J. Acoust. Soc. Jpn. (*E*) **16**, 5 (1995)

Speaker individualities in speech spectral envelopes

Tatsuya Kitamura and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology, Hokuriku,
15 Asahidai Tatsunokuchi-cho, Nomi-gun, Ishikawa, 923-12 Japan

(*Received 2 December 1994*)

The aim of the three psychoacoustic experiments described here was to clarify whether there are speaker individualities in the spectral envelopes, in which frequency bands such individualities exist, and how frequency bands having speaker individualities can be manipulated. The LMA analysis-synthesis system was used to prepare stimuli varied specific frequency bands, and the frequency bands having speaker individualities were estimated experimentally. The results indicate that (1) speaker individualities exist in spectral envelopes, (2) these individualities are mainly at frequencies higher than 22 ERB rate (2212 Hz) and vowel characteristics exist from 12 ERB rate (603 Hz) to 22 ERB rate, and (3) the voice quality can be controlled by replacing the higher frequency band of one talker with that of other talkers. The replace point is the adjacent spectral local minimum below the spectral local maximum around 23 ERB rate in the spectral envelopes.

Key words: Speaker individualities, Spectral envelopes, Voice quality control, LMA analysis-synthesis system

PACS number: 43. 70. Gr, 43. 71. Bp, 43. 72. Ja

1. INTRODUCTION

One of the problems in making a speech recognition system is to cope with speaker individualities, and even the latest systems have difficulties in accurately recognizing phonemes with speaker individualities. People, however, use speaker individualities to recognize speech; they adapt their cognitive systems to the voices of various speakers and perceive phonemes correctly. Modeling this process would facilitate the development of an advanced speaker-independent speech recognition system. Present synthesized speech, on the other hand, is not very natural and is therefore not easy to listen to. The addition of speaker individualities would make synthesized speech more natural and easier to listen.

If we are to improve speech recognition accuracy and the articulation of speech synthesizers, we need manipulate speaker individualities, and to do this we need to specify some physical characteristics related to those individualities. In this paper we assume that the physical characteristics people use to identify speakers are significant physical characteristics representing speaker individualities, and we use psychoacoustical experiments to estimate some of those characteristics in vowels. Specifically, we investigate frequency bands in the spectral envelopes of vowel sounds.

Itoh and Saito[1] and Kuwabara and Ohgushi[2] have reported that there are speaker individualities in spectral envelopes, and Furui and Akagi[3] have shown that these individualities are mainly in the frequency band between 2.5 and 3.5 kHz. These studies, however, have not identified *specific* frequency bands contains speaker individualities. Nor has the relationship between speaker identification and spectral envelope variations in the frequency band between 2.5 and 3.5 kHz been investigated in psychoacoustic experiments. This is because the

LPC analysis-synthesis systems that have been used for manipulating spectral envelopes cannot handle specific frequency bands of the spectral envelopes.[1,2]

The study described here used the log magnitude approximation (LMA) analysis-synthesis system[4,5] which can handle specific frequency bands of the spectral envelopes. The relationship between physical characteristics and the speaker identification rates was studied by using stimuli in which several types of physical characteristics were varied. Experiment 1 shows the existence of speaker individualities in the spectral envelopes, Experiment 2 identifies the specific frequency bands in which these individualities exist, and Experiment 3 shows how these bands can be manipulated.

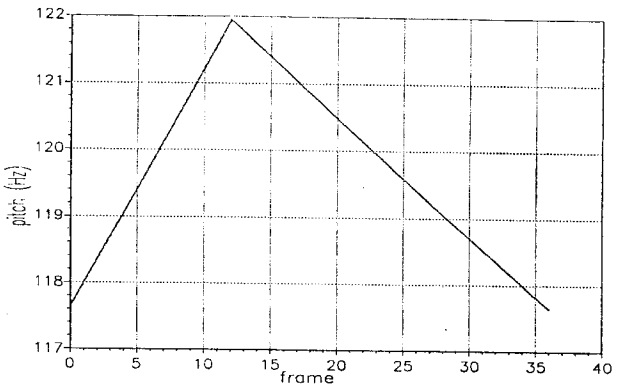

Fig. 1 The pitch contour for the speech waves in stimuli of **1c**, **1d** and **1e**. (Frame length is 51.2 ms and frame shift is 12.8 ms.)

2. EXPERIMENT 1

2.1 Method

Stimuli Five male native Japanese speakers recorded five vowels at a sampling rate of 20 kHz with 16-bit resolution. When uttering vowels, the speakers were forced to tune the pitch of their voices to the same height as that of the 120-Hz pure tone in order to avoid the influence of pitch frequency on the speaker identification tests. The stimuli include several varied types of physical characteristics resynthesized from FFT cepstral data of the original speech waves by the LMA analysis-synthesis system. In the experiment, the following five types of stimuli were used:

1a. original speech waves,
1b. LMA analyzed-synthesized speech waves,
1c. speech waves with fixed power and pitch contour,
1d. speech waves with randomized frame sequence of stimuli **1c** and
1e. speech waves with fixed spectral envelope tilt of stimuli **1d**.

The power of the waves in stimuli **1c**, **1d** and **1e** was fixed to

$$x'(n)=\frac{x(n)}{\sum_{n=1}^{N} x(n)}, \qquad (1)$$

where $x(n)$ is the original speech wave and $x'(n)$ is speech wave with fixed power. The pitch contour for the waves in stimuli **1c**, **1d** and **1e** is shown in Fig. 1. It simulates the pitch contour of the original speech wave. The tilt of the spectral envelopes of **1e** was fixed for each vowel by substituting the averaged first and second FFT cepstra of each vowel for these of the original speech wave. The LMA filter with 60 FFT cepstra was used to synthesize the stimuli, and the duration of each stimuli was approximately 500 ms.

Subjects. The eight listeners (seven males and one female) serving as subjects in Experiment 1 and Experiment 2 were graduate students who were very familiar with speaker voice characteristics. All were native speakers of Japanese and had no known hearing impairments.

Procedure. The stimuli were presented through binaural earphones at a comfortable loudness level in a soundproof room (27.7 dB(A)). Each was presented to the subjects three times randomly at intervals of 5.0 s. The task was to identify vowels and speakers, and when the subjects could not identify speakers or vowels they responded with "X." Speaker identification rates and vowel identification rates for the stimuli were averaged across subjects. This experimental procedure is also used in Experiment 2 and Experiment 3.

2.2 Results and Discussion

The speaker identification rates and the vowel identification rates for Experiment 1 are shown in Fig. 2. Speaker identification rates were tested by using the F-test with 1 and 14 free parameters. The significance level is $F(1, 14; 0.05)=4.60$. The number of factors is two and the number of levels is eight. The results lead to the following four conclusions.

1. There are speaker individualities in the pitch

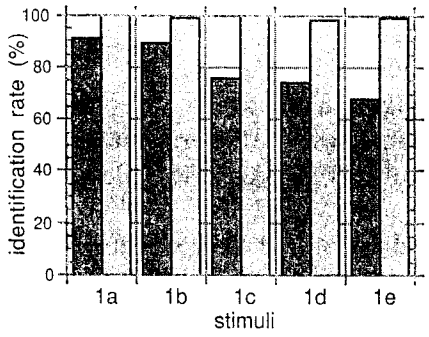

Fig. 2 Speaker (■) and vowel (□) identification rates for Experiment 1.

contour ($F(1, 14)=17.74$ between **1b** and **1c**).

2. There are speaker individualities in the spectral envelope (The speaker identification rate of **1d** is 74.17%).
3. There are no speaker individualities in the spectral envelope sequence or the tilt of the spectral envelopes ($F(1, 14)=0.15$ between **1c** and **1d**, $F(1, 14)=2.56$ between **1d** and **1e**).
4. There are no vowel characteristics in the pitch contour, the spectral envelope sequence, or in the tilt of the spectral envelopes ($F(1, 14)=1.34$ between **1b** and **1c**, $F(1, 14)=0.98$ between **1c** and **1d**, $F(1, 14)=0.11$ between **1d** and **1e**).

3. EXPERIMENT 2

3.1 Analysis of Spectral Envelopes

To identify the frequency bands containing speaker individualities in the spectral envelope, we calculated the variance for the five vowel spectral envelopes of ten male speakers from the ATR speech databases.[6] The spectral envelopes were smoothed with 60 FFT cepstra and the frequency axis was converted into the ERB rate.[7]

Let $E_{ijk}(n)$ be the kth frame log-power spectrum of the jth vowel uttered by the ith speaker at an ERB rate n, where $n=1\sim N$, $k=1\sim K$, $j=1\sim J$ and $i=1\sim I$. (N is the maximum ERB rate, K is the number of frames, J is the number of vowels and I is the number of speakers.) The variance of $E_{ijk}(n)$ with respect to i is given by

$$\sigma_j^2(n)=\frac{1}{I}\sum_{i=1}^{I}\left\{\frac{1}{K}\sum_{k=1}^{K}E_{ijk}(n)-\frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}E_{ijk}(n)\right\}^2, \quad (2)$$

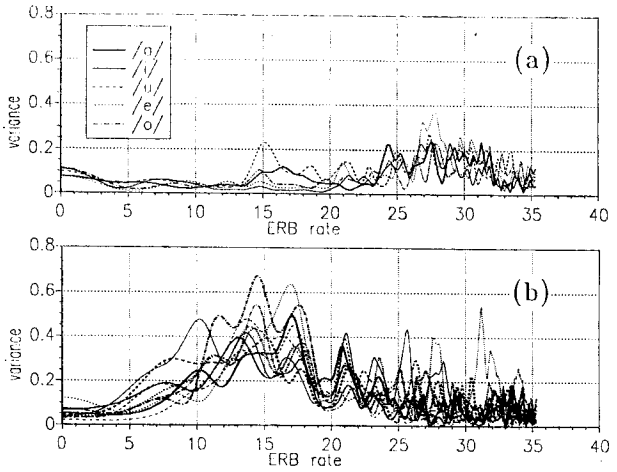

Fig. 3 Variance of $E_{ijk}(n)$: (a) speaker individualities $\sigma_j^2(n)$, (b) vowel characteristics $\sigma_i^2(n)$.

and the frequency bands having large quantities of $\sigma_j^2(n)$ are regarded as providing the speaker individualities.

The variance of $E_{ijk}(n)$ with respect to j is given by

$$\sigma_i^2(n)=\frac{1}{J}\sum_{j=1}^{J}\left\{\frac{1}{K}\sum_{k=1}^{K}E_{ijk}(n)-\frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}E_{ijk}(n)\right\}^2, \quad (3)$$

The frequency bands having large quantities of $\sigma_i^2(n)$ are regarded as providing the vowel characteristics.

The variances $\sigma_j^2(n)$ and $\sigma_i^2(n)$ shown in Fig. 3 indicate that speaker individualities are mainly above the 22 ERB rate (2212 Hz[7]) and that vowel characteristics are mainly exist from 12 ERB rate (603 Hz) to 22 ERB rate.

3.2 Experiment

The results of the analyses in Sec. 3.1 indicate that the speaker individualities exist mainly above 22 ERB rate and that the vowel characteristics exist from 12 to 22 ERB rate. We thus assume that the frequency band above 22 ERB rate contains the speaker individualities and the band from 12 to 22 ERB rate contains the vowel characteristics. The second experiment was designed to test this assumption from a psychoacoustic viewpoint.

3.2.1 Method

Stimuli. The spectral envelopes of the **1d** stimuli of Experiment 1 were varied by using the LMA analysis-synthesis system. Two varying methods were used: in Method 1 the spectral envelopes are reversed symmetrically with respect to their auto-

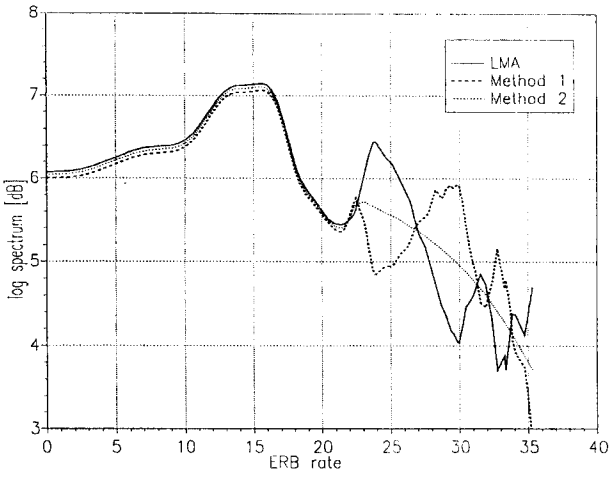

Fig. 4 Spectral envelopes varied by Methods 1 and 2 above 22 ERB rate. 60 FFT cepstra were used to make these envelopes.

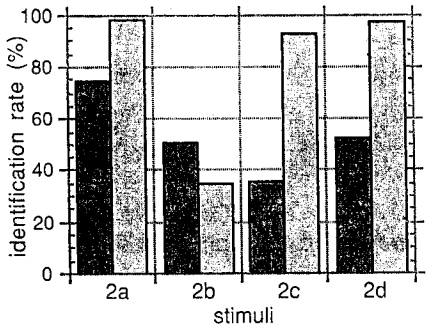

Fig. 5 Speaker (▩) and vowel (□) identification rates for Experiment 2.

regressive line, and in Method 2 the spectral envelopes were replaced by their autoregressive line. The spectral envelopes varied above 22 ERB rate by these two methods are shown in Fig. 4. The connection point between the original spectral envelope and its autoregressive line or between the original spectral envelope and its reversed spectral envelopes is not discontinuous as shown in Fig. 4. This is because the varied spectral envelopes were translated into 1024 FFT cepstra and 60 FFT cepstra were used to make the LMA filter. The types of stimuli were as follows:

2a. LMA analyzed-synthesized speech waves,

2b. speech waves varied by Method 1 from 12 to 22 ERB rate,

2c. speech waves varied by Method 1 above 22 ERB rate and

2d. speech waves varied by Method 2 above 22 ERB rate.

3.2.2 Results and discussion

The speaker identification rates and the vowel identification rates for Experiment 2 are shown in Fig. 5, and they suggest the following four conclusions.

1. The distortion of the spectral envelopes above 22 ERB rate does not affect vowel identification but does affect speaker identification ($F(1, 14)=4.51$ between **2a** and **2c** for the vowel identification rate, $F(1, 14)=88.90$ between **2a** and **2c** for the speaker identification rate).
2. The distortion of the spectral envelopes from 12 to 22 ERB rate affects vowel identification ($F(1, 14)=342.85$ between **2a** and **2b** for the vowel identification rate). This distortion affects speaker identification rates less than does the distortion of the spectral envelopes above 22 ERB rate ($F(1, 14)=11.84$ between **2b** and **2c** for the speaker identification rate).
3. The vowel identification rate is lower than the speaker identification rate of **2b** ($F(1, 14)=12.04$ between the vowel identification rate and the speaker identification of **2b**).
4. Method 1 affects speaker identification rates more than Method 2 does ($F(1, 14)=14.32$ between **2c** and **2d** for the speaker identification rate).

Conclusions 1 and 2 mean that the speaker individualities and the vowel characteristics can be controlled. If the frequency band of the spectral envelopes above 22 ERB rate can be manipulated, speaker normalization and adaptation methods for speech recognition can be constructed without affecting the vowel recognition rates. Conclusion 3 means that humans can identify speakers even when listening to unidentified vowels, and Conclusion 4 means that the relationship between the local maxima and the local minima in the spectral envelopes is important in identifying speakers.

4. EXPERIMENT 3

Experiment 2 indicates that speaker individualities are mainly above 22 ERB rate and that the relationship between the local maxima and the local minima in the spectral envelopes is important in identifying speakers. Observing spectral envelopes above 22 ERB rate, we can see that there is a spectral

local maximum around 23 ERB rate (2489 Hz). The spectral local maximum is regarded as the third formant, F_3, which Kuwabara and Ohgushi[2] have suggested is the most important factor for identifying speakers. We thus designed the third experiment to investigate whether the spectral local maximum around 23 ERB rate is significant for identifying speakers and whether speaker individualities can be controlled by manipulating spectral envelopes.

4.1 Method

Stimuli. The steps for making the stimuli were as follows. Three male native Japanese speakers recorded five vowels under the same conditions as in Experiment 1, but these speakers were different from those of Experiments 1 and 2. Let $E_{ijk}(n)$ be the kth frame spectral envelope of the jth vowel uttered by the ith speaker for an ERB rate n. $E_{ijk}(n)$ is averaged with respect to k:

$$E_{ij}(n)=\frac{1}{K}\sum_{k=1}^{K}E_{ijk}(n)\,. \qquad (4)$$

And $E_{ij}(n)$ is averaged with respect to i:

$$E_{j}(n)=\frac{1}{I}\sum_{i=1}^{I}E_{ij}(n)\,. \qquad (5)$$

The higher frequency band of $E_j(n)$ is replaced by that of $E_{ij}(n)$. The replace points are the adjacent local minima (a) below and (b) above the spectral local maximum around 23 ERB rate (see Fig. 6). The local maximum was chosen by hand. The stimuli were synthesized from the spectral envelopes by using the LMA analysis-synthesis system, and the duration of each stimulus was approximately 500 ms. The types of stimuli were as follows:

3a. original speech waves,

3b. LMA analyzed-synthesized speech waves,

3c. speech waves with randomized frame sequence and fixed power and pitch contour,

3d. speech waves with replaced higher frequency spectral envelopes. The replace point was the adjacent local minimum below the spectral local maximum and

3e. speech waves with replaced higher frequency spectral envelopes. The replace point was the adjacent local minimum above the spectral local maximum.

Subjects. The nine male listeners serving as subjects in this experiment were graduate students who were very familiar with speaker voice characteristics. All subjects were native speakers of Japanese and had no known hearing impairments.

4.2 Results and Discussion

The vowel identification rates are above 99% for all stimuli and there are no significant differences in the vowel identification rates for each stimulus ($F(4, 40)=0.60$, $F(4, 40; 0.05)=2.61$). This indicates that these experimental manipulations of the spectral envelopes do not influence vowel identification.

The speaker identification rates for Experiment 3 are shown in Fig. 7.

1. For /a/, /o/ and /u/ the speaker identification rates of **3d** are close to that of **3c** and different from that of **3e** ($F(1, 16)=9.67$ between **3c** and **3d** for /a/, $F(1, 16)=25.27$ between **3d** and **3e** for /a/, $F(1, 16)=3.66$ between **3c** and **3d** for /o/, $F(1, 16)=95.01$ between **3d** and **3e** for /o/, and $F(1, 16)=1.56$ between **3c** and

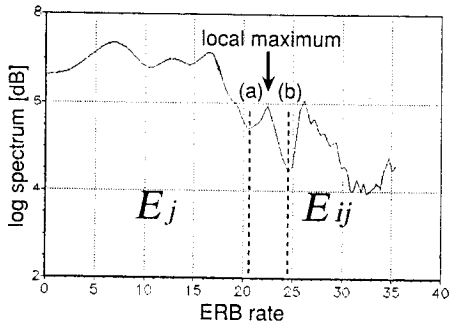

Fig. 6 The higher frequency band of E_j was replaced by that of E_{ij}. The replace points were the adjacent local minima (a) blow and (b) above the spectral local maximum around 23 ERB rate.

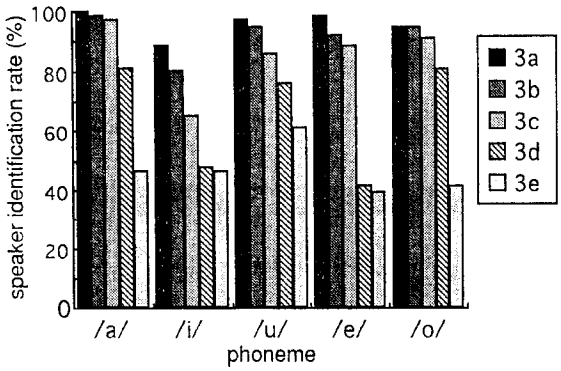

Fig. 7 Speaker identification rates for Experiment 3.

3d for /u/, $F(1, 16)=3.45$ between **3d** and **3e** for /u/). These results indicate that the spectral local maximum around 23 ERB rate is significant for voice quality control. Speaker individualities can thus be controlled without influencing vowel identification by replacing the higher frequency band in which speaker individualities exist.

2. For /i/ there are significant differences in the speaker identification rates between **3c** and **3d** ($F(1, 16)=13.51$), and none between **3d** and **3e** for /i/ ($F(1, 16)=0.03$). One possible reason for low speaker identification rate of **3d** is that the spectral local maximum around 23 ERB rate is not significant for voice quality control for /i/. Another possibility is that the speaker individualities of /i/ exist in the pitch contour and the whole spectral envelopes because of the significant differences between **3b** and **3c** only for /i/ ($F(1, 16)=5.24$).
3. For /e/ there are significant differences in the speaker identification rates between **3c** and **3d** ($F(1, 16)=43.40$) and there are none between **3d** and **3e** ($F(1, 16)=0.15$). This shows that the local maximum is not significant for voice quality control. From another viewpoint, it shows that averaging the lower frequency band reduces the accuracy of speaker identification. Thus, speaker individualities of /e/ may be mainly in the lower frequency band. There is other evidence supporting this conclusion: the speaker identification rate for /e/ of **2c** in Experiment 2 is 46.6% and the mean rate of other phonemes is 32.4%. This shows that the effect of the higher frequency band distortion is less for /e/ than for other phonemes.

Note that $F(1, 16; 0.05)$ is 4.49.

5. GENERAL DISCUSSION

These three experiments show that speaker individualities in the spectral envelopes are mainly above 22 ERB rate and that, under these experimental conditions, they can be controlled without influencing vowel identification by manipulating the frequency band higher than around 23 ERB rate. These results show that people identify speakers of vowels mainly by using the higher frequency band in the spectral envelopes.

If the speaker individualities in the higher frequency band can be used by automatic recognition systems, advanced speaker-independent speech recognition and speaker recognition systems can be constructed. The accuracy of a speaker-independent speech recognition system can be increased by suppressing the influence of speaker individualities in the higher frequency band. Conversely, the accuracy of speaker recognition system may be improved by selectively using the speaker individualities in the higher frequency band.

Speaker individualities in the higher frequency band in spectral envelopes can also be used directly for speech synthesis. Speaker individualities of synthesized speech can be controlled by manipulating of the higher frequency band, thus the synthesis of speech is more natural.

It is likely that the roles of dynamic characteristics play in identifying speakers in continuous speech are more important than those played by the static characteristics investigated in this paper. Our future work will be to investigate the relationship between speaker individualities and spectral dynamic characteristics.

REFERENCES

1) K. Itoh and S. Saito, "Effects of acoustical feature parameters of speech on perceptual identification of speaker," Trans. IEICE **J65-A**, 101–108 (1982) (in Japanese).
2) H. Kuwabara and K. Ohgushi, "The role of formant frequencies and bandwidths in the perception of speaker," Trans. IEICE **J69-A**, 509–517 (1986) (in Japanese).
3) S. Furui and M. Akagi, "Perception of voice individuality and physical correlates," Tech. Rep. Hear. Acoust. Soc. Jpn. H85-18, 1–8 (1985).
4) S. Imai and T. Kitamura, "Speech analysis synthesis system using the log magnitude approximation filter," Trans. IEICE **J61-A**, 527–534 (1978) (in Japanese).
5) S. Imai, "Log magnitude approximation (LMA) filter," Trans. IEICE **J63-A**, 886–893 (1980) (in Japanese).
6) K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe, and H. Kuwabara, "Speech database user's manual," ATR Tech. Rep. TR-I-0028 (1988).
7) B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138 (1990).
8) T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," Proc. ICSLP 94, Vol. 3, 1183–1186 (1994).

Tatsuya Kitamura was born in Yamagata, Japan, on Dec. 25, 1969. He received the B.E. degree in Information Engineering from Yamagata University, Yamagata, Japan, in 1992, and the M.E. in Information Science from Japan Advanced Institute of Science and Technology, Hokuriku (JAIST), Ishikawa, Japan, in 1994. He is presently a graduate student of JAIST and a Research Fellow of the Japan Society for the Promotion of Science. He is a member of the Acoustical Society of Japan, and the Institute of Electronics, Information and Communication Engineers of Japan.

Masato Akagi was born in Okayama, Japan on September 12, 1956. He received a B.E. degree from Nagoya Institute of Technology in 1979, and M.E. and D.E. degrees from Tokyo Institute of Technology in 1981 and 1984, respectively. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been with School of Information Science, Japan Advanced Institute of Science and Technology, Hokuriku (JAIST) and now he is an Associate Professor of JAIST. His research interests include speech perception, modeling of speech perception mechanisms of humans, and signal processing of speech. During 1988, he joined the Research Laboratories of Electronics, MIT as a visiting researcher and in 1993, he studied at the Institute of Phonetic Science, Univ. of Amsterdam. Dr. Akagi is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of Japan (ASJ), the Institute of Electrical and Electronic Engineering (IEEE), the Acoustical Society of America (ASA), and the European Speech Communication Association (ESCA).