Acoustical Society of America and Acoustical Society of Japan

Third Joint Meeting

Honolulu, Hawaii, 2-6 December 1996

Relationship between physical characteristics and speaker individualities in speech spectral envelopes

T. KITAMURA and M. AKAGI

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai Tatsunokuchi, Nomi, Ishikawa 923-12, Japan
E-mail: kitamura@jaist.ac.jp, akagi@jaist.ac.jp

Abstract

Frequency bands having speaker individualities in the spectral envelopes of vowels and physical characteristics representing speaker individualities in these frequency bands were investigated by psychoacoustic experiments. In this study, the relationship between physical characteristics and the speaker identification rates was studied using stimuli, re-synthesized by varying a specific frequency band in the spectral envelopes. The results lead to the following conclusions. 1) The peaks in the spectral envelopes were more significant than the dips for speaker identification. 2) Speaker individualities mainly exist in the frequency band higher than the peak around 20 ERB rate (1740 Hz) and the voice quality can be controlled by replacing the frequency band of one speaker with that of other speakers.

Key words speaker individualities, speaker identification, spectral envelopes

1 Introduction

Physical characteristics representing speaker individualities have been studied as a means of speaker individuality control. Results of these studies are important not only for voice quality control but also for speaker-independent speech recognition. It is also significant from the viewpoint of explanation of human's auditory perception process.

In this paper we assume that the physical characteristics people use to identify speakers are some significant physical characteristics of speech representing speaker individualities, and we use psychoacoustical experiments to estimate some of those characteristics in vowels. Frequency bands having speaker individualities in the spectral envelopes of vowel and significant physical cues for speaker identification in these bands are investigated.

Previous studies have not been able to clarify the frequency band or physical characteristics representing speaker individualities in the spectral envelopes. It is because analysis-synthesis systems used in these studies could not handle a specific frequency band of the spectral envelope[1],[2],[3].

The study described here used the Log Magnitude Approximation (LMA) analysis-synthesis system[4] which can handle specific frequency bands of the spectral envelopes. The relationship between physical characteristics and the speaker identification rates was studied by varying the physical characteristics of the stimuli.

Our previous study[5] has shown that speaker individualities in spectral envelopes of vowel are mainly at above F3 . However, this study used only three male speakers, thus the results might depend on the speaker set. Experiment 1 of this study reconfirms that the speaker individualities mainly exist in higher frequency using vowels of nine male speakers. Experiment 2 investigates the significant physical cues for speaker identification in higher frequency band of the spectral envelopes, and Experiment 3 identifies the specific frequency band in which these individualities exist.

2 Experiment 1

Experiment 1 was an ABX test to find out the frequency band having the speaker individualities in spectral envelopes of vowels.

2.1 Method

Stimuli. In Experiment 1, five Japanese vowels in the ATR speech database[6] uttered by nine Japanese speakers were used. The vowels were recorded at a sampling rate of 20 kHz with 16-bit resolution.

The *stimuli* were vowels re-synthesized from their 60 FFT cepstra by using the Log Magnitude Approximation (LMA) analysis-synthesis system[4]. FFT cepstra were computed by improved cepstral method[7], and averaged with respect to frames in the voiced parts. The acceleration parameter was 1.0 and iteration was 3. Frame length was 25.6 ms and frame shift was 12.8 ms. Let $E_{ij}(n)$ be the spectral envelope of the jth vowel ($j = 1 \sim 5$)uttered by the ith speaker ($i = 1 \sim 9$) for an ERB rate n, and $E'_{ij}(n)$ be that of stimulus. The types of stimuli were as follows:

A. speech waves synthesized using the spectral envelope of one of the speakers,

$$E'_{ij}(n) = E_{ij}(n)$$

B. speech waves synthesized using the averaged spectral envelope across the speakers,

$$E'_{ij}(n) = E_j(n)$$

where,

$$E_j(n) = \tfrac{1}{9}\textstyle\sum_{i=1}^{9} E_{ij}(n),$$

and

X. speech waves synthesized using the spectral envelope of **B** replaced by that of **A** in the following frequency band. Spectral envelopes correspond to the same vowels.

$$E'_{ij}(n) = \begin{cases} E_{ij}(n) & \textit{the following band} \\ E_j(n) & \textit{otherwise} \end{cases}$$

X1. whole frequency band (**X1** was equal to **A**)

X2. from 0 to 10 ERB rate (from 0 to 442 Hz[8])

X3. from 10 to 20 ERB rate (from 442 to 1740 Hz)

X4. from 20 to 30 ERB rate (from 1740 to 5544 Hz)

The pitch frequency of the stimuli was 130 Hz, which was the mean of the pitch frequencies of each speaker. Duration of each stimulus was 0.5 s, and the amplitude was normalized.

Since physical characteristics were normalized expect for the spectral envelops, the results of this experiment, also Experiment 2 and 3, were caused by the difference between the spectral envelopes of stimuli.

Subjects. Ten listeners (nine males and one female) were employed in this experiment. All listeners had no known hearing impairments.

Procedure. The stimuli of the same vowel corresponding to **A**, **B**, and **X**, were presented at intervals of 2.0 s in the order of **ABX**. They were also presented in order of **BAX** to cancel the successive effect. Each set of stimuli was presented three times randomly. Subjects were asked to answer which speaker of stimulus **A** or **B** was similar to that of stimulus **X**.

The low pass filtered stimuli with a cut off frequency of 8 kHz (33.3 ERB rate) were presented through binaural earphones at a comfortable loudness level in a soundproof room (27.7 dB(A)). The subjects were allowed to listen to the stimuli repeatedly.

2.2 Results and Discussion

Figure 1 shows percentages of stimulus **X** identified as speech uttered by the speaker of stimulus **A**. The percentage increases as the frequency band of replacement is higher. This shows that the higher frequency bands are more important for speaker identification.

This results were gotten through an ABX test. So, the subjects might used timbre of stimuli besides the speaker individualities to identify. Hence, in Experiment 2 and 3, a naming test would be used to investigate the effects of the speaker individualities only.

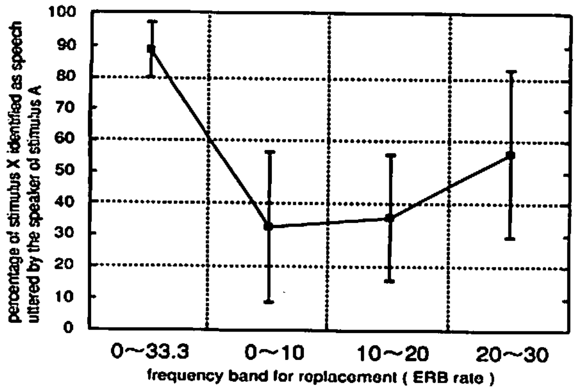

Figure 1: The percentages of stimuli **X** identified as speech uttered by the speaker of stimuli **A**.

3 Experiment 2

Experiment 2 was performed to investigate significant physical cues for speaker identification in the spectral envelope above F3. The effect of peaks and dips, which decide the shape of the spectral envelope, for speaker identification was investigated.

3.1 Method

Stimuli. Five Japanese vowels spoken by five male native Japanese speakers were recorded at a sampling rate of 20 kHz with 16-bit resolution. Speakers' ages were from 24 to 26. When uttering vowels, the speakers were forced to tune the pitch of their voices to the same level as that of the 125-Hz pure tone in order to avoid the influence of pitch frequency on the speaker identification tests. The vowels were stored on a computer disk, and a 200-ms steady part of that was used as the speech wave.

The stimuli were vowels which spectral envelope varied using the LMA analysis-synthesis system as in Experiment 1. The pitch frequency of the stimuli is 125 Hz. Let $E_{ij}(n)$ be the spectral envelope of the jth vowel ($j = 1 \sim 5$) uttered by the ith speaker ($i = 1 \sim 5$) for an ERB rate n, $E'_{ij}(n)$ be that of stimulus, and $R(n)$ be the auto-regressive line of the spectral envelope above F3. The types of stimulus were as follows:

2a. LMA analyzed-synthesized speech waves,

$$E'_{ij}(n) = E_{ij}(n)$$

2b. speech waves with the spectral envelope after eliminated dips above F3,

$$E'_{ij}(n) = \begin{cases} E_{ij}(n) & n < F3 \\ \max[E_{ij}(n), R(n)] & n \geq F3 \end{cases}$$

2c. speech waves with the spectral envelopes after eliminated peaks above F3, and

$$E'_{ij}(n) = \begin{cases} E_{ij}(n) & n < F3 \\ \min[E_{ij}(n), R(n)] & n \geq F3 \end{cases}$$

2d. speech waves with the spectral envelopes after eliminated peaks and dips above F3.

$$E'_{ij}(n) = \begin{cases} E_{ij}(n) & n < F3 \\ R(n) & n \geq F3 \end{cases}$$

F3 was decided manually. Figure 2 shows the spectral envelopes of these stimuli. Before the experiment, it was confirmed that phoneme of these stimuli could be identified as the corresponding original vowels.

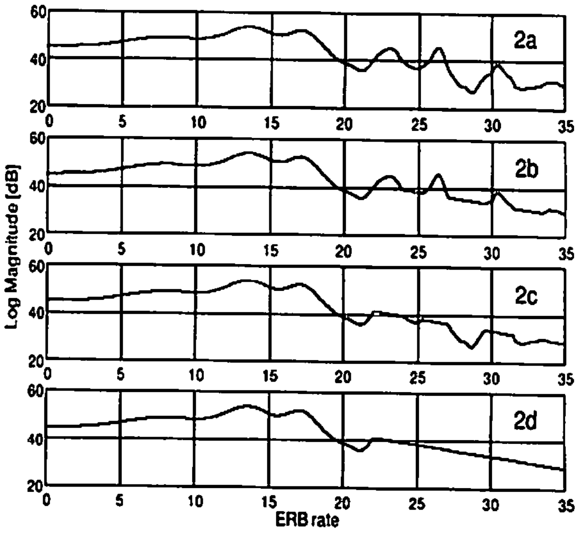

Figure 2: The spectral envelopes of /a/ for **2a**, **2b**, **2c**, and **2d**.

Subjects. Six male listeners were employed in the experiment. They were very familiar with speaker voice characteristics. All listeners had no known hearing impairments.

Procedure. The low pass filtered stimuli with a cut off frequency of 8 kHz were presented to the subjects through binaural earphones at a comfortable loudness level in a soundproof room. Each was presented to the subjects five times randomly. The subjects were allowed to listen repeatedly and they were asked to identify the speaker of the stimuli.

3.2 Results and Discussion

Table 1 shows speaker identification rates of **2a** averaged across subjects and standard deviation. The large standard deviations indicate that the speaker identification accuracy of each subject, using only difference of the spectral envelopes, are widely distributed. Thus, differences between the speaker identification rate of **2a** and that of each other were used to compare the effects of the elimination of the peaks and/or the dips.

The differences averaged across the subjects are shown in Figure 3. The differences were tested by F-test. The critical F value is $F(1,58) = 4.01, p < .05$. The results show that there are significant differences between **2a** and **2b** ($F(1,58) = 21.1$), between **2b** and **2c** ($F(1,58) = 8.45$), and between **2c** and **2d**($F(1,58) = 7.27$), so it can be concluded that **2a** > **2b** > **2c** > **2d** in the speaker identification rates. This indicates that both the peaks and the dips are significant for speaker identification, but the peaks are more significant than the dips. The results also implied that the power gap between the peaks and the dips could be significant.

Table 1: Speaker identification rates of **2a** across subjects (%) and standard deviation.

	/a/	/i/	/u/	/e/	/o/
Speaker identification rate	91.3	92.0	83.3	94.0	67.3
Standard deviation	6.3	7.7	11.2	11.7	11.2

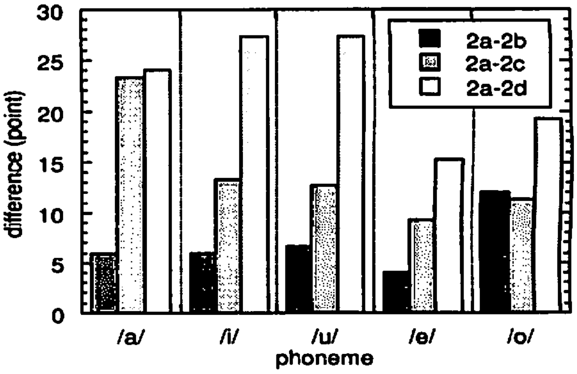

Figure 3: The difference, **2a-2b**, **2a-2c**, and **2a-2d** averaged across the subjects.

4 Experiment 3

Our previous study[5] had shown that the speaker individualities in the spectral envelopes were mainly above F3 for /a/, /o/, and /u/, but it is not the case for /e/ and /i/. On the other hand, the results of Experiment 1 indicate that the speaker individualities of vowels are mainly above 20 ERB rate (1740 Hz), and F2 of /e/ and /i/ is around 20 ERB rate.

From these points, we thus assume that the speaker individualities in the spectral envelopes of vowel exist mainly in and above the peak region around 20 ERB rate. Experiment 3 was designed to test this assumption.

Here after, the frequency band in and above the peak region around 20 ERB rate is denoted in italic as the *higher band*, and that under the peak region is denoted as the *lower band*. These bands are shown in Figure 4.

4.1 Method

Stimuli. The five vowels used in Experiment 2 were used. The stimuli were vowels which the spectral envelope varied using the LMA analysis-synthesis system. The types of stimuli were as follows:

3a. speech waves with the spectral envelope averaged across the speakers for the *lower band.*

Our previous study[9] had shown that if large magnitude existed in the spectral envelopes in the frequency band from 0 to 10 ERB rate, these were significant for speaker identification. Hence, for the frequency band from 0 to 10 ERB rate, the

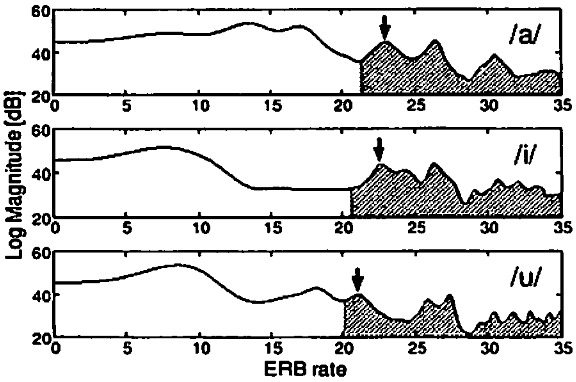

Figure 4: Darken area is the *higher band* and undarken area is the *lower band*. Arrow points at the peak around 20 ERB rate.

spectral envelopes of two subjects having large magnitude on this frequency band and that of other three subjects were averaged respectively. For the frequency band from 10 ERB rate to the peak around 20 ERB rate, that of all the five subject were averaged.

For the *higher band*,

$$E'_{ij}(n) = \max[E_{ij}(n), R(n)]$$

where, $R(n)$ is the auto-regressive line of the *higher band*.

3b. speech waves with peaks of the spectral envelope in the *higher band* of **3a** apploximated by triangls whose vertex and width were same as those of the original peaks.

The *higher band* was decided manually. Figure 5 shows the spectral envelopes of these stimuli. Before the experiment, it was confirmed that phoneme of these stimuli could be identified as the corresponding original vowels.

Subjects. The six male listeners of Experiment 2 were employed.

Procedure. Procedure was same as that of Experiment 2.

4.2 Results and Discussion

Differences between the speaker identification rate of **2a** and that of **3a** and **3b**, were also used. Figure 6 shows that the differences averaged across the subjects.

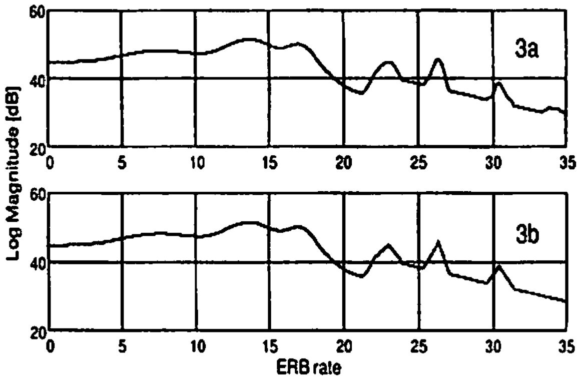

Figure 5: The spectral envelopes of /a/ of **3a** and **3b**.

The differences were tested by F-test. The results lead to the following conclusions.

1. There is no significant difference between **3a** and **2b** ($F(1,58) = 3.76$). The critical F value is $F(1,58) = 4.01, p < .05$. This indicates that averaging the spectral envelopes in the *lower band* across the speakers does not affect speaker identification accuracy, and therefore the *higher band* is more significant. It means that voice quality can be controlled by manipulating the *higher band*.

2. There is no significant difference among vowels for **3a** ($F(4,25) = 0.91$). The critical F value is $F(4,25) = 2.76, p < .05$. This suggests that the speaker individuality in the spectral envelope of each vowel exist mainly in the *higher band*, justifying the assumption made above.

3. There is no significant difference between **3a** and **3b** ($F(1,58) = 1.26$). This implies that the speaker individuality in the spectral envelope can be approximated to a significant degree by frequency and bandwidth of the peaks in the *higher band*, and these parameters of the peaks are significant for speaker identification.

But still, almost all the subjects reported a difficulty in identifying the speaker for stimulus **3b**, it seems that the approximation reduces the speaker individualities.

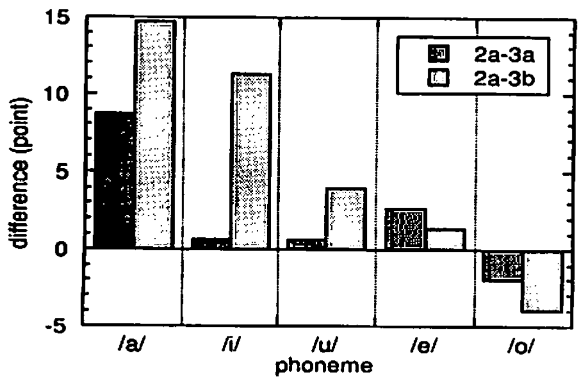

Figure 6: The difference, **2a-3a** and **2a-3b** averaged across the subjects.

5 General Discussion

Under these experimental conditions, the three experiments show that speaker individualities in the spectral envelopes of the vowels are mainly in and above the peak around 20 ERB rate (*higher band*). This also means that humans mainly used the *higher band* for speaker identification.

In the *higher band*, both the peaks and the dips are significant for speaker identification, but the peaks are more significant than the dips. This agrees with the previous knowledge that peaks are more significant for the human's auditory perception.

Even if these peaks were approximated by triangles, speaker individuality still remained. This means that the speaker individualities in the spectral envelopes can be approximated to a significant degree by the frequency and bandwidth of these peaks.

Furui and Akagi[2] have shown that the speaker individualities exists mainly in frequency band from 2.5 to 3.5 kHz, which in fact is included in the *higher band*.

In addition, there is a possible agreement between the results of this study and that of Dang and Honda's study[10] from the standpoint of speech production. They had shown that the acoustic effect of the pyriform fossa, which is anatomically part of the vocal tract, extend widely over the frequency band from 2 to 6 kHz. This band is in good agreement with the *higher band*. It implies that the fossa affects the speaker individualities in the spectral envelopes.

In this study, static characteristics in speaker individualities in steady vowels were investigated. It seems that dynamic characteristics in speaker individualities in continuous speech are also significant for speaker identification. Our future work will be to investigate the relationship between speaker individualities and spectral dynamic characteristics.

Acknowledgment

This study was partially supported by Japan Society for the Promotion of Science (No. 6157) and Grant-in-Aid for Scientific Research (No. 07680388).

References

[1] Itoh, K. and Saito, S., "Effects of Acoustical Feature Parameters of Speech on Perceptual Identification of Speaker," IEICE, Vol. J65-A No. 1, 101-108 (1982) (in Japanese).

[2] S. Furui and M. Akagi, "Perception of voice individuality and physical correlates", Tech. Rep. ASJ, H85-18 (1985).

[3] Kuwabara, H. and Ohgushi, K., "The Role of Formant Frequencies and Bandwidths in the Perception of Speaker," Trans. IEICE, Vol. J69-A No. 4, pp. 509-517 (1986) (in Japanese).

[4] S. Imai and T. Kitamura, "Speech analysis synthesis system using the log magnitude approximation filter," Trans. IEICE, Vol. J61-A No. 6, pp. 527-534 (1978) (in Japanese).

[5] T. Kitamura and M. Akagi, "Speaker Individualities in Speech Spectral Envelopes", J. Acoust. Soc. Jpn., Vol. 16, No. 5, pp. 283-289 (1995).

[6] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe, and H. Kuwabara, "Speech database user's manual", ATR Tech. Rep., TR-I-0028 (1988) (in Japanese).

[7] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method", Trans. IEICE, Vol. J62-A No. 4, pp. 217-223 (in Japanese).

[8] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data", Hearing Research, 47, pp. 103-138 (1990).

[9] T. Kitamura, N. Takagi, and M. Akagi "Frequency bands having speaker individualities", Tech Rep. IEICE, SP95-37 (1995) (in Japanese).

[10] J. Dang and K. Honda, "Acoustic effects of the pyriform fossa on vowel spectra", Tech. Rep. IEICE, SP95-10 (1995) (in Japanese).