PAPER

Enhancement of esophageal speech using formant synthesis

Kenji Matsui^{1,*}, Noriyo Hara¹, Noriko Kobayashi² and Hajime Hirose²

¹*Advanced Technology Research Laboratories, Matsushita Electric Ind. Co., Ltd.*

²*Kitasato University School of Allied Health Sciences*

(Received 1 March 2001, Accepted for publication 19 October 2001)

Abstract: The feasibility of using the formant analysis-synthesis approach to replace the voicing sources of esophageal speech was explored. Using inverse-filtered signals extracted from normal speakers provided the voicing sources. Pitch extraction was tested with various pitch extraction methods, and then a computationally simple, band-limited auto-correlation method was chosen. To accomplish stable and practical speech enhancement, the input signal was divided into low- and high-frequency channels, then only the low-frequency channel was processed by the formant analysis-synthesis method. A special purpose DSP-hardware unit was designed to perform the proposed analysis-synthesis process in real-time. Subjective evaluation tests (rating scale method) have been made with seven well-trained esophageal speakers and three speech therapists. Results of the subjective test showed that the synthesized speech was significantly improved, especially in cases of “loudness”, “sonority”, “strained”, “stoma noise”, “choppy”, “stability”, “intelligibility”, “recognizability”, and “duration” features.

Keywords: Speech enhancement, Esophageal speech, Formant synthesis, Digital signal processor, Voice source

PACS number: 43.72.Ew, 43.72.Ja, 43.70.Dn

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 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 level. 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. The formant patterns of esophageal speech, however, are found to be similar to those of normal speakers, except for overall elevations of formant frequencies attributed to the shortening of the vocal tract as a result of the surgical method.

A system that converts esophageal speech into normal speech could be useful for enhancing communication by

alaryngeal talkers, especially in noisy environments [1,2]. It is well documented that LPC analysis is an effective tool for estimating the parameters of the linear system for normal speech production. To enhance the quality of esophageal speech, Qi attempted replacing the voicing source of esophageal speech using a LPC method [2–4].

However, actual usability by alaryngeal talkers, subjective impressions of both talkers and listeners, and noise robustness have not yet been reported, and these factors are very important in determining the best design of a practical enhancement device.

The present study was undertaken to explore the feasibility of using the formant synthesis method to replace the voicing source of esophageal speech. Also, an enhancement device for esophageal speech running in real-time was designed and evaluated to determine the actual speech conversion performance in a real environment. The specific goals were: (1) to determine the practical algorithm and control parameters for analysis-synthesis of esophageal speech, and (2) to determine if the synthesized speech has high intelligibility; [5,7,9–17].

2. SPEECH DATA COLLECTION

As the first step, we collected esophageal speech to perform several basic experiments and evaluation. Eso-

*e-mail: matsui@crl.mei.co.jp

phageal speakers were recruited through contacts with the GINREI-KAI club in the Tokyo area. A total of 15-esophageal speakers, 11 men and 4 women, provided the speech data for this study. The speakers were rated by two speech therapists to have speech skill levels ranging from proficient to intermediate. The 15 speakers had various power levels, skills, and speech qualities.

Recordings were made of each subject producing 17 short sentences, shown in Table 1, at a comfortable level of pitch and loudness. The test sentences were designed by a speech therapist to be able to test different phoneme categories.

The recordings were made using a DAT recorder (TEAC RD-125T) in a quiet room with the microphone (B&K TYPE 4133) placed about 5 cm from the mouth of each talker. These recorded materials were used as the database for testing the various analysis methods and evaluating the effects of the speech enhancement methods.

3. SPEECH ANALYSIS-SYNTHESIS (I)

To enhance the speech quality of alaryngeal speech, Qi *et al.* tried to replace the voicing source using the LPC method [2,4]. Their basic assumptions are: 1) articulatory-based acoustic features of alaryngeal speech are not significantly modified by laryngectomy, and 2) vocal tract transfer functions of alaryngeal speech could be accurately determined using LPC analysis. Based on the same assumptions, we performed a simple experiment with a straightforward formant analysis-synthesis method using an inverse-filtered voicing source [8].

The formant synthesis method has played an important role in various areas of speech perception and speech disorders. This is because the synthesizer simulates the acoustic speech production process to a reasonable degree

of approximation and can smooth and modify the formant parameters. In some cases, both source- and articulation-related properties of esophageal speech need to be modified to achieve enhancement. Therefore, parametric synthesis methods such as formant synthesis must be utilized in order to modify the spectrum. Also, among the parametric synthesis methods, formant synthesis has the advantage of being able to smooth the spectrum, which could minimize discontinuities of the extracted formant parameters in the esophageal speech.

A block diagram of the analysis-synthesis process is illustrated in Fig. 1. The speech data was down sampled at 10kHz. Every 10ms, a 10th order LPC analysis was performed on 30ms Hanning windowed speech samples. Due to the unstable esophageal speech input, only the power level was used to discriminate voiced and unvoiced speech frames. Then, only voicing frames were analyzed and re-synthesized. The unvoiced speech frames were not processed and were accordingly mixed with the synthesis filter output at the final stage. The voiced signal was pre-emphasized and was used for processing.

A. Fundamental Frequency Contour Extraction

Esophageal speakers have less stability in F_0 control during phonation, and less variability during the intentional F_0 changes of natural speech. In order to figure out an effective way to extract the F_0 contour of esophageal speech, first, we performed a comparison study using several different pitch extraction methods [7]. The experimental results showed that there is not much difference among the methods, and it is difficult to obtain a stable

Table 1 Phonetically balanced short sentences for perceptual evaluations written in roman letters.

1	Papa mo mama mo minna de mamemaki o shita.
2	Sasayaku yô na asase no seseragi.
3	Karada ga daruku te daruku te shikata ga nai.
4	Takai takai tokoro e nobotte iku tokoro da.
5	Achirakara mo kochirakara mo dochirakara mo ikukoto ga dekiru.
6	Aoi o ueru.
7	Pitchâ ga dai ikkyû o nage mashita.
8	Anohito wa bunkajin to yobareru no ga fusawashî.
9	Ikko niko to kazoe te iki mashita.
10	Shichi gatsu kara hanshin densha de tûkin shite ima su.
11	Gasorin ga haitta bin ga ware te kiken da.
12	Ginkô mo gakkô mo aruite ikeru kyorî ni ari masu.
13	Byôsyô de mo gyûnyû o nonde mita.
14	Kinkô ga tore te iru no de kakkô ga yoi.
15	Shijû gamu o kamuno ga syûkan ni natte iru.
16	Hana o ottari ana o hottari sanzanna meni atta.
17	Chôdo hyakuban ni natta no wa zyôdeki da.

Fig. 1 A block diagram of the speech analysis-synthesis (Method I).

estimate of F_0 due to the inherent in-stability of the speakers' esophageal speech F_0 . Therefore, we chose a computationally simple autocorrelation method [5] with pre-selected upper and lower frequency limits of the F_0 range, so that the F_0 extraction process is suitable for the computationally limited speech enhancement devices, which should work in real-time. Three-point median smoothing followed the F_0 extraction process.

B. Formant Estimation

Our objective is to realize a speech enhancement device running in real time. Considering the processing speed, LPC analysis with Bairstow's method [6], a technique of root finding, was chosen for estimating the first three formant frequencies and bandwidths. Three-point median smoothing was also used to reduce the prediction error.

C. Voice Source Generation

Only voiced frames were analyzed and re-synthesized with the extracted formant parameters and pre-selected inverse-filtered voicing source [8]. The new excitation source was generated from one-pitch-period glottal waveform samples extracted from normal male and female talkers. The inverse filtered source waveform was obtained using a software package developed in our laboratory. First, a pitch-synchronized LP analysis was completed on a steady part of a vowel. An inverse filter was constructed from the LP coefficients. Then, the locations and the bandwidths of the inverse filters were hand-tuned by looking at the smoothness of the FFT spectrum. Aliasing distortion was considered as substitute for the higher formants effect. One-pitch-period waveform was Hamming windowed and used as the new excitation source with the pitch-synchronized overlap add method. The fundamental frequency can be increased or decreased from the preset condition.

D. Speech Synthesis

Fourth formant and fifth formant are usually set to constant values based on the analysis result of the users. A cascade formant synthesizer with five resonators was used to synthesize the voiced speech. The output from the synthesis filters was merged with the output from the unvoiced channels.

4. PERCEPTUAL EVALUATION (I)

A paired comparison test was conducted to assess if the synthesized speech was preferred over the original unprocessed speech. Six listeners participated in the evaluation task. The test material was a short sentence [aoi o ueru]. The speech from 15 speakers was used from the speech database. The order of the pairs in the list was randomized. The sound level was adjusted to 60 dBA at listener's position. Also, the same task was performed with 55 dBA pink noise. Testing the speech in noisy environ-

Fig. 2 A paired comparison test result of the speech analysis-synthesis (Method I) and the original speech. Six subjects judged 15 speakers' speech. The figures are average preference scores, standard deviations for six listeners (SD_L), and standard deviations for 15 speakers (SD_S).

ment is very important because one of the objectives of the speech enhancement is to improve the speech intelligibility in noise. Average scores, standard deviations for six listeners (SD_L), and standard deviations for 15 speakers (SD_S) are shown in Fig. 2. The preference score differences among listeners were not significant. On the other hand, significant score differences among speakers existed. In quiet environment, there was almost no improvement except for 5 of the speakers. However, in noisy environment, the synthesized speech was preferred for all except 3 speakers. The result of this simple study indicates that the replacement of the voice source could be effective to improve the vowel intelligibility in noisy environment. On the other hand, the synthesized speech has an un-natural, machine like sound in the quiet environment.

5. SPEECH ANALYSIS-SYNTHESIS (II)

The previous power threshold-based V/UV detection often failed the discrimination in the case of strong voiceless fricatives. To avoid the problem, we performed a V/UV detection experiment using several discriminate functions [9]. The result showed about 90% discrimination performance for a very stable, well-trained esophageal speaker. However, discrimination errors still occurred frequently and degraded the speech quality. To accomplish more stable and more natural-sounding voice conversion, we modified the previous analysis-synthesis method by adding a by-pass channel for the high frequency speech components, as shown in Fig. 3. This modification has two major effects. Blending the higher frequency component into the synthesized speech gives more natural sounding and intelligible consonants. Also, voiced/unvoiced decisions using the lower frequency component power information gives better results. For example, the confusion between unvoiced fricatives and voiced segments is often avoidable by utilizing the lower band power information. On the other hand, if the original fundamental frequency is

Fig. 3 A block diagram of the speech analysis-synthesis (Method II).

modified significantly, the synthesized speech sometimes sounds like mixed speech from two speakers. In this modified approach, the input speech signal was digitized by sampling at 10 kHz. The input signal was divided into two channels at 2.5 kHz with 4th order low pass and high pass filters. Only the low frequency channel was used for the synthesis processing. The high frequency channel was mixed with the synthesis filter output at the final stage. The lower channel power level was used to discriminate voiced and unvoiced speech frames. The unvoiced speech frames were not processed. Only voicing frames were analyzed and synthesized with a pre-selected inverse-filtered voicing source. The analyzed parameters (pitch, power, formant frequencies, bandwidths) were filtered by 3-point median smoothing. Fourth formant and fifth formant were usually set to constant values, which were obtained from the formant analysis beforehand.

6. PERCEPTUAL EVALUATION (II)

The same comparison test was conducted to determine whether enhancement of esophageal speech using the method II gave better results than method I. Each original sentence and its synthetic counterparts from method I and method II were presented to the listeners. The order of the sets was randomized. The listeners had to assess which sample was preferred. Six listeners participated in the evaluation task. Again, the test material was a short

sentence [aoi o ueru]. The speech from 15 speakers was used from the database. Average preference scores, standard deviations for six listeners (SD_L), and standard deviations for 15 speakers (SD_S) are shown in Fig. 4. As can be seen from the figure, the results were similar to those obtained by method I. The preference score differences among listeners were not significant. On the other hand, significant score differences among speakers existed. Both in quiet and noisy environments, method II was preferred over both method I and the original speech. Enhanced speech was more preferred in noisy environment than in quiet one. The result of this simple study indicates that both method I and method II could be effective to improve the vowel intelligibility in noisy environment. Method II was more preferred than method I. Based on this evaluation result, we utilized method II for later evaluation.

7. DSP IMPLEMENTATION

The improved speech enhancement algorithm (method II) described above was implemented on a DSP system based on TMS320C32. A block diagram of the hardware architecture is shown in Fig. 5. PC-based speech analysis

	Synthesis (I)	Synthesis (II)	Original Speech
Without Noise	32% ($SD_L=10\%$, $SD_S=27\%$)	40% ($SD_L=16\%$, $SD_S=30\%$)	27% ($SD_L=10\%$, $SD_S=16\%$)
With Noise	34% ($SD_L=8\%$, $SD_S=23\%$)	43% ($SD_L=3\%$, $SD_S=25\%$)	22% ($SD_L=10\%$, $SD_S=25\%$)

Fig. 4 A paired comparison test result of the speech analysis-synthesis (Methods I and II) and the original speech. Six subjects judged 15 speakers' speech. The figures are average preference scores, standard deviations for six listeners (SD_L), and standard deviations for 15 speakers (SD_S).

Fig. 5 A block diagram of the DSP hardware unit.

tools have also been developed to analyze all of the necessary pre-adjustment parameters, (i.e. average power, pitch range, V/UV power threshold) in advance. We introduced the DSP to realize all of the analysis-synthesis process in real time. The DSP hardware unit consists of a small board with a 60MHz C32, a 32 bit floating point

DSP, a fourteen bit A/D and D/A with 10kHz sampling, a parallel host interface, 16Mbit SRAM with 20ns access time, analog circuits for a microphone and loudspeaker, auto-voltage regulator, and a Li-ion battery. Figure 6 shows the DSP unit. With the combination of the PC-based speech analysis tools and the DSP hardware unit, a new user can first adjust all of the necessary parameters, and then download the adjusted DSP software into the hardware unit. Some overall signal processing delay is unavoidable. Because of the analysis frame length (30ms) and the 3-point median smoothing, the system requires 40ms delay. Figure 7 shows an example of the analysis result using the DSP hardware unit. The input speech is “aoi o ueru”. The input speech shows noisy and unstable pitch period. On the other hand, the processed speech shows stable pitch period. Due to the median smoothing, the F_0 and formant frequencies are well stabilized, as shown in the Fig. 7.

8. PERCEPTUAL EVALUATION (III)

Subjective evaluation tests (by rating scale method) have been made with 5 male and 2 female well-trained esophageal speakers, and 3 speech therapists. Each speaker first spoke “aoi o ueru” twice to adjust all of the necessary parameters. (Table 2) During the adjustment, fundamental frequencies of the female speakers were raised. After pre-

Fig. 6 The DSP hardware unit. The unit size is 65 mm × 65 mm.

Fig. 7 An example of the analysis tool view. From the top, (1) the input speech waveform “Aoi o ueru”, (2) the enhanced speech waveform, (3) Extracted pitch contour, (4) Extracted formants, F_1 , F_2 , and F_3 .

Table 2 Preset parameter values of each esophageal speaker, and *t*-test result.

Skill	Proficient					Intermediate	
Speaker	W	MI	S	A	H	N	MO
Sex	Male	Male	Male	Female	Female	Male	Male
LPC order	10	10	10	10	10	10	10
F_4	Fixed	Fixed	Fixed	Extracted	Extracted	Fixed	Fixed
F_0 detection range (Hz)	60 ~ 150	60 ~ 150	60 ~ 150	80 ~ 170	60 ~ 130	60 ~ 130	60 ~ 120
Power threshold (dB)	30	30	30	40	30	30	30
F_0 -shift	1	1	1	2	2	1	1
<i>t</i> -test result	⊙	○	⊙	NSD	⊙	⊙	⊙

⊙: The speech was improved ($p < 0.01$), ○: The speech was improved ($p < 0.05$).

setting the hardware unit, the speaker read the test materials without using the enhancement device first, and then spoke the same materials with the device. The test materials consist of seven sentences which contain samples of all phoneme classes. The sentences are 6, 8, 10, 12, 14, 15, 16 from Table 1.

The speech therapists rated the quality of the unprocessed speech stimuli and the processed ones using multidimensional scaling. Problems of various speech disorders are related to pitch, loudness, intonation, rhythm, resonance, and speech quality [18]. These features were selected from the reference to be able to assess the problems. Also we added some features which characterize a wide diversity of esophageal speech quality [19,20]. As shown in Table 2, a *t*-test of the subjective evaluation indicated that the synthesized speech obtained better scores except in the case of one female speaker who has almost normal speech quality. Figure 8 graphically shows the

average preference score differences between processed speech and unprocessed speech. The scores from each speech therapist, the averages, and standard deviations are summarized in Table 3. The preference score differences among speech therapists are not significant.

9. DISCUSSION

The results of this study indicate that the speech quality of esophageal speakers could be improved by LPC based formant analysis and synthesis processing. The ability to improve the voice quality with the proposed method implies that glottal source replacement may be adequate for esophageal vowel reproduction. The quality improvements were observed in eleven features out of nineteen, such as “loudness”, “sonority”, “strained”, “stoma noise”, “choppy”, “stability”, “intelligibility”, “recognizability”, and “duration” features. No significant improvement, however, was found for the “breathy” feature, although we

Fig. 8 Subjective evaluation result (The average scores from 3 speech therapists). **($p < 0.01$), *($p < 0.05$).

Intelligibility scale: 1: Clear, 2: Almost clear, 3: Clear if known phrase, 4: Almost unclear, 5: Very unclear. **Other features scale:** 1: Normal, 2: Slightly Annoying, 3: Annoying, 4: Very Annoying.

Table 3 Subjective evaluation result (The scores from 3 speech therapists, average scores and standard deviations).

	Unprocessed					Processed				
	Speech therapists					Speech therapists				
	NK	HH	NH	Average	(SD)	NK	HH	NH	Average	(SD)
Sonority**	2.43	1.57	2.14	2.05	0.67	2.00	1.00	2.00	1.67	0.48
Roughness*	3.00	1.86	2.29	2.38	0.59	2.71	1.57	2.14	2.14	0.65
Breathy	2.00	1.29	1.71	1.67	0.66	2.14	1.00	1.29	1.48	0.68
Strained**	2.43	1.29	2.29	2.00	0.77	1.43	1.00	1.14	1.19	0.40
Naturalness	2.71	1.43	2.14	2.10	0.70	2.29	1.43	1.86	1.86	0.57
Stoma noise**	2.14	2.00	2.14	2.10	0.94	1.57	1.14	1.29	1.33	0.48
Injection noise	1.43	1.29	2.00	1.57	0.81	1.57	1.43	1.71	1.57	0.51
Duration**	2.00	1.29	2.29	1.86	0.73	1.71	1.00	1.71	1.48	0.60
Too low F_0 *	2.00	2.29	1.86	2.05	0.97	1.57	1.86	1.57	1.67	0.48
Too high F_0 *	1.00	1.00	1.14	1.05	0.22	1.00	1.00	1.00	1.00	0.00
F_0 perturbation	1.00	1.29	1.71	1.33	0.58	1.29	1.71	1.43	1.48	0.51
Loudness**	2.43	1.43	2.43	2.10	0.83	1.00	1.00	1.14	1.05	0.22
Loudness perturbation	1.86	1.00	2.00	1.62	0.74	1.57	1.00	1.29	1.29	0.46
Choppy**	1.57	1.43	2.57	1.86	0.79	1.29	1.14	1.71	1.38	0.50
Consonant intelligibility	2.29	2.00	2.43	2.24	0.54	2.00	1.86	2.14	2.00	0.32
Vowel intelligibility	1.00	1.00	1.57	1.19	0.40	1.00	1.00	1.43	1.14	0.36
Stability**	2.86	2.00	2.57	2.48	0.51	2.29	1.86	2.00	2.05	0.38
Intelligibility (overall)**	2.00	1.43	2.71	2.05	0.80	1.86	1.14	2.14	1.71	0.56
Recognizability**	2.71	1.57	2.43	2.24	0.70	2.00	1.43	2.00	1.81	0.40

**($p < 0.01$), *($p < 0.05$).

had expected to observe some improvements. This was because one of the female speakers gave a “breathy” impression when the fundamental frequency was increased. On the other hand, “choppy”, “strained” and “duration” features, which are not directly related to the signal processing effect, obtained high scores. This is because just amplifying the input speech may relieve the speaker’s tension. Stoma noise was significantly reduced because the close-talking microphone characteristics effectively reduce ambient noise. As expected, the proposed method did not produce the same perceptual effects across all talkers. For example, in cases with less low-frequency power, rough and noisy talkers, the enhanced speech was clearly preferred, most likely because the proposed method boosts the lower frequency range. On the other hand, for talkers who have relatively rich vowel power, their original speech was preferred over the enhanced one.

Every speaker was asked to fill out a form to state his or her impressions. Significant comments included:

- (1) The device may require some training period to feel comfortable. For example, adapting to the 40 ms delay.
- (2) Increasing the original fundamental frequency gives better sound impression.
- (3) The processed speech sounds like some other person’s voice.
- (4) Speaking in a relaxed voice gives better output voice quality.

Also, the speech therapists were asked to summarize

their impressions.

- (1) Loudness and overall speech quality are improved.
- (2) The device is effective especially for hoarse voice.
- (3) Sometime the injection noise is amplified.
- (4) The input and output mixed voice is heard due to the time delay.
- (5) In the case of female speech, increasing the original pitch gives better sound.
- (6) Using the device can be effective for the esophageal speech training.

10. CONCLUSION

LPC based formant analysis synthesis method for esophageal speech was proposed, and a real-time DSP unit was designed. The device was tested in a practical environment by speech therapists. Results of the subjective evaluation indicated that the processed speech was preferred in most of the evaluation criteria. However, the esophageal speakers may be required to do some training before utilizing the device comfortably. The device also needs further evaluation from both listeners and speakers under various practical conditions.

ACKNOWLEDGMENTS

The authors would like to express our gratitude to the members of GINREIKAI for their cooperation with the recording and subjective evaluation. We also would like to thank Mr. K. Ohira, Ashida Sound Co., Ltd. Mr. Kubota and Ms. M. Jin, TEAC Corporation for collaborating on

this work with us.

This Work was performed as a part of the Research & Development of Medical & Welfare Apparatus supported by NEDO (New Energy and Industrial Technology of Development Organization) of Japan.

REFERENCES

- [1] H. Javkin, M. Galler and N. Niedielski, "Enhancement of esophageal speech by injection noise rejection", *Proc. IEEE ICASSP*, **97**, 1207 (1997).
- [2] N. Bi and Y. Qi, "Application of speech conversion to alaryngeal speech enhancement", *IEEE Trans. Speech Audio Proc.*, **5**, 97 (1997).
- [3] Y. Qi and B. Weinberg, "Characteristics of voicing source waveforms produced by esophageal and tracheoesophageal speakers", *J. Speech Hear. Res.*, **38**, 536 (1995).
- [4] Y. Qi, "Replacing traceoesophageal voicing source using LPC synthesis", *J. Acoust. Soc. Am.*, **88**, 1228 (1990).
- [5] K. Matsui and N. Hara, "Development and evaluation of speech aid system for esophageal speakers", *Proc. Autumn Meet. Acoust. Soc. Jpn.*, p. 333 (1998).
- [6] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical Recipes in C* (Cambridge University Press, Cambridge, 1988), p. 283.
- [7] N. Hara and K. Matsui, "Pitch detection performance for esophageal speech and amplitude to pitch-contour conversion method", *Proc. Autumn Meet. Acoust. Soc. Jpn.*, p. 331 (1998).
- [8] S. Pearson, H. Javkin, K. Matsui and T. Kamai, "Text-to-speech synthesis using a natural voice source", *ICSLP*, **90**, 193 (1990).
- [9] N. Hara and K. Matsui, "V/UV/N detection by using discriminant function for esophageal speech", *Proc. Spring Meet. Acoust. Soc. Jpn.*, p. 335 (1998).
- [10] N. Hara and K. Matsui, "Adaptive gain control method for esophageal speech", *Proc. Autumn Meet. Acoust. Soc. Jpn.*, p. 329 (1997).
- [11] K. Matsui, N. Hara and H. Javkin, "Real-time formant analysis-synthesis method for esophageal speakers", *Proc. Autumn Meet. Acoust. Soc. Jpn.*, p. 331 (1997).
- [12] K. Matsui, E. Noguchi and Y. Kato, "Enhancement of esophageal speech using formant synthesis method", *Proc. Spring Meet. Acoust. Soc. Jpn.*, p. 311 (1997).
- [13] K. Matsui, E. Noguchi and H. Javkin, "Enhancement of esophageal speech", *Proc. Autumn Meet. Acoust. Soc. Jpn.*, p. 423 (1997).
- [14] E. Noguchi and K. Matsui, "An evaluation of esophageal speech enhancement", *Proc. Autumn Meet. Acoust. Soc. Jpn.*, p. 421 (1997).
- [15] N. Hara, K. Matsui, K. Kubota, M. Jin and I. Ohira, "Total performance of the esophageal speech aid", *Proc. Spring Meet. Acoust. Soc. Jpn.*, p. 183 (1999).
- [16] N. Hara, K. Matsui, K. Kubota, M. Jin and I. Ohira, "The esophageal speech aid system", *Tech. Rep. IEICE*, **SP98-150**, 33 (1999).
- [17] K. Matsui and N. Hara, "Enhancement of esophageal speech using formant synthesis", *Proc. ICASSP*, **99**, 81 (1999).
- [18] The Japan Society of Logopedics and Phoniatrics, *Voice Examination* (Ishiyaku Publishers Inc., Tokyo, 1979), p. 191.
- [19] J. K. Casper and R. H. Colton, *Clinical Manual for Laryngectomy and Head/Neck Cancer Rehabilitation* (Singular Publishing Group Inc., San Diego, 1998), p. 157.
- [20] P. C. Doyle, *Foundations of Voice and Speech Rehabilitation Following Laryngeal Cancer* (Singular Publishing Group Inc., San Diego, 1994), p. 143.

Kenji Matsui received his B.E., M.E. degrees in Communication Engineering from Osaka University in 1979 and 1981. He joined Matsushita Electric Ind. Co., Ltd. in 1981. From 1986 to 1988, he was a research engineer at Panasonic Speech Technology Laboratory, USA. Currently, he is a general manager at Advanced Technology Research Labs. He is a member of ASJ.

Noriyo Hara received a B.E degree in Mathematics from Nara Women's University in 1979. She joined Matsushita Electric Ind. Co., Ltd. in 1979. Currently, she is a staff researcher at Advanced Technology Research Labs. She is a member of ASJ, IEICE and PSJ.

Noriko Kobayashi is a professor of School of Allied Health Sciences, Kitasato University. She also works for patients with voice and speech disorders in Kitasato University Hospital as a licensed speech therapist. She graduated from Tsuda College with a B.A. degree in Liberal Arts in 1970 and from Graduate School and University Center of the City University of New York with a PH.D. in speech sciences in 1988.

Her interests are physiological mechanisms of speech production and speech pathology (speech disorders).

Hajime Hirose received his M.D. degree from Faculty of Medicine, University of Tokyo in 1957 and Dr. Med. Sciences degree from Section of Otolaryngology, Graduate School, University of Tokyo in 1962. From 1970 to 1973, he was a research associate at Haskins Laboratories, New Haven, USA. From 1983 to 1993, he was a professor of Research Institute of Logopedics and Phoniatrics, Faculty of Medicine, University of Tokyo. He is now a professor emeritus, University of Tokyo, and a professor in School of Allied Health Sciences, Kitasato University. He received the Manuel Garcia Prize (International Association of Logopedics and Phoniatrics) in 1983. He is a member of IALP.