the extraction of the first formant is executed to enhance the clarity of the vowel sounds without changing the speaker's individuality. 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 real-time for every input to generate clarified speech outputs.



(*a*) System configuration (*b*) Clarification filter in Max/MSP environment

Figure 2. *Flow diagram of filtering.*

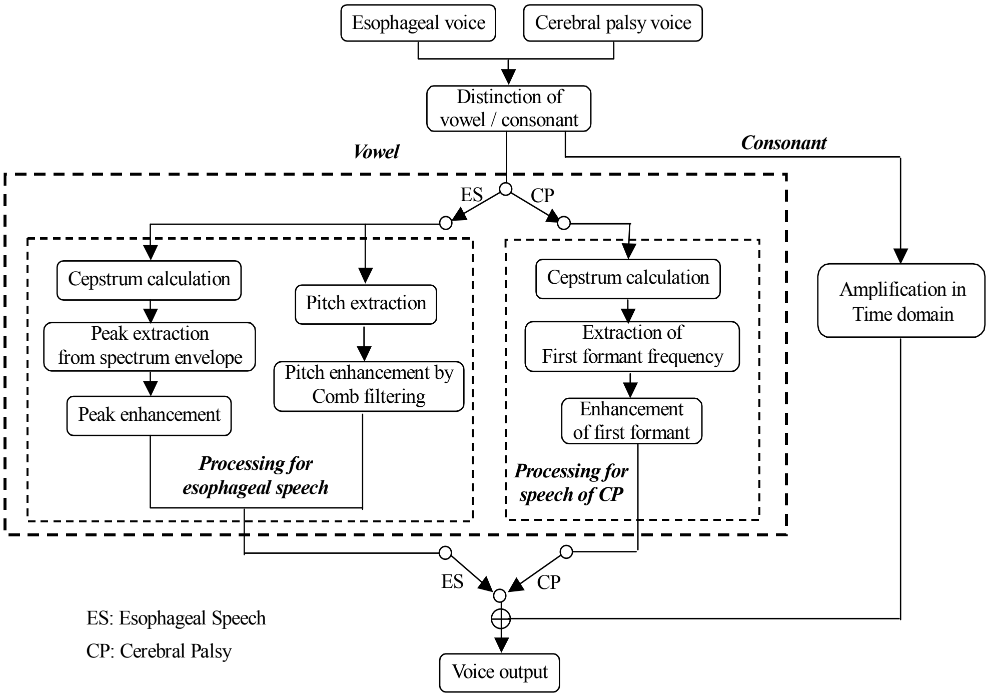

Esophageal voice
Cerebral palsy voice
Distinction of vowel / consonant
Vowel
Consonant
ES
CP
Cepstrum calculation
Pitch extraction
Peak extraction from spectrum envelope
Pitch enhancement by Comb filtering
Peak enhancement
Processing for esophageal speech
Cepstrum calculation
Extraction of First formant frequency
Enhancement of first formant
Processing for speech of CP
Amplification in Time domain
ES
CP
ES: Esophageal Speech
CP: Cerebral Palsy
Voice output

Figure 3. *Flow diagram of filtering.*

4.2 Distinction between vowels and consonants

Comb filter is able to enhance pitch components, and works to clarify vowel sounds in the oesophageal speech. Since the consonant sound basically doesn't have the overtone structure, the comb filtering performs inadequately. Because of the different characteristics of phonemes to be clarified, vowels and consonants are required to be classified in advance to employ 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.

The spectrum envelope of the vowel sound shows several local 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 particular 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 B_{index} as

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

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.

4.3 Comb filtering

A digital comb filter shown in Figure 4 enhances a fundamental frequency and its harmony components of input sounds. By suitably operating comb filters with the control of the parameters M, $g1$ and $g2$, oesophageal 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 preserved, since the comb filter enhances the overtones without changing the formants and the spectrum envelope.

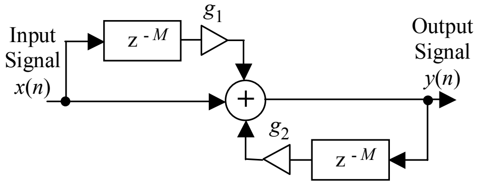
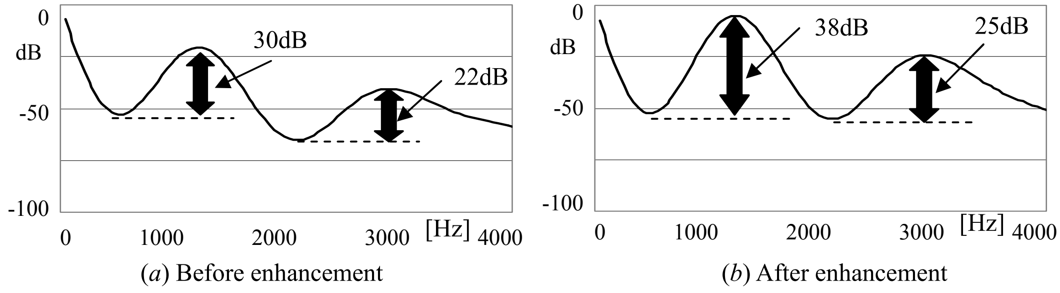


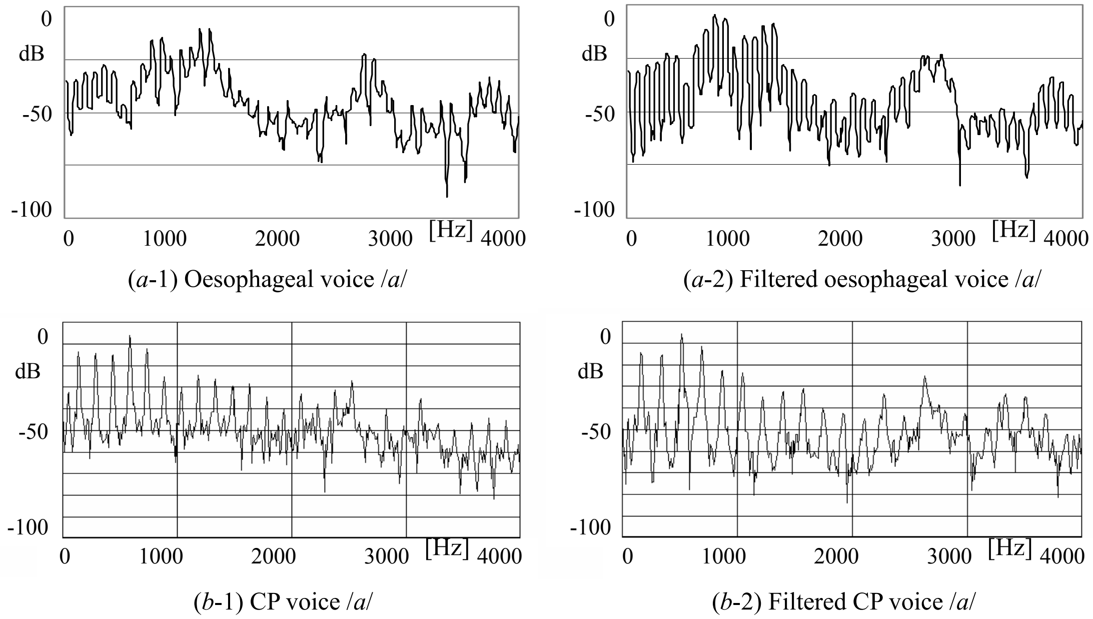
Figure 4. *Digital comb filter*



(*a*) Before enhancement

(*b*) After enhancement

Figure 5. *Spectrum envelope of /a/ sound before and after the enhancement*



(*a*-1) Oesophageal voice /*a*/

(*a*-2) Filtered oesophageal voice /*a*/

(*b*-1) CP voice /*a*/

(*b*-2) Filtered CP voice /*a*/

Figure 6. *Results of clarification filtering for CP voice /a/.*

4.4 Enhancement of peaks in spectrum envelope

A spectrum envelope characterizes the resonance characteristics, which are perceived as different vowel sounds. Vowel sounds are formed and articulated according to the inner shape of the vocal tract, which is governed by the complex movements of the jaw, the tongue and the muscles. Due to the insufficient articulation, the spectrum envelope of a dysphonic voice is flat comparing with a laryngeal voice. In this study, we tried to clarify vowel sounds by enhancing the spectrum peaks and modifying the resonance characteristics.

Spectrum envelope is obtained by the inverse Fourier transform of the 1st to 32nd coefficients extracted by the cepstrum calculation. Formants are the local maxima in the spectrum envelope, and the first and the second largest maxima in the lower frequency range are extracted every 30 ms. Then two band-pass filters to enhance the two peaks are generated in real-time, and are applied to emphasize the envelope. Since the higher-order formant frequencies are unstable due to the instability of vocalization, the peak emphasis filtering is applied only to the first and second formants. Figure 5 shows the spectrum envelope of /*a*/ sound before and after the enhancement.

4.5 Consonant enhancement

The volume of the oesophageal speech is generally low because of the lack of the expiration pressure, and the consonant pronunciations are insufficient and difficult to be distinguished. The starting utterance and consonants of CP speech are also indistinct as listed in CP3). The consonant sound is enhanced in the time domain by multiplying its amplitude by a suitable coefficient. The amplitude can be varied from 1.0 to 3.0 in the program not to distort the vocal characteristics, and a suitable value is determined according to the characteristics of input voice based on listening experiments conducted before the use.

4.6 Filtering experiments

Filtering experiments for oesophageal speech and CP voices were conducted. Figure 6 (*a*) and (*b*) show the power spectra of oesophageal speech /*a*/ and CP voice /*a*/ before and after the filtering, respectively.

The proposed filter could form the clear overtone structures of the fundamental frequencies for oesophageal speeches. The basic form of a spectrum envelope and formant positions were well preserved throughout the frequency range, and the peaks and valleys of the overtones were enhanced, which contributed to the clarification of vowel sounds and the noise reduction. The starting of utterance and consonants were emphasized and clarified, owing to the enhancement of the signal in the time domain. The individuality of the speaker was satisfactorily preserved, and the improvement of the clarification was also recognized.

5. LISTENING EXPERIMENTS

For the assessment of the filtering ability, listening experiments were conducted and filtered voices were evaluated by questionnaires.

10 able-bodied subjects listened to 12 pairs of unfiltered and filtered oesophageal speeches given by 6 speakers, and evaluated them in 7 levels from 12 points of view. The contents of the voices are

Voices A_1 - F_1: Speech of Japanese 5 vowels /aiueo/ by 5 speakers A to F.
Voices A_2 - F_2: Reading of Japanese sentence /Chiisana Otokonoko ga Teeburuno ueni/ (A little boy is sitting on a table) by 5 speakers A to F.

To examine the filtering ability to the difference of training experience, the speakers A to F were selected from the experience of a half, 3, 4, 7, 17 and 18 years, respectively. Only the speaker E was a female.

For CP voices, the subjects listened to 6 pairs of unfiltered and filtered speeches given by 3 patients, and scored them in 7 levels. The contents of the voices are

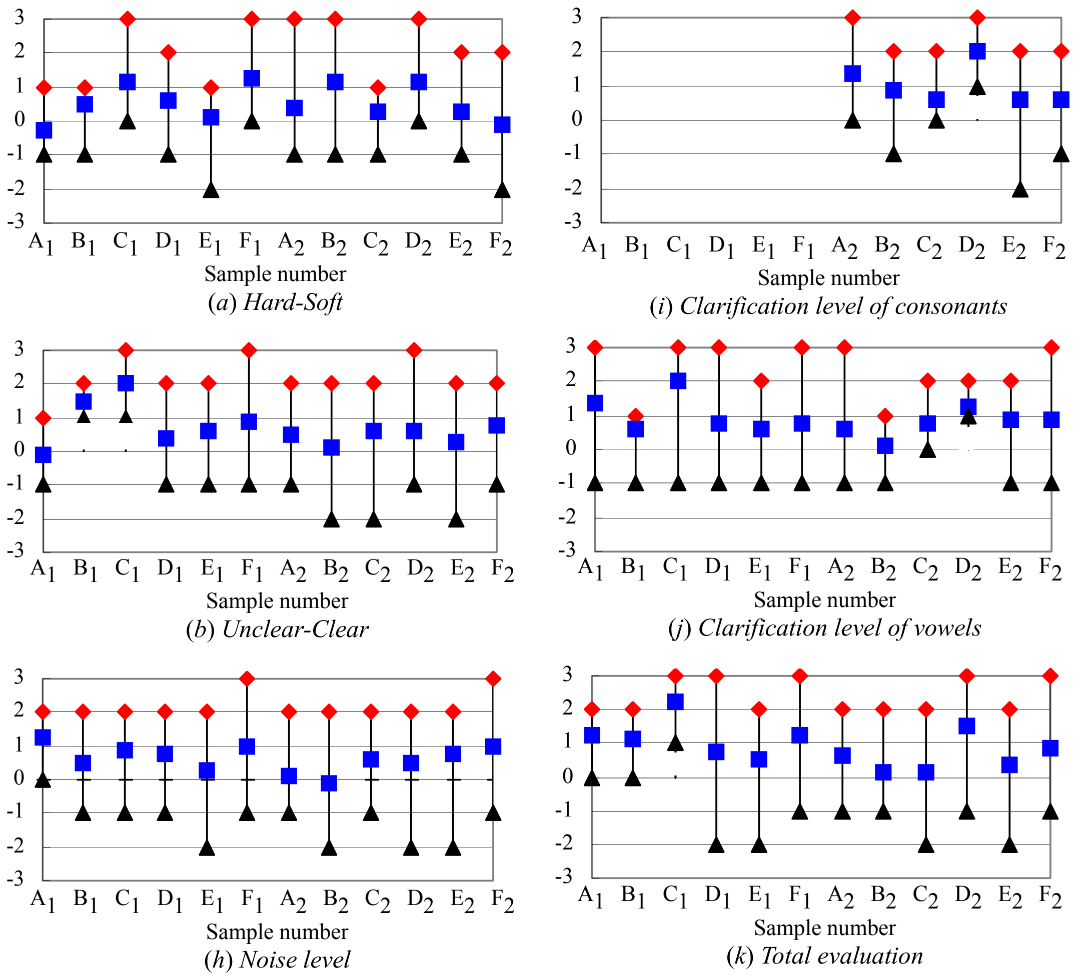
Voices G_1 - I_1: Speech of Japanese 5 vowels /aiueo/ by 3 speakers G to I
Voices G_2 - I_2: Reading of Japanese sentence /Chiisana Otokonoko ga Teeburuno ueni/ (A little boy is sitting on a table) by 3 speakers G to I

Twelve evaluation points are

(a) Hard-Soft, (b) Unclear-Clear, (c) Rough-Smooth, (d) Blur-Sharp,
(e) Unpleasant-Pleasant, (f) Echo level, (g) Electronic-Human, (h) Noise level,
(i) Clarification level of consonants, (j) Clarification level of vowels,
(k) Total evaluation, (l) Preservation of individuality.

A method of pair comparisons was employed for the evaluation. First, a subject listened to a pair of unfiltered and filtered speech, and was let know which is the filtered voice. After listening, he evaluated the filtered voice by comparing with the unfiltered speech, based on the twelve evaluation points listed above. In the adjective pairs, the former was assigned as negative impression, and the latter was rated as positive impression. A subject evaluated the speech with scores between -3 and +3. The greater the points are, the higher the filtering results were better evaluated.

Figure 7 and 8 show a part of results of the listening experiments for the clarification of oesophageal speech and CP speech, respectively. The maximum scores are plotted with the mark ◆, the average scores with ■, and minimum scores with ▲, summed up by the 10 subjects' evaluations. If the average plot is allocated upper than 0 level, the filtering is considered to be effective.



(*a*) *Hard-Soft*

(*i*) *Clarification level of consonants*

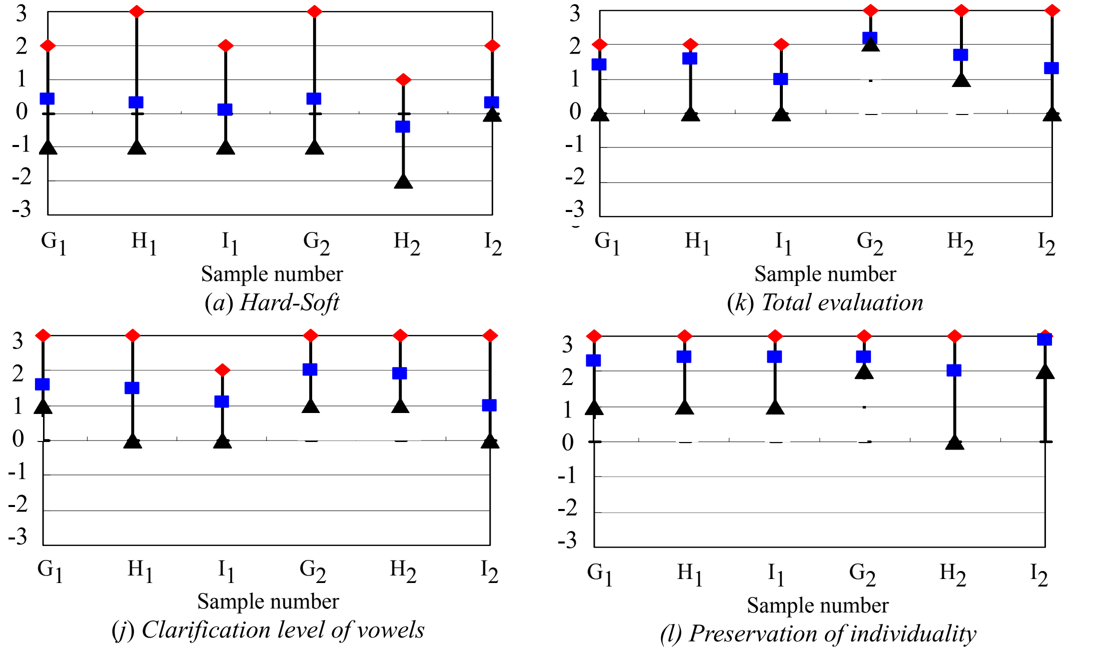
(*b*) *Unclear-Clear*

(*j*) *Clarification level of vowels*

(*h*) *Noise level*

(*k*) *Total evaluation*

Figure 7. Results of listening experiment for oesophageal speech



(*a*) *Hard-Soft*

(*k*) *Total evaluation*

(*j*) *Clarification level of vowels*

(*l*) *Preservation of individuality*

Figure 8. Results of listening experiment for the speech of CP

In almost all view points, filtered voices obtained higher evaluations, especially in the points of smoothness and clearness. For oesophageal speech, the filtering was especially effective for the speakers with the training experience of 3 to 4 years. In the evaluations of ***clarification level of consonants***, ***clarification level of vowels*** and ***total evaluation***, better assessments were obtained owing to the proposed filtering algorithms.

6. CONCLUSIONS

This paper introduced a real-time software filtering algorithm which clarifies oesophageal speech and the speech of cerebral palsy patients with the speaker's individuality preserved, and presented the results of the listening experiments for the assessment of the filtering ability. The filtering results were preferably evaluated by the listeners in the points of the smoothness, clearness and the clarification of vowels and consonants, as we had designed to complement the unclear factors and to enhance the clarity. There still remain the problems such as the echo effects and electronic impressions caused by the mis-detection of voice characteristics and mis-control of the filtering parameters in the real-time processing.

We continue to construct the precise filtering algorithm to clarify the dysphonic speech with the speakers individuality preserved. We are currently working to develop a portable device using a digital signal processor (DSP) to generate naturally clarified vocalization at any time and place. The proposed software algorithm will 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 Dr. Yoichi Nakatsuka, the director of the Kagawa Prefectural Rehabilitation centre for the Physically Handicapped, Dr. Rieko Goto in the department of medicine, Kagawa University, and the members of Kagawa Kouyukai for their helpful supports for the experiment and the useful advice.

7. REFERENCES

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

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

A Hisada and H Sawada, Real-time Clarification of Oesophageal Speech Using a Comb Filter *International Conference on Disability, Virtual Reality and Associated Technologies*, pp.39-46, 2002

E Noguchi and K Matsui, An evaluation of oesophageal 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 oesophageal 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 Oesophageal 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 Oesophageal, 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, oesophageal and tracheoesophageal speech production . Journal of Speech and Hearing Disorders, 49, pp. 202-210, 1984

Kokoro Resource Book 2003-2004 Kokoro Resource Book Publishing, http://www.kokoroweb.org/index.html, 2003.

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