Evaluation of Extremely Small Sound Source Signals Used in Speaking-Aid System with Statistical Voice Conversion

Keigo NAKAMURA(a), Student Member, Tomoki TODA(b), Hiroshi SARUWATARI(c), Members, and Kiyohiro SHIKANO(d), Fellow

SUMMARY We have so far proposed a speaking-aid system for laryngectomees using a statistical voice conversion technique. In the proposed system, artificial speech articulated with extremely small sound source signals is detected with a Non-Audible Murmur (NAM) microphone, and then, the detected artificial speech is converted into more natural voice in a probabilistic manner. Although this system basically allows laryngectomees to speak while keeping the external source signals silent, it is still questionable how much these new sound source signals affect the converted speech quality. In this paper, we investigate the impact of various sound source signals on voice conversion accuracy. Various small sound source signals are designed by changing the spectral envelope and the waveform power independently. We conduct objective and subjective evaluations. The results of these experimental evaluations demonstrate that voice conversion accepts 1) various sound source signals with different spectral envelopes and 2) large degree of power of the sound source signals unless the power of speaking parts is almost equal to that of silence parts. Moreover, we also investigate the effectiveness of enhancing auditory feedback during speaking with the extremely small source signals.

key words: laryngectomy, speaking aid, electrolarynx, voice conversion, NAM, enhancing auditory feedback

1. Introduction

Speech is one of our most basic communication tools in our daily life, and therefore, speech-impaired people who have difficulties with phonation have serious problems in their speech communication. Laryngectomees are a kind of speech-impaired people whose vocal folds were removed as a result of laryngectomy. In particular, because total laryngectomees completely lose their vocal folds, they have to find an alternative method to utter without vocal folds vibration.

An electrolarynx is a battery-driven medical device that imitates our vocal folds vibration electrically [1]. Some electrolarynxes such as 'yourtone' and 'myvoice', which are developed in Japan, are designed so that the user can speak more naturally with it [2]-[5]. As Fig. 1 shows, these electrolarynxes basically have same structures. 'yourtone' enables us to speak with a very natural voice by controlling the pitch frequency [3], [4]. 'myvoice' allows the user to speak with pre-defined built-in $F_0$ contours [5]. Although the user can easily master the usage of these electrolarynxes, some disadvantages have been indicated so far. Because it is indeed difficult to generate natural sound source signals in particular exhibiting a natural $F_0$ contour, the articulated speech using an external sound source device that is called artificial speech in this paper sounds mechanical. Moreover, because these devices are developed so that the users can speak with the volume of normal speech in their daily life, they need to generate large power of sound source signals. Those sound source signals would be noisy for people around the user especially in a quiet environment such as a library.

Our study focuses on two problems of the conventional electrolarynxes: 1) noisy sound source signals and 2) unnaturalness of the artificial speech. In order to address these problems, we have proposed a novel speaking-aid system for laryngectomees using some innovative devices and techniques [6]. In our proposed system, first, the user is supposed to utter using extremely small sound source signals. The artificial speech is captured by a body-conductive microphone. The detected body-conducted artificial speech is converted into natural speech by a statistical voice conversion technique. Finally, the converted speech is presented as the user's voice. An advantage of our system is that the people around the user would hear neither noisy sound source signals nor the produced artificial speech. This system is supposed to be used in speech communications of the user's daily life.

Because the proposed system includes the voice conversion part, the sound source signal can be designed in different view points compared to the conventional electrolarynxes. In order to keep the sound source signals from annoying people around the user, it is desired to make its power as small as possible or to set its frequency components so as to be easily masked by background noises. However, it is still questionable how much the voice conversion accuracy is affected by using different sound sources. Therefore, we investigate the impact of various sound source signals on voice conversion accuracy. To demonstrate the acceptable ranges of the sound source signals in our proposed system, we evaluate voice conversion accuracy when using some source signals designed by changing the spectral envelope and the total power independently.

Copyright © 2010 The Institute of Electronics, Information and Communication Engineers
One concern in our proposed system is also unrevealed. Our proposed system employs statistical voice conversion that constructs of training and conversion part. To use our system, a conversion model needs to be trained in advance using the recorded training data. The user would be able to get the auditory feedback from the output converted speech. However, the user cannot get it during the speech recording of the training data for the voice conversion. Therefore, the user has to get the auditory feedback from only the produced artificial speech in the training data recording. Because the power of the sound source signal used in our system is extremely small, it might be difficult for the speaker to get sufficient auditory feedback. This would make the articulation of the speaker unstable and the voice conversion accuracy degraded. In this paper, our system directly provides the user with artificial speech enhanced by an amplifier as the amplified auditory feedback, and the effectiveness of enhancing the artificial speech is experimentally evaluated.

This paper consists of following parts. The Sect. 2 describes details about our aid system. The motivation of enhancing the auditory feedback is described in Sect. 3. Some sound source signals are designed in Sect. 4. Effectiveness of enhancing the auditory feedback and the acceptance ranges of sound source signals are experimentally evaluated in Sect. 5. Finally, this paper is concluded in Sect. 6.

2. Proposed Speaking-Aid System

We have proposed a novel speaking-aid system for laryngectomees as shown in Fig. 2[6]. Our system consists of the following four components: 1) extremely small sound source signals [7], 2) Non-Audible Murmur (NAM) microphone [8], 3) voice conversion [9], and 4) output.

The user articulates using sound source signals that are generated from the unit attached on the lower jaw and are conducted through the skins toward the oral cavity. Note that the attaching location of the sound source unit is same as conventional electrolarynxes. The power of the sound source signal is extremely small so that people around the user can hardly hear both the sound source signal itself and the artificial speech. Therefore, the use of the extremely small sound source signals is effective to alleviate the problem of noisy sound source signals. However, it causes another problem, i.e., the difficulty of detecting the artificial speech because its power is too small.

A novel body-conductive microphone called NAM microphone [8] is introduced as the next step to capture the soft artificial speech. NAM is defined as articulated respiratory sound without vocal-fold vibration through the soft tissues of the head [10], and the NAM microphone has been developed as a special sensor to detect NAM. Because the low dynamic range of the NAM microphone is much larger than that of the conventional air-conductive microphones, it enables us to detect artificial speech as a kind of body-conducted speech even if using the extremely small sound source signals.

The detected body-conducted artificial speech is converted into more natural speech by a voice conversion technique. Voice conversion is a technique for converting voices of one speaker (so-called source speaker) as if they are uttered by another speaker (so-called target speaker) [11]. This paper adopts the statistical voice conversion method based on maximum likelihood estimation of parameter trajectory [9]. This method includes the training and the conversion part. Therefore, a user needs to record his or her artificial speech as the training data before using our proposed speaking-aid system. In the training part, first, utterance-pairs of a source and a target speaker are time-aligned by dynamic time warping procedure. Next, joint probability density of the source and the target features are modeled by a Gaussian mixture model (GMM) of which a parameter set is trained by Expectation-Maximization (EM) algorithm. In the conversion part, the target features are determined by
maximizing the conditional probability density function of the target features given the source features, which are derived from the trained GMM. Finally, the converted speech is presented as the user’s new voice.

Because the artificial speech does not have effective \( F_0 \) information, the accurate estimation of natural \( F_0 \) contours is difficult \([16]\). To avoid this problem, a voice conversion framework from NAM into whispered voice that also does not have \( F_0 \) information had been proposed and evaluated \([13]\). Inspired by this method, we also convert the body-conducted artificial speech into whispered voice. In our previous work, the proposed system was evaluated using speech data uttered by one non-laryngectomized \([6]\). Our experimental results have demonstrated that both the naturalness and the intelligibility of the converted whispered voice are much better than not only those of the source body-conducted artificial speech with the small sound source signals but also those of another artificial speech using the conventional electrolarynx.

3. Enhancing Auditory Feedback

Auditory feedback plays very important roles to make our articulation stable \([14]\). It is assumed that the user of our system gets the auditory feedback from the output converted speech. However, our system needs to record training data for building the conversion model, and the user cannot get the converted speech as the auditory feedback in the training data recording. Therefore, the user is forced to get the auditory feedback from only the produced soft artificial speech in the training data recording.

The artificial speech for the training data has so far been recorded in a sound-proof room \([6]\). We know from our experiences that the user can get the auditory feedback only from the produced soft artificial speech in the sound-proof room. However, it is not always that the user can record speech data in such a significantly quiet room. Moreover, it would be more practical and convenient for the user that the training data recording is available in other places than the sound-proof room, e.g., the user’s own room. However, it is still questionable whether or not the training data recorded in noisy environments with getting the auditory feedback from only the produced soft artificial speech affects the voice conversion accuracy. If it is difficult for the user to get the auditory feedback, the articulation would be unclear. The use of those training data would cause the insufficient converted speech quality. Therefore, it is meaningful to investigate the effectiveness of a method to keep the speaker’s articulation stable when the training data are recorded under the existence of noisy background noises compared to the sound-proof room.

This paper proposes a simple method to provide the user with amplified auditory feedback to keep the articulation stable. To enhance the auditory feedback, our system directly gives the user’s one ear detected body-conducted artificial speech passed through an amplifier. The effectiveness of providing the amplified auditory feedback is experimentally evaluated in Sect. 5.

4. Designing Sound Source Signals

4.1 Designing Idea

Because the proposed system has the voice conversion part inside, we can design sound source signals from different view points compared to conventional electrolarynxes. This paper designs sound source signals not to degrade the voice conversion accuracy by changing the spectrum and the power independently.

4.2 Changing Spectrum

This paper designs three different types of spectra of pulse train, sierra wave, and compensation wave into the target data. Figure 3 shows the waveforms and the spectra of the designed signals. Fundamental frequency of every sound source signals is set to 100 Hz. Averaged power of each signal is aligned to that of the pulse train.

1 Pulse train

Pulse train is introduced as one of the most basic excitations. It is expected that the pulse train most exactly extracts the impulse response of the user’s neck.

2 Sierra wave

Sierra waves are used to approximate our vocal folds vibration. Figure 3 shows that sierra waves have much more powers in the lower frequency components compared to the pulse train.

3 Compensation wave into the target data

In the training procedure of the GMM in the voice conversion part, we automatically perform the time-alignment be-
between the source and the target data, and then, a parameter set of the GMM is estimated by EM algorithm. Therefore, it might be useful to use the source data of which acoustic characteristics are close to that of the target data in order to improve the time-alignment performance and initial estimation accuracy of the GMM parameter set.

Figure 4 shows the flow chart of designing the compensation wave. First, body-conducted artificial speech using pulse train as described above are recorded. Next, in order to compensate the long term spectral envelope, the difference of the averaged mel-cepstra between the source and the target speech is convolved to a pulse train. As the result, the compensation wave into the target data, whispered voice in this paper, are designed. The designed compensation wave has much more powers in the higher frequency components compared to other two sound source signals. Usually, the power of higher frequency components of speech data detected with a NAM microphone is severely attenuated because of lack of radiation characteristics of lips and so on [10]. On the other hand, whispered voice has much power in the higher frequency components. Therefore, the compensation wave is appropriate to make the source features much closer to the target ones.

4.3 Changing Power

When the signal power is changed, the spectrum is fixed only to the sierra wave described in Sect. 4.2 because the sierra wave has the largest dynamic range in the above three spectra. Fundamental frequency of every sound source signals is set to 100 Hz. The basic power of the sierra wave is defined so that the averaged power of which is equivalent to that of the pulse trains (i.e. same as the sierra wave designed in Sect. 4.2). The maximum and the minimum power of the sierra wave are respectively defined as +18 dB and −27 dB compared to the basic power. Figure 5 shows power histograms of recorded body-conducted artificial speech data using the basic, the minimum, and the maximum power of the sierra wave. For each sound source signal, the distribution at the left side mainly shows the power histogram of the silence part and the other distribution at the right side mainly shows the power histogram of the speaking part, respectively. As the power of the sound source signals is larger, the distance between these two distributions is also larger. On the other hand, when the power of the sound source signal is set to the minimum power, these two distributions are obviously overlapped. This is because the power of the speaking part is almost the same power as that of the silence part.

5. Experimental Evaluations

Two experimental evaluations are conducted to investigate 1) the effectiveness of enhancing the auditory feedback and 2) the impact of the voice conversion accuracy using different sound source signals.

5.1 Effectiveness of Enhancing Auditory Feedback

5.1.1 Experimental Conditions

The user was one non-laryngectomee of a Japanese male speaker who had trained the way to speak using an electrolarynx for 21 days. We recorded his body-conducted artificial speech using the pulse train shown in Fig. 3 in a soundproof room as shown in Fig. 6. The noise level in the soundproof room was almost 31 dB. Assuming indoor noise environments in our daily life, air-conditioner noise [15] was presented from two loud speakers while its power was adjusted so that the noise level at the user’s location was set to 50 dB, 55 dB, 60 dB, and 65 dB, respectively. When enhancing the auditory feedback, the body-conducted artificial speech passed through an amplifier was directly given to the user's one ear with a closed-type ear phone. The volume of the amplified auditory feedback was controlled by the user according to each noise level so that he comfortably listened to it. When not enhancing the auditory feedback, he got the auditory feedback only from the produced artificial speech. Nine sentences of newspaper articles were uttered in each condition and totally 90 utterances (9 sentences, 5 noise levels, with/without auditory feedback) were
In order to clarify whether or not presenting amplified body-conducted artificial speech was effective to enhance the auditory feedback, we firstly conducted a subjective evaluation of the stability of user’s articulation by eight non-laryngectomies. Each listener evaluated all 90 sentences with five-scaled opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). The audio format of the stimuli was set to 16 bit and 16000 Hz. Stimuli were randomly presented to both ears of the listeners by a headphone in the sound-proof room.

In order to evaluate the effectiveness of the enhancing the auditory feedback, we also conducted a voice conversion experiment in which the body-conducted artificial speech was set to the source speech and whispered voice recorded using a head-set microphone uttered by the same person as the source speaker was set to the target speech. The user additionally recorded 75 sentences including 49 phoneme-balanced sentences for the training and 26 newspaper articles for the test. The noise levels were set to 50 dBA and 55 dBA, respectively. The other recording conditions were the same as described above.

The 0th through 24th mel-cepstral coefficients were extracted from the source and the target speech, respectively. In order to alleviate the degradation of the voice conversion accuracy due to the lack of information because of recording the artificial speech using a NAM microphone as described in Sect. 2, segmental feature vectors were used in the previous study about a voice conversion from NAM to normal speech [16]. This paper also used spectral segmental feature vectors which were constructed by the following frame-by-frame procedures: we prepared a vector by concatenating the extracted static feature vectors at a current ±8 frames, and then, the dimension of the concatenated feature vector was reduced by principal component analysis (PCA). Finally, 50-dimensional segmental feature vectors were established as the source data. A joint feature vector of the static and the delta of the first-order information was constructed at each frame to be set to the target data. The number of GMM mixture components was set to 32. The mel-cepstral distortion between the target and the converted mel-cepstra was used as an evaluation metric of the voice conversion accuracy, which was given by:

$$\text{Melcd} \ [\text{dB}] = \frac{1}{T} \sum_{t} \left[ \frac{10}{\ln 10} \sqrt{ \sum_{d=0}^{24} (\text{tar}_t[d] - \text{conv}_t[d])^2} \right]$$

where \(\text{tar}_t[d]\) and \(\text{conv}_t[d]\) are \(d\)th coefficients of the target and the converted mel-cepstrum in the frame \(t\), respectively and \(T\) denotes the number of total frames in each utterance.

5.1.2 Experimental Results

Figure 7 shows the result of the subjective evaluation. When the auditory feedback is not enhanced, the stability of the user’s articulation is obviously degraded. On the other hand, when the auditory feedback is enhanced, the stability of the articulation is kept stable even if the noise level is 65 dBA. Therefore, our method of providing the user with the amplified auditory feedback is effective to enhance the auditory feedback.

Figure 8 shows a waveform example of the recorded...
body-conducted artificial speech with or without amplified auditory feedback. When the auditory feedback is not enhanced, many bursts supposed to appear in consonants are suppressed. On the other hand, the user can clearly utter those phonemes even under the large external noise environment by enhancing the auditory feedback. Therefore, our method of enhancing the auditory feedback is effective of making the articulation stable in the training data recording.

Figure 9 shows the result of a voice conversion experiment. When the auditory feedback is not enhanced, the voice conversion accuracy is significantly degraded. This degradation is still observed even if the noise level is set to 50 dBA. On the other hand, the auditory feedback enhancement yields significant improvements of the voice conversion accuracy. Therefore, our method of enhancing the auditory feedback is significantly useful to record the body-conducted artificial speech used for training the conversion model.

5.2 Impact of Using Different Sound Source Signals on Voice Conversion Accuracy

5.2.1 Experimental Conditions

In order to investigate the impact of using different sound source signals on the voice conversion accuracy, voice conversion experiments were carried out.

As described in Sect. 4, three different spectra of the pulse train, the sierra wave, and the compensation wave were used to investigate the influences of the spectral changes of the sound source signals. Moreover, six types of the sierra wave of which power was set to −27 dB, −18 dB, −9 dB, 0 dB, +9 dB, and +18 dB were also used to investigate the influences of the power changes of the sound source signals. 0 dB meant the basic power described in Sect. 4.3. We recorded the body-conducted artificial speech uttered by the same user as described in Sect. 5.1 using each sound source signal. We also recorded the body-conducted faint speech produced by the user without using any external sound source signals. In this recording, the user articulated only using sounds generated in the oral cavity when uttering some consonants such as /k/. Consequently, the power of the recorded speech without using any external sound source signals was much smaller than that of the body-conducted artificial speech with the −27 dB sierra wave, and this speech was used as the body-conducted artificial speech using the lower limit of the power of sound source signals. Moreover, we also recorded body-conducted artificial speech using a conventional electrolarynx. The power of the body-conducted artificial speech using the conventional electrolarynx was significantly larger than that of the artificial speech using the sierra wave which power was +18 dB, and therefore, this speech was used as the upper limit of the power of the sound source signals. Totally 10 kinds of different artificial speech were recorded. A fundamental frequency of all small sound source signals was set to 100 Hz.

Cross validation using 70 sentences of newspaper articles in which 50 sentences were used for the training and the other 20 sentences were for evaluation was conducted. As the target data in the voice conversion, whispered voice uttered by the same speaker as Sect. 5.1 was recorded with a head-set microphone. The other conditions of speech analysis and the converted model setting were same as Sect. 5.1. We conducted objective evaluations using the same mel-cepstral distortion as described in Sect. 5.1 as an evaluation metric of the voice conversion accuracy.

We also conducted subjective evaluations. 10 non-laryngectomees evaluated the naturalness of the converted whispered voice with a five-graded opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). Stimuli were seven kinds of converted whispered voice which were derived from different sound source signals of only articulation, an electrolarynx, the minimum (−27 dBA), basic (0 dBA), and the maximum (+18 dBA) power of the sierra wave, the compensation wave, and the pulse train. Totally 140 utterances (20 utterances, 7 excitations) were randomly presented to listeners.

5.2.2 Experimental Results

1 Changing spectrum

Figure 10 shows the mel-cepstral distortion between the target and the converted features. Mel-cepstral distortion before voice conversion using the compensation wave (10.0 dB) is slightly better than others (10.8 dB and 11.4 dB for the pulse train and the sierra wave, respectively). On the other hand, after conversion, all conversion accuracies are almost same. We think that the GMM absorbs the differences between source data.

Figure 11 shows the result of the subjective evaluation. The impact by changing the spectrum of the small sound source signals is few just as same as objective result, and we confirm that the voice conversion can accept some sound
source signals with different spectra.

2 Changing power

Figure 12 shows the mel-cepstral distortion between the target and the converted features. The averaged mel-cepstral distortions of -27 dB and articulation are worse than others. The power of the speaking part of these data is almost same as that of the silence part (see Fig. 5). It indicates that we cannot use sound source signals with which the power of the speaking part is almost same as that of the silence part.

From the result of the subjective evaluation shown in Fig. 11, the converted voice qualities of the sierra wave (-27 dB) and articulation are worse than others, which is the same trend as the objective experiment. Although the converted voice quality using +18 dB is slightly degraded, we believe that this slight degradation is caused by the variations of the user’s articulation because the converted voice quality using the conventional electrolarynx is almost same as the converted voice quality using other small sound source signals with the basic power. As the result of objective and subjective evaluations, voice conversion is difficult to accept sound source signals with extremely small powers, which derives almost same speaking power as the silence power as the power histograms shown in Fig. 5. The voice conversion, however, can accept other sound source signals with much power.

From those results, we conclude that the voice conversion can allow large degree of the sound source signals in terms of the spectrum and the power.

6. Conclusion

This paper made two investigations on our speaking-aid system for laryngectomees; 1) the effectiveness of enhancing the auditory feedback, and 2) the impact of the voice conversion accuracy using several sound source signals. About the first concern, the artificial speech detected with a NAM microphone passed through an amplifier was given to the user’s one ear directly. From subjective evaluation, we confirmed that enhancing the auditory feedback helped the user to utter with stable articulation under existing external noises. Moreover, the enhancement of the auditory feedback improved the voice conversion accuracy. About the second concern, the spectrum and the power of the sound source signals were operated independently. In the spectral changes, three kinds of spectrum of pulse train, sierra wave, and compensation wave into target data were designed. In the power changes, the averaged power of the sierra wave was changed every 9 dB, and totally 6 kinds of sierra wave with different powers were designed. From the results of objective and subjective experimental evaluations of voice conversions from body-conducted speech into whispered voice using non-laryngectomee’s data, it was revealed that the voice conversion in our proposed system can accept large degree of both spectra and powers.

In our later works, we will evaluate the proposed system using real laryngectomee’s data, and improve the con-
verted voice quality. Moreover, the source speech will be converted into not only whispered voice but also normal speech.

Acknowledgements

This research was supported in part by MIC SCOPE. This research was also supported in part by Grant-in-Aid for JSPS Fellows.

References


Keigo Nakamura graduated Course of Multimedia Studies, Faculty of Education and Human Sciences, Yokohama National University in Japan in 2005. He received Master of Engineering from graduate Graduate School of Information Science, Nara Institute of Science and Technology (NAIST) in Japan in 2007. Now he is currently in the doctoral course as a Research Fellow of the JSPS in NAIST from 2009. He mainly studies speaking-aid systems for laryngectomees. He has been a student member of ASJ.

Tomoki Toda was born in Aichi, Japan on January 18, 1977. He received the B.E. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 1999 and the M.E. and D.E. degrees in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan, in 2001 and 2003, respectively. He was a Research Fellow of the JSPS in the Graduate School of Engineering, Nagoya Institute of Technology from 2003 to 2005. He is currently an Assistant Professor of the Graduate School of Information Science, NAIST. From March 2001 to March 2003, he was an Intern Researcher at the ATR Spoken Language Translation Research Laboratories, Kyoto, Japan, and then he was a Visiting Researcher at the ATR until March 2006. He has been a Visiting Researcher at the NICT, Kyoto, Japan, since May 2006. He was also a Visiting Researcher at the Language Technologies Institute, CMU, Pittsburgh, USA, from October 2003 to September 2004 and at the Department of Engineering, University of Cambridge, Cambridge, UK, from March 2008 to August 2008, respectively. His research interests include statistical approaches to speech processing such as voice transformation, speech synthesis, speech analysis, speech production, and speech recognition. He received the TELECOM System Technology Award for Students and the TELECOM System Technology Award from the TAF in 2003 and 2008, respectively. He also received the ISS Best Paper Award from the IEICE, Japan, in 2008, the Ericsson Young Scientist Award from Nippon Ericsson K. K. in 2008, and both the Awaay Prize Young Researcher Award and the Hakura Prize Innovative Young Researcher Award from the ASJ in 2009. He has been a member of the Speech and Language Technical Committee of the IEEE SPS since January 2007. He is a member of IEEE, ISCA, IPSJ, and ASJ.

Hiromi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E. and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991, 1993 and 2000, respectively. He joined Intelligent Systems Laboratory, SECOM CO., LTD., Mitaka, Tokyo, Japan, in 1993, where he engaged in the research and development on the ultrasonic array system for the acoustic imaging. He is currently an associate professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include array signal processing, blind source separation, and sound field reproduction. He received the Paper Awards from IEICE in 2001 and 2006. He is a member of the IEEE and the Acoustical Society of Japan.
Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a professor of Nara Institute of Science and Technology (NAIST), where he is directing speech and acoustics laboratory. From 1972 to 1993, he had been working at NTT Laboratories. During 1986-1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories. During 1984-1986, he was a visiting scientist in Carnegie Mellon University. He received the Yonezawa Prize from IEICE in 1975, the Signal Processing Society 1990 Senior Award from IEEE in 1991, the Technical Development Award from ASJ in 1994, IPSJ Yamashita SIG Research Award in 2000, and Paper Award from the Virtual Reality Society of Japan in 2001, IEICE paper award in 2005 and 2006, and Inose award in 2005. He is a fellow of the Institute of Electrical and Electronics, Engineers (IEEE), and Information Processing Society of Japan, and a member of the Acoustical Society of Japan (ASJ) and International Speech Communication Society (ISCA).