Acoust. Sci. & Tech. **28**, 6 (2007)

ACOUSTICAL LETTER

Effects of acoustic modification on perception of speaker characteristics for sustained vowels

Tatsuya Kitamura[*,†] and Takeshi Saitou[‡,§]

ATR Cognitive Information Science Laboratories,
2–2–2 Hikaridai, "Keihanna Science City," Kyoto, 619–0288 Japan

(*Received 12 March 2007, Accepted for publication 15 May 2007*)

Keywords: Speaker individuality, Interval scale, Glottal source, Speech spectra
PACS number: 43.71.Bp [doi:10.1250/ast.28.434]

1. Introduction

Humans can identify speakers of speech sounds even though the acoustic properties of speech sounds vary considerably within a speaker; that is, our ability to identify speakers exhibits robustness against intraspeaker variations. If the perceptual cues used for speaker identification by adapting to intraspeaker variation are clarified, these acoustic parameters will facilitate the development of advanced speech signal processing techniques. In the present study, we therefore aim to explore such cues on the basis of the hypothesis that humans perceive speaker characteristics by focusing on less arbitrary and more invariant acoustic parameters.

Intraspeaker variation can appear in many acoustic properties including the pitch frequency, power, speaking rate, and spectra, and the variations of these acoustic properties have been studied from the viewpoint of the effects of the emotion and speaking style of the speaker. In general, it is regarded that the pitch frequency and power are more arbitrary and controllable, while the frequency properties of the glottal source and spectra in higher-frequency regions are less arbitrary and controllable.

Previous studies have revealed the effects of the acoustic modification of speech sounds on the perception of speaker characteristics [1–3]. In these studies, the authors mainly investigated the effects of the modification of several acoustic properties on speaker identification rate; however, if the perceptual contributions of the acoustic properties were mapped on a scale, they would give a more clear-cut model of the perception of speaker individuality.

In the present study, psychoacoustic experiments were thus carried out to obtain interval scales for the contribution to the perception of speaker individuality of the following acoustic parameters: the time pattern of the amplitude and pitch frequency, the mean of the pitch frequency, the frequency characteristics of the glottal source, and the spectra in higher-frequency regions. According to the hypothesis mentioned above, it is expected that the contributions of the last two parameters will be larger than the others.

The acoustic modification of stimuli used in psychoacoustic experiments sometimes gives rise to the degradation of sound quality; however, the degradation was not considered sufficiently in previous studies. Thus, the sound quality of the stimuli was also evaluated subjectively to assess whether the degradation affects experimental results.

*Currently with Konan University
†e-mail: t-kitamu@konan-u.ac.jp
‡Currently with Advanced Industrial Science and Technology
§e-mail: saitou-t@aist.go.jp

2. Experiment 1

Experiment 1 was carried out to obtain an interval scale of the acoustic properties for the similarity of speaker characteristics. A sustained vowel /a/ uttered by ten male speakers was used, and the experiment was conducted in accordance with the Thurstone paired-comparison methodology [4,5].

2.1. Method

2.1.1. Stimuli

Stimuli were made from a sustained vowel /a/ of ten male native Japanese speakers. To avoid the effect of the pitch frequency and duration of the stimuli on the experiment, the speakers were asked to tune their pitch frequency to the same pitch and keep the duration of their voices to the same length as that of a 0.7 s harmonic complex tone presented through headphones. Prior to recording, the speakers rehearsed by listening to the harmonic complex tone, and thereafter, vowels uttered without listening to the tone were recorded. The fundamental frequency of the harmonic complex tone was 123 Hz or 124 Hz (the difference in the fundamental frequencies was unintentionally caused by a mistake).

The sustained vowels were recorded at a sampling rate of 48 kHz with 24-bit resolution using a microphone (Audio-Technica AT4041) and a solid-state recorder (Marantz PMD671). The data was downsampled to 16 kHz and converted to 16-bit resolution using a personal computer. Three tokens were used in the experiment. The following eight types of stimuli were used:

- **A** speech waves with normalized maximum amplitude,
- **B** speech waves with normalized time pattern of amplitude; in addition, the pitch frequency was fixed to the speakers' mean value,
- $\mathbf{C_{0.9}}$ speech waves with a pitch frequency 0.9-fold that of stimulus B,
- $\mathbf{C_{1.1}}$ speech waves with a pitch frequency 1.1-fold that of stimulus B,
- $\mathbf{D_{1.0}}$ speech waves with STRAIGHT cepstra beyond the 35th order fixed to zero that of stimulus B,
- $\mathbf{D_{0.9}}$ speech waves with a first-order STRAIGHT cepstrum

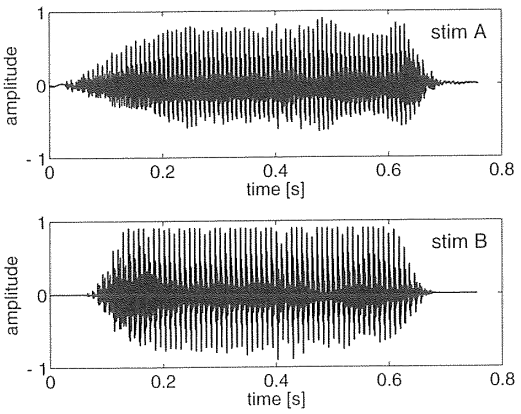

Fig. 1 Waveforms of stimuli A (upper panel) and B (lower panel) of a speaker.

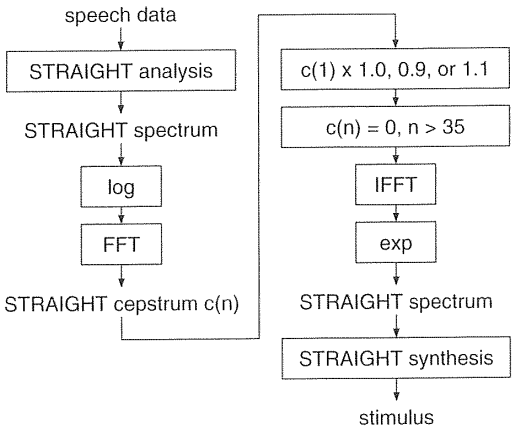

Fig. 2 Procedure for resynthesizing stimuli $D_{1.0}$, $D_{0.9}$ and $D_{1.1}$.

0.9-fold that of stimulus $D_{1.0}$,

$D_{1.1}$ speech waves with a first-order STRAIGHT cepstrum 1.1-fold that of stimulus $D_{1.0}$,

E speech waves with logarithmic STRAIGHT spectra beyond 2.6 kHz replaced by the autoregressive lines of stimulus B.

Stimuli B, $C_{0.9}$, $C_{1.1}$, $D_{1.0}$, $D_{0.9}$, $D_{1.1}$, and E were synthesized using the STRAIGHT analysis-synthesis system [6]. A "STRAIGHT spectrum" is a spectrum calculated using the system and a "STRAIGHT cepstrum" is a cepstrum calculated from a STRAIGHT spectrum. Hereafter, stimuli $C_{0.9}$ and $C_{1.1}$ are referred to together as C_*, and stimuli $D_{1.0}$, $D_{0.9}$, and $D_{1.1}$ are referred to together as D_*.

Stimulus B is used to investigate the effects of the time pattern of the amplitude and pitch frequency. First, a speech wave was resynthesized after fixing the pitch frequency of the voiced frames to each speaker's mean value. The maximum amplitude of every 15 ms voiced frame of the resynthesized speech wave was then normalized. Lastly, the amplitudes of the onset and offset of the voiced section were weighted by a $\cos^2$ function. The waveforms of stimuli A and B of a speaker are shown in Fig. 1.

Stimuli C_* are used to examine the effects of the mean of the pitch frequency. The pitch frequencies of stimuli $C_{0.9}$ and $C_{1.1}$ are fixed to 0.9-fold and 1.1-fold the mean value of each speaker, respectively.

Stimuli D_* are used to clarify the effects of the frequency characteristics of the glottal source, on the basis of the concept of homomorphic filtering [7]. The procedure for resynthesizing stimuli D_* consists of the calculation and liftering of STRAIGHT cepstra, and the inversion of STRAIGHT spectra followed by resynthesis, as shown in Fig. 2. A STRAIGHT cepstrum was calculated by taking the inverse Fourier transform of a logarithmic STRAIGHT spectrum. For stimuli D_*, STRAIGHT cepstra beyond the 35th order were fixed to zero to eliminate glottal source characteristics. Additionally, the first-order STRAIGHT cepstrum was multiplied by 0.9 (high-frequency emphasis) and 1.1 (high-frequency deemphasis) for stimuli $D_{0.9}$ and $D_{1.1}$, respectively, to vary the frequency properties of the glottal source.

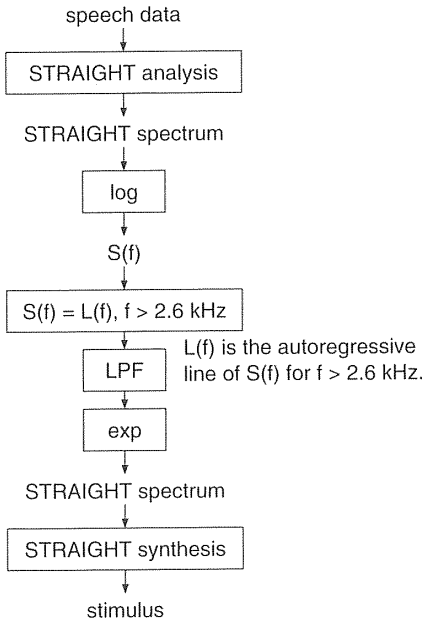

Fig. 3 Procedure for resynthesizing stimulus E.

Stimulus E was resynthesized after replacing a logarithmic STRAIGHT spectrum beyond 2.6 kHz with its autoregressive line to investigate the effects of the spectrum for frequencies above 2.6 kHz. Figure 3 illustrates the procedure for resynthesizing the stimulus. Frequencies above 2.6 kHz were determined so that the frequencies include the third formant of the five Japanese vowels /a/, /e/, /i/, /o/, and /ɯ/ of the speakers. Because the replacement can cause a discontinuity at 2.6 kHz, each logarithmic STRAIGHT spectrum was low-pass filtered after replacement. Figure 4 shows the FFT spectra of stimuli B, $D_{1.0}$, $D_{0.9}$, $D_{1.1}$, and E.

2.1.2. Participants

Sixteen female listeners participated in experiments 1 and 2. They had never met the speakers or listened to their voices. None of the participants had hearing impairments.

2.1.3. Procedure

In the experiments, the participants randomly listened to triplets of stimuli (S1, S2, and S3) of a speaker at intervals of

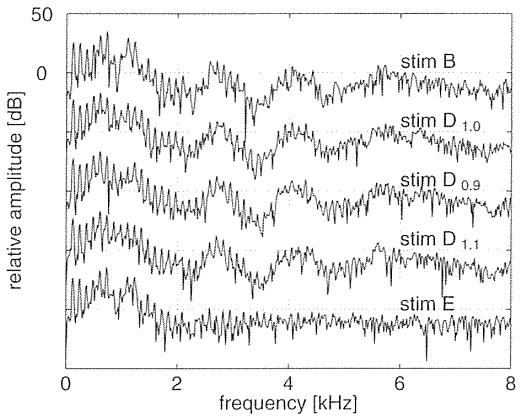

Fig. 4 FFT spectra of stimuli B, $D_{1.0}$, $D_{0.9}$, $D_{1.1}$, and E.

500 m. The first stimulus (S1) of a triplet was stimulus A, and the second and third ones (S2 and S3) were two of stimuli B, C_*, D_*, and E. Different tokens were used for these three stimuli. The stimuli were also presented in the order S1, S3, S2 to counterbalance any effects due to the order of presentation. The total number of trials was 420 ($= {}_7P_2 \times 10$ speakers). The stimuli were presented through binaural earphones (Sennheizer HDA200) at a comfortable loudness level. The listeners were not allowed to listen to each triplet more than once.

The listeners were asked to select which of the last two stimuli had speaker characteristics closer to the first one. We reminded the listeners to judge the stimuli in terms of speaker characteristics, and not by pitch height or sound quality. To confirm the criterion of judgment, the participants had a rehearsal session with ten triplets prior to each actual experiment.

2.2. Results

An interval scale for the similarity of speaker characteristics of the stimuli, obtained by adopting Thurstone case V, is shown in Fig. 5. This result is consistent with the model of the Thurstone paired-comparison methodology ($\chi^2 = 0.712 < \chi^2(15, 0.05) = 24.995$). The value for stimulus B is 0.92, that for stimulus $C_{0.9}$ is 0.14, that for stimulus $C_{1.1}$ is 0.27, that for stimulus $D_{1.0}$ is -0.26, that for stimulus $D_{0.9}$ is -0.26, that for stimulus $D_{1.1}$ is -0.37, and that for stimulus E is -0.45. The experimental results indicate that the closest speaker characteristics to stimulus A are B, C_*, D_*, and E, in that order.

Interval scales for each speaker show that there are four different orders of the similarity of speaker characteristics:

(1) $B > C_* > D_* > E$ for five speakers,
(2) $B > C_* > E > D_*$ for two speakers,
(3) $B > C_* >$ (D_* and E) for two speakers,
(4) $B > C_{1.1} > D_* > C_{0.9} > E$ for one speaker.

There is no clear difference in value between D_* and E in the third case. These results demonstrate that the perceptual contributions of the frequency properties of the glottal source and speech spectra in the higher-frequency regions to perceived speaker characteristics are not the same for all the speakers.

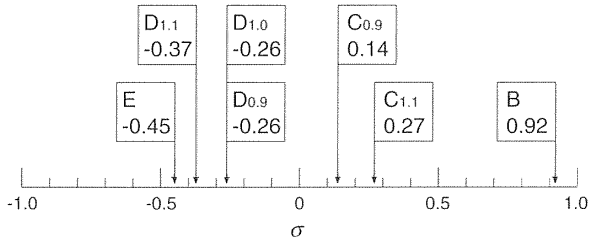

Fig. 5 Interval scale for similarity of speaker characteristics of the stimuli.

3. Experiment 2

There were some stimuli for which the sound quality was degraded by the resynthesis process, and the degradation could have affected the results of experiment 1. A subjective evaluation of the sound quality of the stimuli was thus conducted in accordance with the Thurstone paired-comparison methodology.

3.1. Method

3.1.1. Stimuli

Stimuli B, C_*, D_*, and E were used in this experiment.

3.1.2. Procedure

The participants randomly listened to pairs of stimuli of a speaker at intervals of 500 m. These stimuli were two of stimuli B, C_*, D_*, and E. Different tokens were used for the two stimuli. The total number of pairs was 420. The listeners were asked to select which of the two stimuli had better sound quality.

3.2. Results

Figure 6 shows an interval scale for the sound quality of the stimuli obtained by adopting Thurstone case V. This result is consistent with the model of the Thurstone paired-comparison methodology ($\chi^2 = 0.634 < \chi^2(15, 0.05) = 24.995$). The value for stimulus B is 0.68, that for stimulus $C_{0.9}$ is 0.21, that for stimulus $C_{1.1}$ is 1.04, that for stimulus $D_{1.0}$ is -0.50, that for stimulus $D_{0.9}$ is -0.51, that for stimulus $D_{1.1}$ is -0.55, and that for stimulus E is -0.37. Pearson's correlation coefficient between the interval scale for the similarity of speaker characteristics (Fig. 5) and that for sound quality (Fig. 6) is 0.83, indicating that there is a strong positive correlation between them.

4. Discussion

The results of experiment 1 (Fig. 5) indicate that the perceptual contribution of the acoustic properties to the similarity of speaker characteristics decreases in the following order under the experimental conditions:

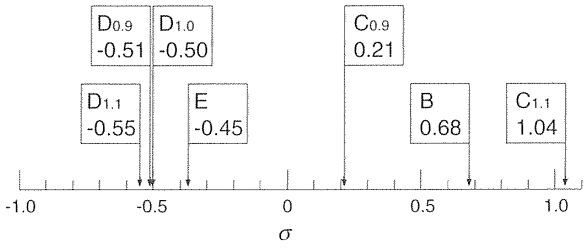

Fig. 6 Interval scale for sound quality of the stimuli.

(1) spectra in the higher-frequency regions,
(2) frequency properties of the glottal source,
(3) mean of the pitch frequency,
(4) time pattern of the amplitude and pitch frequency.

The order of this list suggests that the higher the acoustic property in the list, the more dependent it is on the innate characteristics of the speech organs, and thus, the less controllable it is by the speaker. The shape of the hypopharyngeal cavity, for example, is relatively stable during vowel production, and the interspeaker variation of the cavity affects spectra in the frequency range beyond approximately 2.5 kHz [8]. Therefore, it is probable that humans perceive speaker individuality by focusing on more invariant acoustic properties, as hypothesized in the introduction.

Experiment 1 also revealed that the amount of the perceptual contribution of each of the acoustic properties changes from speaker to speaker. These results support those of Lavner *et al.* [3], who demonstrated that information regarding speaker individuality is not coded for all speakers in the same way by conducting psychoacoustic experiments using the vowel /a/ of twenty speakers.

The contribution of the frequency characteristics of the glottal source was relatively large for judging the similarity of speaker characteristics in experiment 1, whereas previous studies [1,3] revealed that its contribution is relatively small. This contrast might be due to the difference of the task in the experiments; the listeners in the present study were asked to assess the similarity of speaker characteristics, while a naming task and an ABX test were employed in the two previous studies.

The results of experiment 2 imply that the results of experiment 1 are possibly affected by the degradation of the sound quality of the stimuli. There was a strong positive correlation between the interval scales of the similarity of speaker characteristics (Fig. 5) and sound quality (Fig. 6). We thus cannot deny the possibility that sound quality may have been a cue in experiment 1. However, because the order of stimuli B and $C_{1.1}$ and that of stimuli D_* and E are reversed on the two interval scales (Figs. 5 and 6), it is probable that sound quality was not the only cue for the judgment in experiment 1.

The constant values 0.9 and 1.1, which were used in resynthesizing stimuli C_* and D_*, were selected for expedience, and the order of the stimuli on the interval scale could change if other values were used. In addition, the perceptual sensitivity to the acoustic properties may also change. Therefore, the amount of the perceptual contribution of each of the acoustic properties to speaker individuality cannot be determined only from the results of the present study.

5. Conclusions

To investigate the perceptual contributions of various acoustic properties to speaker individuality, an interval scale was obtained to examine the hypothesis that humans perceive speaker characteristics by focusing on more invariant acoustic properties. The interval scale obtained under the experimental conditions revealed that less arbitrary acoustic properties were more important for judging the similarity of speaker characteristics, and the hypothesis was verified. However, there was a strong positive correlation between the interval scales of the similarity of speaker characteristics and sound quality implying that sound quality may have been a cue in the experiment.

Acknowledgments

This research was supported by the Ministry of Internal Affairs and Communications as part of their Strategic Information and Communications R&D Programme (SCOPE). We thank Drs. Akemi Iida and Yuichi Ishimoto of Tokyo University of Technology for their help in recording speech data.

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).
[2] M. Hashimoto, S. Kitagawa and N. Higuchi, "Quantitative analysis of acoustic features affecting speaker identification," *J. Acoust. Soc. Jpn. (J)*, **54**, 169–178 (1998).
[3] Y. Lavner, I. Gath and J. Rosenhouse, "The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels," *Speech Commun.*, **30**, 9–26 (2000).
[4] L. L. Thurstone, "Psychophysical analysis," *Am. J. Psychol.*, **38**, 386–389 (1927).
[5] L. L. Thurstone, "A law of comparative judgment," *Psychol. Rev.*, **34**, 273–286 (1927).
[6] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, **27**, 187–207 (1999).
[7] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Am.*, **45**, 458–465 (1969).
[8] T. Kitamura, K. Honda and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoust. Sci. & Tech.*, **26**, 16–26 (2005).