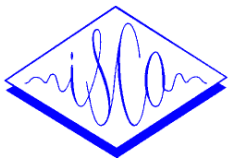

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Speaker Individualities in Speech Spectral Envelopes

Tatsuya Kitamura and Masato Akagi

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

Abstract

Physical characteristics representing speaker individualities embedded in the spectral envelopes of vowels are investigated through four psychoacoustic experiments. The LMA analysis-synthesis system is used to prepare stimuli varying specific frequency bands in the spectral envelopes and the frequency bands having speaker individualities are estimated. The experimental results suggest that speaker individualities mainly exist at above the 23.5 ERB rate (2340 Hz) in the spectral envelopes and that they can be controlled without influencing vowel identification. More detailed information on the spectral envelopes is required for speaker identification than for vowel identification.

1 Introduction

One of the problems in implementing a speech recognition system is to improve speech recognition accuracy against speech individualities. The recognition of phonemes with speaker individualities is difficult for even the latest systems. Humans, however, can adapt their cognitive organs to various talker voices and perceive phonemes correctly. If this process can be modeled, an advanced speaker independent speech recognition system can be constructed.

This paper, assuming that the physical characteristics used for speaker identification by humans are significant physical characteristics representing speaker individualities, estimates some physical characteristics representing speaker individualities through psychoacostical experiments. Specifically, frequency bands in the spectral envelopes are used for investigation.

Previous studies have not clarified the frequency bands in the spectral envelopes that contain speaker individualities because the LPC analysis-synthesis systems that have been used for manipulating spectral envelopes cannot handle specific frequency bands of the spectral envelope[1][2]. The study described here, however, uses the Log Magnitude Approximation (LMA) analysis-synthesis system[3][4], which can handle specific frequency bands of the spectral envelopes.

The relationship between physical characteristics and the speaker identification rate is studied by using stimuli in which several types of physical characteristics are varied. Experiment 1 shows whether speaker individualities exist in the spectral envelopes. Experiment 2 investigates the relationship between detailed structures of the spectral envelopes and speaker individualities. Experiment 3 shows in which frequency bands speaker individualities mainly exist. Experiment 4 shows how the frequency bands having speaker individualities are manipulated.

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 are forced to tune to the same height as the 120 Hz pure tone to avoid the influence of pitch frequency on the speaker identification tests. The stimuli are include several varied types of physical characteristics re-synthesized from their FFT cepstral data by the LMA analysis-synthesis system. The LMA filter with 60 FFT cepstrums is adopted to synthesize the stimuli. The duration of each stimuli is approximately 500 ms. Five types of stimuli are used.

1a. original speech waves

1b. LMA analyzed-synthesized speech waves

1c. speech waves with fixed power and pitch contour

1d. speech waves randomized frame sequence of 1c

1e. speech waves fixed the tilt of the spectral envelopes of 1d

Subjects. The eight listeners (seven males and one female) sarving as subjects in the four experiments were graduate students at JAIST who are very familiar with speaker voice characteristics. All subjects are native speakers of Japanese with no known hearing impairments.

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

2.2 Results and Discussion

The speaker identification rates and the vowel identification rates for Experiment 1 are shown in Figure 1. The results lead to the following four conclusions.

1. There are speaker individualities in the pitch 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 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).

Note that F(1,14;0.05) is 4.60.

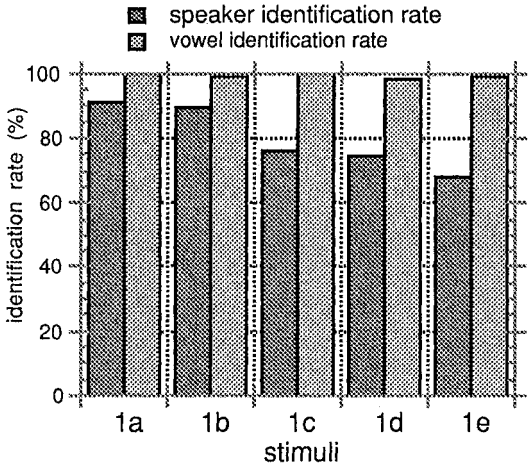

Figure 1: Speaker and vowel identification rate for Experiment 1.

3 Experiment 2

3.1 Method

Stimuli. The stimuli are the same as 1d of Experiment 1, except that the number of FFT cepstrums of the LMA filter is varied from 10 to 60, where 60 represents a more detailed spectral envelope structure than 10.

3.2 Results and Discussion

The speaker identification rates and the vowel identification rates for Experiment 2 are shown in Figure 2.

The results suggest the following two conclusions.

1. As the number of FFT cepstrums decreases, the first significant difference appears in the speaker identification rates between 25 and 20 FFT cepstrums (F(1,14)=0.21 between 30 and 25, F(1,14)=6.03 between 25 and 20).
2. As the number of FFT cepstrums decreases, the first significant difference appears in the vowel identification rates between 20 and 15 FFT cepstrums with (F(1,14)=4.23 between 25 and 20, F(1,14)=10.51 between 20 and 15).

The results show that speaker identification needs more detailed information on the spectral envelopes than vowel identification. If the FFT cepstrum is used in a speaker recognition system, more than 25 FFT cepstrums have to be used at 20 kHz sampling rate, whereas 20 FFT cepstrums are adequate for a vowel recognition system at that sampling rate.

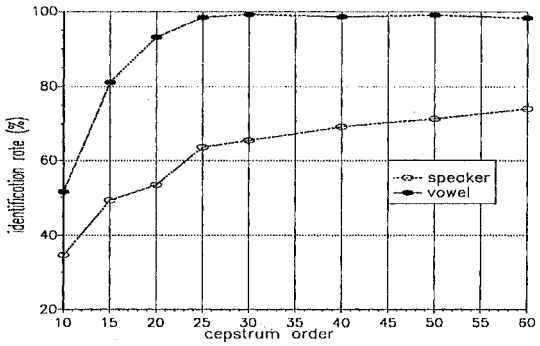

Figure 2: Relationship between number of FFT cepstrums and speaker and vowel identification rates.

4 Experiment 3

4.1 The analysis of spectral envelopes

To identify the frequency bands having speaker individualities in the spectral envelope, the variance for the five vowel spectral envelopes of ten male speakers from the ATR speech databases are calculated. The spectral envelopes are smoothed with 60 FFT cepstrums and the frequency axis is converted into the ERB rate.

Let $E_{ij}(n)$ $(n = 1 \sim N)$ be the i $(i = 1 \sim I)$th speaker spectral envelope of the j $(j = 1 \sim J)$th vowel for an n ERB rate. The variance of $E_{ij}(n)$ with respect to i is

$$\sigma_j^2(n) = \frac{1}{I}\sum_{i=1}^{I}\{E_{ij}(n) - \mu_j(n)\}^2 \qquad (1)$$

where $\mu_j(n)$ is the average of the spectral envelopes of the jth vowel. The frequency bands having quantities of $\sigma_j^2(n)$ are regarded to provide speaker individualities relatively.

The variance of $E_{ij}(n)$ with respect to j is

$$\sigma_i^2(n) = \frac{1}{J}\sum_{j=1}^{J}\{E_{ij}(n) - \mu_i(n)\}^2 \qquad (2)$$

where $\mu_i(n)$ is the average of the spectral envelopes of the ith speaker. The frequency bands having quantities of $\sigma_i^2(n)$ are regarded to provide vowel characteristics relatively.

The variances $\sigma_j^2(n)$ and $\sigma_i^2(n)$ shown in Figure 3 indicate that speaker individualities exist mainly above the 22 ERB rate (2212 Hz[5]) and that vowel characteristics will mainly exist from 12 ERB rate (603 Hz) to 22 ERB rate.

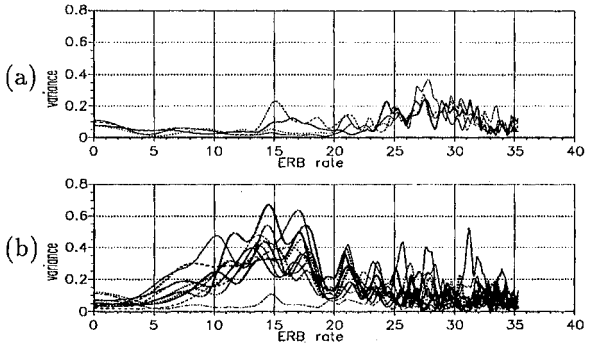

Figure 3: Variance of E_{ij} (a) $\sigma_j^2(n)$ and (b) $\sigma_i^2(n)$.

4.2 Experiment

The results of the above analyses suggest that speaker individualities exist mainly above 22 ERB rate and 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 third experiment is designed to test this assumption from a psychoacostic side.

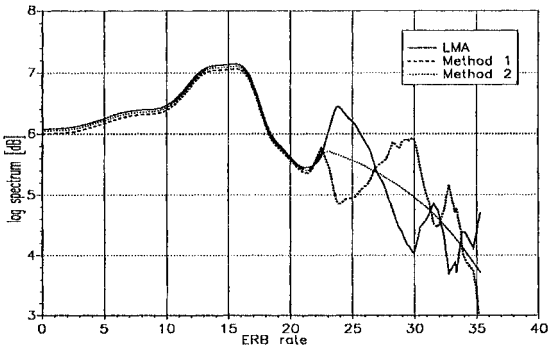

Figure 4: Spectral envelopes varied by Method 1 and 2 above 22 ERB rate. 60 FFT cepstrums are used to make these spectrum envelopes.

4.2.1 Method

Stimuli. The spectral envelopes of the 1d of Experiment 1 are varied by using the LMA analysis-synthesis system. An LMA filter with 60 FFT cepstrums is adopted to make the stimuli. Two varying methods are adopted. The spectral envelopes are reversed symmetrically with respect to their autoregressive line (Method 1) and the spectral envelopes are replaced with their autoregressive line (Method 2). Figure 4 shows spectral envelopes varied with by the two methods above 22 ERB rate. The types of stimuli are as follows.

3a. LMA analyzed-synthesized speech waves

3b. speech waves varied by Method 1 from 12 to 22 ERB rate

3c. speech waves varied by Method 1 above 22 ERB rate

3d. speech waves varied by Method 2 above 22 ERB rate

4.3 Results and Discussion

The speaker identification rates and the vowel identification rates for Experiment 3 are shown in Figure 5. The results suggest the following four conclusions.

1. 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 3a and 3c for the vowel identification rate, $F(1,14)=88.90$ between 3a and 3c for the speaker identification rate).

2. Distortion of the spectral envelopes from 12 to 22 ERB rate affects vowel identification ($F(1,14)=342.85$ between 3a and 3b for the vowel identification rate). The distortion affects speaker identification rates less than the distortion of the spectral envelopes above 22 ERB rate ($F(1,14)=11.84$ between 3b and 3c for the speaker identification rate).

3. The vowel identification rate is lower than the speaker identification rate for 3b ($F(1,14)=12.04$ between the vowel identification rate and the speaker identification of 3b).

4. Method 1 affects speaker identification rates more than Method 2 ($F(1,14)=14.32$ between 3c and 3d for the speaker identification rate).

Results 1 and 2 suggest that speaker individualities and vowel characteristics can be controlled. If the frequency band of the spectral envelopes above 22 ERB rate can be manipulated, speaker normalization and speaker adaptation methods can be constructed that do not affect vowel recognition rates and that may improve speaker-independent speech recognition performance. Result 3 shows that humans can identify speakers even when listening to unidentified vowels. Result 4 suggests that the relationship between the local maxima and the minima in the spectral envelopes is important in identifying speakers.

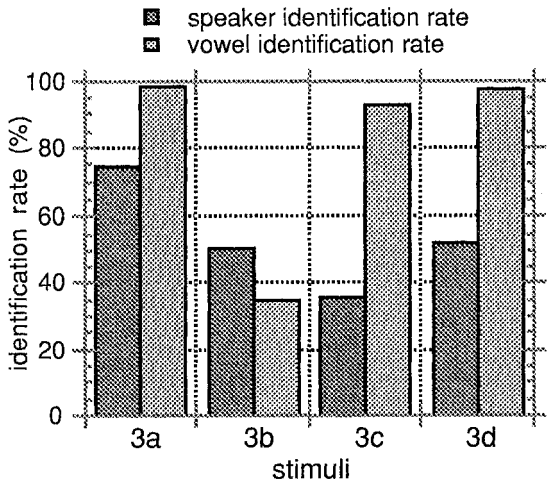

Figure 5: Speaker and vowel identification rates for Experiment 3.

5 Experiment 4

5.1 Method

Stimuli. Two speakers of five speakers are chosen and their five vowels are used. The higher frequency band in the spectral envelopes of one talker is replaced with that of the other talker (Figure 6). The boundary of replacement is varied from 21 to 25 ERB rate. An LMA filter with 60 FFT cepstrums is adopted to synthesize the stimuli.

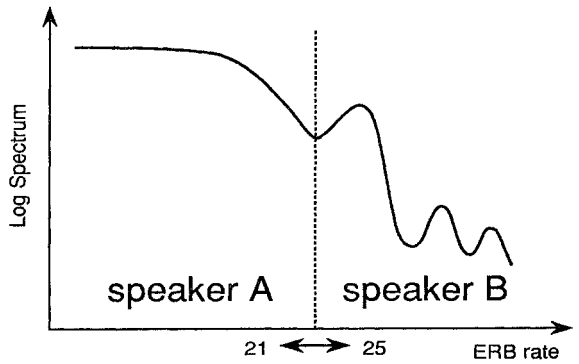

Figure 6: Spectral envelopes of speaker A replaced with those of speaker B in higher frequency band.

5.2 Results and Discussion

The speaker identification rates of Experiment 4 (the averages of 8 subjects responses) are shown in Figure 7. The average vowel identification rate is 98.08%. The results show that the boundary changing the identified speakers is at 23.5 ERB rate (2640 Hz). The boundaries between subjects exist from 22 to 25 ERB rate; accordingly the cues for speaker identification are different for each subject. However, the result suggests that the voice quality can be controlled without influencing vowel recognition if the spectral envelope above 23.5 ERB rate is replaced with that of other speakers.

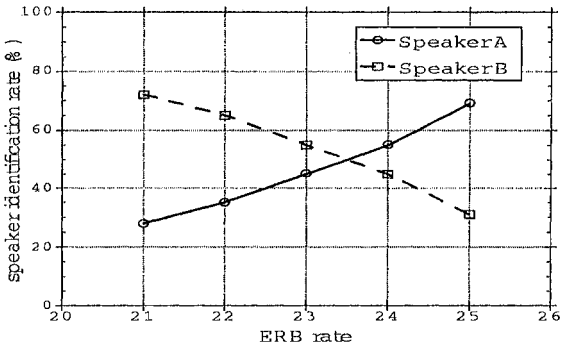

Figure 7: Speaker identification rate for Experiment 4.

6 General Discussion

These four experiments show that speaker individualities exist mainly above 23.5 ERB rate of the spectral envelopes and that they can be controlled without influencing vowel identification. The results suggest that speaker normalization, speaker adaptation and voice quality control can be constructed without affecting the vowel recognition .

Additionally, more detailed information on the spectral envelopes is required for speaker identification than for vowel identification. If FFT cepstrums are used in a speaker recognition system, more than 25 FFT cepstrums must be used at sampling frequency 20 kHz, while 20 FFT cepstrums are adequate for a vowel recognition system.

References

[1] Itoh, K. and Saito, S., "Effects of Acoustical Feature Parameters of Speech on Perceptual Identification of Speaker", IEICE, '82/1 Vol.J65-A No.1

[2] Kuwabara, H. and Ohgushi, K., "The Role of Formant Frequencies and Bandwidths in the Perception of Speaker", IEICE, '86/4 Vol.J69-A No.4

[3] Imai, S. and Kitamura, T., "Speech Analysis Synthesis System Using the Log Magnitude Approximation Filter ", IEICE, '78/6 Vol.J61-A No.6

[4] Imai, S., "Log Magnitude Approximation (LMA) Filter ", IEICE, '80/12 Vol.J63-A No.12

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

[6] Furui, S. and Akagi, M., "Perception of voice individuality and physical correlates", JASJ, H85-18 (1985)