Reversal of relationship between impression of voice pitch and height of fundamental frequency: Its appearance and disappearance

Teruhisa Uchida

The National Center for University Entrance Examinations, Research Division, 2–19–23 Komaba, Meguro-ku, Tokyo, 153–8501 Japan

(Received 5 June 2018, Accepted for publication 4 January 2019)

Abstract: This study investigated the cognitive biases related to the impression of voice pitch caused by changes in tonal quality. According to the vocal tube model, changing the vocal-tract length (VTL) systematically alters the tonal quality. In one experiment, the fundamental frequency ($f_o$) of the speech samples was raised and lowered on a mel-scale axis. Then the spectral-frequency scale was expanded and contracted to simulate reducing and increasing the VTL. In a second experiment, the width of the $f_o$ range was changed in addition to changing the $f_o$ height and VTL scaling. Noise-vocoded speech samples were generated to measure the independent effects of the VTL scaling. The participants rated their impressions of the pitch using paired comparison. The results revealed a reversal of the relationship between impression of voice pitch and height of $f_o$ when the effects of $f_o$ height and VTL scaling on pitch impression were opposite to each other and when the range of the $f_o$ contour was equivalent to that of natural speech. VTL scaling played a dominant role in this reversal. However, as the $f_o$ contour became flat, this reversal phenomenon disappeared, and the $f_o$ height factor came to play the dominant role.

Keywords: Pitch, Tonal quality, Vocal tract length, Spectral centroid, WORLD

PACS number: 43.66.Hg, 43.66Lj, 43.72.Ja [doi:10.1250/ast.40.198]

1. INTRODUCTION

This study investigated cognitive biases regarding the effect of changes in the tonal quality of speech on the impression of voice pitch. The aim was to clarify whether the biases are strong enough to reverse the relationship between the height of the fundamental frequency ($f_o$) and the listener’s impression of voice pitch.

According to the vocal tube model, changing the vocal-tract length (VTL) systematically alters the tonal quality of speech. This manipulation is accomplished by expanding or contracting the spectral envelope of the speech. Therefore, changing the spectral frequency axis represents actual manipulation of the VTL [1,2].

In previous studies, higher pitched voices were stably perceived in speech converted by reducing the VTL even though the $f_o$ patterns remained the same [3,4]. This indicates that changes in voice quality related to VTL manipulation can introduce cognitive bias into the impression of voice pitch [4].

In contrast, the $f_o$ of a singing voice plays a dominant role in the impression of musical pitch height despite the fact that it is also a human voice. The subjective pitch corresponds to the $f_o$ height regardless of voice quality.

This study measured the strength of cognitive biases in relation to pitch impression correlated with the change in VTL. The effects of changing $f_o$ and VTL were evaluated by systematically converting the speech stimuli used for paired comparison in a subjective evaluation task.

One experiment, “Experiment 1,” was conducted to determine whether cognitive biases can reverse the relationship between the height of the $f_o$ pattern and the impression of the voice pitch. Another experiment, “Experiment 2,” was conducted to evaluate the effects of $f_o$ fluctuation on the time axis, which corresponds to the width of the $f_o$ range, and to examine whether the pitch impression corresponds to the $f_o$ height against the effect of VTL scaling, especially for a flat $f_o$ contour because a flat $f_o$ contour might play a dominant role in pitch cognition.

2. EXPERIMENT 1

Experiment 1 was designed to measure the effects of $f_o$ height and VTL scaling on the impression of voice pitch by using paired comparison.
2.1. Generation of Experimental Speech Samples

2.1.1. Original speech data

Monologue speech samples uttered by eight women and eight men were selected from the JNAS speech corpus [5,6]. The details are listed in Table 1.

2.1.2. Transformation of fundamental frequency ($f_o$) and vocal tract length (VTL)

The $f_o$ and VTL of the original speech samples were converted to obtain experimental speech stimuli.

First, their $f_o$ patterns were converted by raising or lowering their $f_o$ values on the psychophysical mel-scale axis [7,8]. The expression for converting frequency $f$ (Hz) into mel value $m$ is given by

$$m = 2.595 \log_{10}(1 + f/700).$$  \hspace{1cm} (1)

Second, their spectral-frequency scales were expanded and contracted to simulate reducing and increasing the VTL.

To generate experimental speech stimuli, the WORLD vocoder was mainly used to manipulate $f_o$ and the spectral envelope and to synthesize the converted speech stimuli [9,10]. The TANDEM-STRAIGHT vocoder was partially used in the first $f_o$ extraction step [11–13].

The WORLD and TANDEM-STRAIGHT speech analysis/synthesis systems have been used in various applications, such as voice conversion and statistical parametric speech synthesis, that use a high-quality vocoder-based system to accurately decompose the speech waveform into the $f_o$ spectral envelope and the aperiodicity. They are able to synthesize new voices integrating the transformed $f_o$, spectral envelope or aperiodicity.

2.1.3. Manipulation of $f_o$ height

The $f_o$ patterns of the original speech samples were shifted along the mel-scale axis by a fixed amount. Three $f_o$ shift conditions were applied: 1) original $f_o$, 2) higher $f_o$, and 3) lower $f_o$. The range of the shift amount ranged from small (+5 mel to −5 mel) to large (+20 mel to −20 mel). Within these ranges, the conversion steps were ±5, ±10, and ±20 mel, as shown in Table 1.

After $f_o$ was transformed on the mel scale, the mel values (mel) were converted into physical frequencies (Hz). The expression for converting mel value $m$ into frequency $f$ (Hz) is given by

$$f = 700(10^{m/2.595} - 1).$$ \hspace{1cm} (2)

Figure 1 shows an example of $f_o$ manipulation for an experimental speech stimulus pair. The lower graphs illustrate $f_o$ patterns of converted speech stimuli.

2.1.4. Manipulation of VTL

The original speech samples were transformed into both short-VTL speech samples and long-VTL speech samples. The VTL targets for the female speech samples were 12.96 cm and 15.1 cm. Those for the male speech samples were 15.1 cm and 17.6 cm. The corresponding intervals are equal on a logarithmic scale. In a previous study, the effects of VTL conversion on speaking naturalness were examined. The VTL was varied from 12 cm to 19 cm [3]. The target VTLs here were selected on the basis of the results of that study [4].

Kawahara et al. developed an algorithm that makes it possible to estimate the VTL from naturally spoken voice [14]. It is based on a database of Japanese vowels with

### Table 1

| Original ID | Age | Duration (s) | $f_o$ shift (mel) |
|-------------|-----|--------------|-------------------|
| F1: NF005072 | 30–39 | 10.2 | +20, 0, −20 |
| F2: NF008052 | 20–29 | 10.2 | +15, 0, −15 |
| F3: NF017058 | 20–29 | 10.8 | +10, 0, −10 |
| F4: NF018051 | 40–49 | 10.6 | +5, 0, −5 |
| F5: NF023053 | 50–59 | 10.2 | +20, 0, −20 |
| F6: NF027086 | 30–39 | 10.0 | +15, 0, −15 |
| F7: NF028089 | 30–39 | 10.5 | +10, 0, −10 |
| F8: NF051088 | 30–39 | 10.5 | +5, 0, −5 |

| M1: NM007014 | 30–39 | 9.8 | +5, 0, −5 |
| M2: NM010090 | 20–29 | 10.7 | +10, 0, −10 |
| M3: NM026056 | 20–29 | 10.4 | +15, 0, −15 |
| M4: NM042090 | 20–29 | 10.2 | +20, 0, −20 |
| M5: NM071090 | 30–39 | 9.7 | +5, 0, −5 |
| M6: NM084089 | 40–49 | 10.2 | +10, 0, −10 |
| M7: NM086013 | 20–29 | 10.7 | +15, 0, −15 |
| M8: NM092056 | 30–39 | 10.4 | +20, 0, −20 |

Sampling frequency: 16 kHz, Resolution: 16 bit linear

**Fig. 1** Example spectra and $f_o$ patterns of converted speech stimuli (original M8 sample: “karekara no tegami niwa”); $f_o$ raised or lowered along mel-scale axis in accordance with three experimental conditions: 1) original $f_o$, 2) higher $f_o$, and 3) lower $f_o$. VTL of original sample was converted into two target VTLs by expanding or contracting spectral-frequency scale.
MRI-based vocal tract shape information. Application of this algorithm to the current voice conversion transformed the original samples into experimental stimuli that matched the two VTL targets. The stimuli were practically converted by expanding and contracting the spectral-frequency scale and using a suitable estimated ratio to obtain the target VTLs. The spectral frequency band ranged up to 7 kHz for all experimental stimuli.

Figure 1 shows example spectra and $f_0$ patterns of converted speech stimuli. Those on the left figure represent expansion of the spectral-frequency scale, resulting in reduction of the VTL and movement of the overall formant toward a higher frequency than for those on the right, which represent contraction of the scale. This means that the left side represents shortening of the vocal tube while the right side represents lengthening of the vocal tube.

2.1.5. Assignment of comparison pairs

For each original sample, six experimental stimuli were prepared by using various combinations of the three $f_0$ shifts and two target VTLs. A total of 96 stimuli were prepared. Figure 2 illustrates the distribution of experimental stimuli used in the paired comparison task. Each stimulus pair was defined by the difference in $f_0$ ($f_0$ diff) and the difference in VTL (VTL diff). These two parameters express the relative relationships of all pairs.

Five comparison pairs were selected from the six experimental speech stimuli for each original sample. One pair had the same $f_0$ (no shift) and different VTLs, two pairs had a different $f_0$ and the same VTL, and the last two pairs had a different $f_0$ and different VTLs. A total of 80 stimulus pairs were used for the experiment.

2.2. Method

2.2.1. Participants

A total of 141 first-year university students, 110 men and 31 women with unimpaired hearing, participated in the experiment. Their ages ranged from 18 to 22. They were divided into ten groups, and eight unique experimental stimulus pairs were assigned to each group.

2.2.2. Evaluation items

The adjectives used for expressing the participant’s subjective impressions of three traits are shown in Table 2. There were ten for voice quality, including one for evaluating voice pitch (“high”) [15,16], four for naturalness, and four for speaker body size [3,4].

2.2.3. Procedure

For each stimulus pair, the first stimulus presented was called stimulus A, and the second one was called stimulus B. Figure 2 shows an example of the presentation orders (A–B) using the stimuli based on the M8 original sample. The stimulus pairs generated from the original M4 sample with the same $f_0$ shift as that for the M8 sample were played in the opposite order. The presentation orders were thereby counter-balanced among pairs which had same gender and same $f_0$ shift.

The participants listened to the stimuli in each pair one after the other through stereo earphones connected to a portable CD player (Sony: D-EJ002) at the most comfortable sound level.

They were asked to compare the two stimuli for each of the 18 adjectives in Table 1 and record on an answer sheet which stimulus better matched each adjective. They gave a rating score on a scale of 1 to 10 for each adjective. They were given 60 s for each comparison and were free to compare them in any order of adjectives.

| Table 2 Adjectives for expressing subjective impression |
|---------------------------------|----------------|
| Trait                           | Adjectives     |
| Voice Quality                   | High, Manly, Husky, Placid, Powerful, Young, Deep, Forceful, Nasal, Falsetto |
| Naturalness                     | Natural, Comprehensible, Clear, Listenable |
| Speaker Body Size               | Large, Tall, Fat, Robust |

Fig. 2 Upper graph illustrates estimated VTLs and $f_0$ shift of converted stimuli. Assignments of comparison pairs are shown on right for experimental stimuli based on original M8 sample as an example. Difference in $f_0$ ($f_0$ diff) and difference in VTL (VTL diff) between paired experimental stimuli are shown at top and bottom. Lower graph illustrates estimated VTLs of original stimuli. Overlapping marker points were vertically shifted.
2.3. Results

2.3.1. Evaluation indexes

For each stimulus pair, 12 evaluation indexes were calculated: ten indexes for voice quality were calculated from the ratings for each of the ten corresponding adjectives, and indexes for naturalness and speaker body size were calculated by averaging the ratings for the four corresponding adjectives for each trait.

Experiment 1 focused on the height of the $f_0$ pattern and the impression of the voice pitch. More specifically, rating scores (1–10) for the “high” items were converted into evaluation indexes (pitch scores), which ranged from −4.5 to 4.5. A positive value means that the participant evaluated stimulus A higher than stimulus B. A negative value means the opposite. The higher the absolute value, the greater the degree of conviction.

The 80 stimulus pairs were divided equally into ten sets, and each set was copied to a CD (i.e., eight pairs per CD). The participants were also divided into 10 groups, with 12–15 participants in each group. For each stimulus pair, the participants in each group rated the stimuli against the 18 adjectives. Then 12 evaluation indexes were calculated for each participant for each stimulus pair from the 18 adjectives. Then 12 evaluation indexes were calculated: ten indexes for voice quality were calculated for each participant for each stimulus pair from the 12–15 participants in each group. For each stimulus pair, the participants in each group rated the stimuli against 12–15 participants in each group. For each stimulus pair, the participants in each group rated the stimuli against each stimulus pair from his or her ratings. The mean values of these personal evaluation indexes were then calculated and defined as the final integrated evaluation indexes for each stimulus pair.

2.3.2. Additive effects of $f_0$ diff and VTL diff

Figure 3 illustrates the evaluation indexes for voice pitch (pitch scores) on the $f_0$ diff − VTL diff plane by speaker gender. VTL diff was zero for the middle series of pitch scores; in other words, each stimulus in the pair had the same VTL. The effect of $f_0$ diff on this series was to linearly increase the pitch score. Naturally, a higher $f_0$ resulted in the perception of a higher pitch.

In the left series of pitch scores, the VTL of stimulus A was lower than that of stimulus B, and the pitch scores were much higher than for the middle series. The distributions for the left and middle series ran parallel to the $f_0$ diff − VTL diff plane while the left series had a higher pitch impression rating. VTL diff could have additionally affected the pitch scores because the slope of the increase caused by $f_0$ diff for the middle series was similar to that for the left series. The shorter VTL led the participants to perceive a higher pitch.

In the right series of pitch scores, the VTL of stimulus A was longer than that of stimulus B, resulting in the perception of a much lower pitch than for the middle series. The longer VTL led the participants to perceive a lower pitch.

2.3.3. Reversal of relationship between pitch impression and $f_0$ height

In the front two quadrants of the $f_0$ diff − VTL diff plane in Fig. 3, the $f_0$ of stimulus A was lower than that of stimulus B. Nevertheless, the pitch of stimulus A was perceived to be higher than that of stimulus B, shown in the front left quadrant.

As discussed in the previous section, when $f_0$ was raised, higher pitches were perceived, and when VTL was increased, lower pitches were perceived. This is illustrated in Fig. 4. There were 16 stimulus pairs in which there was a conflict between the direction of the $f_0$ effect and the direction of the VTL effect on pitch. The perceived result was opposite to what was expected from $f_0$ diff for 15 of these pairs. This result revealed a reversed relationship between the height of the fundamental frequency and the impression of the voice pitch.

2.3.4. Cognitive biases caused by conversion of VTL

The pitch scores were additionally affected by $f_0$ diff and VTL diff. Multiple regression analysis was used to predict the pitch score. In the model used, $f_0$ diff (mel) and VTL diff (mm) are independent variables, and the pitch score is a dependent variable. The $f_0$ diff and VTL diff variables were selected because the term for interaction between them was not significant in a preliminary analysis.

Model fitting was highly significant ($p < 0.0001$), and the adjusted $R^2$ was 0.87. The regression function computed from this analysis is

\[
\text{pitch score} = -0.077 \text{ VTL diff } + 0.031 f_0 \text{ diff } - 0.004.
\]

This regression equation shows that reducing VTL by 10 mm cognitively biases pitch impression and that the effect is the same as that caused by raising $f_0$ by 24.4 mel. The cognitive bias is enough to reverse the relationship between $f_0$ height and pitch impression.
2.3.5. Body size impression and conversion of VTL

Previous studies found that VTL affects the impression of speaker body size [1,2]. The evaluation indexes for speaker body size (body-size scores) were thus analyzed to verify the findings.

The body-size scores on the $f_0$ diff / VTL diff plane are illustrated in Fig. 5. Then Fig. 6 shows the scores by VTL ratios. Shorter VTLs created an impression of a smaller speaker while longer VTLs created an impression of a larger speaker (Fig. 5). Moreover, stimuli with a lower $f_0$ created an impression of a slightly large body (Fig. 6). However, the effect was relatively small. To summarize, VTL had the greatest effect on the body-size scores.

3. EXPERIMENT 2

As mentioned in the introduction, one purpose of Experiment 2 was to evaluate the effect of $f_0$ fluctuation on the time axis, which corresponds to the width of the $f_0$ range. The experiment was thus designed to measure the compensatory effect of the $f_0$ range on the cognitive biasing of the voice pitch impression. As also mentioned in the introduction, another purpose was to examine whether
the pitch impression corresponds to the $f_0$ height against the effects of VTL scaling, especially in a flat $f_0$ contour. The flat $f_0$ contour represents the most stable $f_0$ pattern and has the narrowest $f_0$ range. Since the $f_0$ of a singing voice is partially stable on the time axis, the flat $f_0$ contour might play a dominant role in pitch cognition. Therefore, the combination of the relationship between $f_0$ height and VTL scaling was controlled as an experimental factor.

### 3.1. Generation of Experimental Speech Samples

#### 3.1.1. Original speech data

The same 16 stimulus-pair dataset used in Experiment 1 was also used in Experiment 2. Table 3 shows the identification numbers of the speech samples and the assigned conditions. The $f_0$ and VTL of the original speech samples were converted to obtain experimental speech stimuli. The WORLD and TANDEM-STRAIGHT vocoders were used to transform the original samples into experimental stimuli.

#### 3.1.2. Manipulation of $f_0$ height and range

The $f_0$ patterns were converted in two ways. The first was shifting $f_0$, which raised or lowered the values of $f_0$ on the mel scale. The $f_0$ shift amount was within the range $+10$ mel (high) to $-10$ mel (low).

The second conversion was manipulating the $f_0$ range width. The standard deviation (SD) of the $f_0$ patterns was set to four levels, 40 mel, 20 mel, 2 mel, and $r.p.$ 20 mel ($r.p.$ is short for reversed-phased $f_0$ height contour). Conversion at SD 40 mel represents ordinary speech intonation on the whole. Conversion at SD 2 mel provides almost flat intonation.

Figure 7 shows sample two-way conversions of $f_0$ shift and $f_0$ range. The lower graphs illustrate $f_0$ patterns of converted speech stimuli.

#### 3.1.3. Manipulation of VTL

Three target VTLs, short, middle, and long, were set. The target VTLs for female speech stimuli were 12.96 cm, 13.33 cm, and 15.1 cm. Those for male speech stimuli were 15.1 cm, 16.3 cm, and 17.6 cm. The original speech samples were converted by expanding or contracting spectral-frequency scale. Results for shortened and lengthened VTLs are shown.

#### 3.1.4. Generation of noise-vocoded speech samples

To measure the independent effects of VTL, $f_0$ and its harmonic structure had to be eliminated in the speech samples. Therefore, noise-vocoded speech samples were generated from the original speech samples using white noise as a source signal instead of a modeled glottal sound containing the $f_0$ pattern [17,18]. A smoothed spectrum of both long and short VTL voices was then applied to the synthesized experimental VTL voices.

#### 3.1.5. Factor 1: $f_0$ height and VTL scaling

Two experimental factors were defined. Factor 1 was

| Original ID | Factor 1: Combinations of $f_0$ height and VTL scaling | Factor 2: $f_0$ range [SD] [40, 20, 2, r.p.20 (mel)] |
|-------------|---------------------------------------------------------|-----------------------------------------------------|
| M1: NM007014 | Parallel dir. $r.p.$ 20 S-L | Underline: reversed order presentation |
| M2: NM010090 | Parallel dir. 2 S-L | r.p.: reversed-phase $f_0$ contour |
| M3: NM026056 | Parallel dir. 20 S-L |
| M4: NM042090 | Parallel dir. 40 S-L |
| M5: NM071090 | Parallel dir. $r.p.$ 20 L-S |
| M6: NM084089 | Parallel dir. 2 S-L |
| M7: NM086013 | Parallel dir. 20 L-S |
| M8: NM092056 | Parallel dir. 40 L-S |
| F1: NF005072 | Opposite dir. 40 L-S |
| F2: NF008052 | Opposite dir. 20 S-L |
| F3: NF017058 | Parallel dir. 2 L-S |
| F4: NF018051 | Parallel dir. $r.p.$ 20 S-L |
| F5: NF023053 | Opposite dir. 40 S-L |
| F6: NF027086 | Opposite dir. 20 S-L |
| F7: NF028089 | Parallel dir. 2 S-L |
| F8: NF051088 | Parallel dir. $r.p.$ 20 S-L |

1Factor 2: $f_0$ range [SD] [40, 20, 2, r.p.20 (mel)]

Underline: reversed order presentation
r.p.: reversed-phase $f_0$ contour
L-S: long-short, S-L: short-long (presentation order)
The parameter was the standard deviation (SD) of the patterns. These variations in the $f_o$ range with were applied to 1) the parallel direction condition, 2) the opposite direction condition, and 3) the neutral direction condition as for factor 1. They were not applied to the noise-vocoded condition. The configurations of these conditions are also shown in Table 3.

3.1.7. Assignment of comparison pairs

The first set of 128 stimuli was generated as follows: 8 original speech samples $\times$ 2 conditions (factor 1: parallel and opposite conditions) $\times$ 4 conditions (factor 2: all four width conditions) $\times$ 2 transformations (a comparison pair: evaluation and comparison stimuli). These 128 stimuli were used to define 64 comparison pairs.

The second set of 32 stimuli was generated as follows: 16 original speech samples $\times$ 1 condition (factor 1: neutral condition) $\times$ 1 condition (factor 2: one of width conditions indicated in Table 3) $\times$ 2 transformations (a comparison pair: evaluation and comparison stimuli). These 32 stimuli were used to define 16 comparison pairs.

The third set of 16 stimuli was for the noise-vocoded condition in factor 1, generated from 8 original speech samples $\times$ 1 condition (factor 1: noise-vocoded condition) $\times$ 2 transformations (a comparison pair: evaluation and comparison stimuli). These 16 stimuli were used to define 8 comparison pairs.

3.1.8. Representation of comparison pairs and stimuli

As shown in Fig. 8, the VTL intervals (short, middle, long) were systematically different between the female and male stimuli. However, the VTL ratio for each comparison pair described their relative distance well. Therefore, the differences in $f_o$ shift amounts ($f_o$ diff) and VTL ratios were defined as representations of comparison pairs.

Each stimulus in a comparison pair was defined as either an evaluation stimulus or a comparison stimulus. A short-VTL stimulus was defined as an evaluation stimulus, and a long-VTL one was defined as a comparison stimulus. When each stimulus in a pair had the same VTL, a high-$f_o$ stimulus was treated as an evaluation stimulus.

In sum, the effects of the $f_o$ range and the combination of $f_o$ height and VTL scaling were investigated. The interaction effects between the $f_o$ range and the combination of $f_o$ height and VTL scaling were also investigated.

3.2. Method

3.2.1. Participants

A total of 140 first-year university students, 106 men and 34 women, participated in the experiment. Their ages ranged from 18 to 23. They were divided into ten groups, and unique experimental comparison pairs were assigned to each group. None of the participants had participated in Experiment 1.

3.2.2. Procedure

As in Experiment 1, each participant listened to the evaluation stimulus and comparison stimulus one after the other in each comparison pair through stereo earphones.
connected to a portable CD player (Sony: D-EJ002) at the most comfortable sound level. The presentation orders were counter-balanced among same gender for each experimental condition, as shown in Table 3.

The participants were asked to compare the two stimuli and record on an answer sheet which stimulus better matched each adjective in Table 2. They were given 60 s to complete each comparison and were free to compare them in any order of adjectives. They marked a rating score on a scale of 1 to 10 for each adjective.

3.3. Results

3.3.1. Evaluation index

Experiment 2 mainly focused on the voice pitch impression, similar to Experiment 1. The ratings for the 18 adjectives were converted into 12 evaluation indexes, as in Experiment 1 (Sect. 2.3.1). These indexes ranged from −4.5 to 4.5. A positive value indicates that the participant evaluated the evaluation stimulus more conformable than the comparison stimulus.

3.3.2. Interaction between factors 1 and 2

Figure 9 shows the means and standard errors of the evaluation indexes for voice pitch (pitch scores) against the \( f_0 \) range width. The horizontal axis represents the converted \( f_0 \) range width (factor 2), and the vertical axis represents the evaluation index. The changes in the index were plotted for three conditions (parallel, opposite, and neutral); the index for the noise-vocoded condition is shown on the right (factor 1).

Two-factor analysis of variance was applied to the pitch scores. Factor 1 was the combination of \( f_0 \) height and VTL scaling with three levels (parallel, opposite, and neutral conditions). Factor 2 was the \( f_0 \) range with four levels (40, 20, 2, \( r.p.20 \) mel). Because the cells were of unequal size, a modified error term (type III) was calculated using SAS software and used in the analysis of variance.

The main effect of the \( f_0 \) height and VTL scaling combination (factor 1) was significant \( (F_{(2,68)} = 40.31, p < 0.0001) \). The results of Tukey’s multiple comparison showed that the mean pitch score was the highest for the parallel condition followed by the neutral condition and was the lowest for the opposite condition.

The main effect of \( f_0 \) range width (factor 2) was not significant \( (F_{(3,68)} = 1.02, n.s.) \). However, interactions among the combinations of \( f_0 \) height and VTL scaling (factor 1) and \( f_0 \) range (factor 2) were significant \( (F_{(6,68)} = 4.35, p < 0.001) \).

The mean pitch scores for positive-phase 20 mel and \( r.p.20 \) mel were not significant for any condition according to multiple comparison analysis. Thus, the \( f_0 \) range of \( r.p.20 \) mel was treated as 20 mel. This result is reflected in Fig. 9.

3.3.3. Asymmetry of \( f_0 \) range effect

As shown in Fig. 9, the evaluation stimuli consisted of short VTLs in parallel or opposite conditions. In the neutral condition, the VTLs of the evaluation and comparison stimuli were equal. For the parallel and neutral conditions, the evaluation stimuli had a higher \( f_0 \) (±20 mel) than the comparison ones. For the opposite condition, the evaluation stimuli had a lower \( f_0 \) (−20 mel).

For the parallel condition, both \( f_0 \) height and VTL scaling affected pitch impression in the accelerative direction. The mean pitch scores did not change significantly, even when the \( f_0 \) ranges differed.

For the opposite condition, the \( f_0 \) height affected pitch impressions in the suppressive direction while VTL scaling affected it in the accelerative direction. The \( f_0 \) range factor clearly affected pitch impressions. When the \( f_0 \) range was wide, although the evaluation stimuli were lower than the comparison stimuli for \( f_0 \) height, the evaluation stimuli were in fact evaluated as higher voice. This trend was the reverse of the relationship between \( f_0 \) height and pitch impression. This phenomenon was also observed in Experiment 1.

However, when the \( f_0 \) range was narrower for the opposite condition, the pitch impressions were reversed. Consequently, the evaluation stimuli were perceived as lower voices. In sum, only when the intonation was flat did the perceived pitch reflect the \( f_0 \) height relation, and the reversal phenomena disappeared.

3.3.4. Effect of \( f_0 \) height

The results of the noise-vocoded condition are plotted on the right in Fig. 9. Although the noise-vocoded speech sample had neither an \( f_0 \) pattern nor a harmonic structure,
an evaluation stimulus, which had a short-VTL spectrum shape, was perceived as a higher voice. The mean pitch score for the noise-vocoded condition was located halfway between the mean score for the parallel condition and the mean score for the opposite condition.

The mean pitch score for the noise-vocoded condition is assumed to be an approximation of the effect-size independently caused by the VTL scaling. If this assumption is correct, differences between the mean scores for the noise-vocoded condition and those for the parallel or opposite condition can be considered as the effect of \( f_0 \) height. When the width of the \( f_0 \) range was 40 mel, the difference score was 0.59 between the mean score for the parallel condition (2.28) and that for the noise-vocoded one (1.69). For the parallel condition, the evaluation stimulus had a high \( f_0 \) pattern.

The difference score was –0.27 between the mean score for the opposite condition (1.42) and that for the noise-vocoded condition (1.69). These difference scores could be approximate values of the effect of \( f_0 \) height.

3.3.5. Cognitive biases caused by VTL scaling

For the neutral condition, because the stimuli in a comparison pair have equal VTLs, there is no effect from VTL scaling.

A change in the plus or minus sign of the mean pitch score leads to a reversal in roles between evaluation stimuli and comparison stimuli. Such sign inversion for the neutral condition indicates that the redefined evaluation stimulus had a low \( f_0 \) pattern (–20 mel) with no VTL scaling effect. Only in this case was the \( f_0 \) height for the neutral condition equivalent to the \( f_0 \) height for the opposite condition. Therefore, differences between negated values of the mean score for the neutral condition and mean score for the opposite condition may indicate the size of the cognitive bias caused by VTL scaling.

When the \( f_0 \) range was 40 mel, the difference score was 2.49 between the mean score for the opposite condition (1.42) and the negated mean score for the neutral condition (–1.07). This cognitive bias caused by VTL scaling was nearly 9 times greater than the absolute effect of \( f_0 \) height (–0.27).

3.3.6. Compensatory effect of \( f_0 \) range

For the opposite condition, evaluation of voice pitch obviously changed, as the width of the \( f_0 \) range became narrower. The mean pitch scores decreased from 1.42 at 40 mel to –0.40 at 2 mel in the \( f_0 \) range. The difference score was –1.82. This decrease could indicate the amount of compensation for the \( f_0 \) height effect as the \( f_0 \) pattern became flatter. In addition, the difference score between the mean score for the noise-vocoded condition (1.69) and that of pitch impression (–0.40) at 2 mel was –2.09. This difference score could indicate the degree to which the VTL scaling effect is suppressed.

When the width of the \( f_0 \) range was a flat 2 mel, reversal of pitch impression and \( f_0 \) height disappeared. In this condition, cognitive biasing caused by VTL scaling was the difference between the mean score for the opposite condition (–0.40) and the sign-changed mean score for the neutral condition (–1.83). The score for cognitive bias was 1.43, which is less than 60% of that for an \( f_0 \) range of 40 mel (2.49). However, the effects of VTL scaling still remained in this condition.

3.3.7. Peculiarity of \( f_0 \) range effect

Previous studies found that the evaluations of four traits were distinctively affected by \( f_0 \) height and VTL scaling [3,4]: voice height, voice deepness, speaking naturalness, and speaker body size. The means and standard errors of the evaluation indexes against the converted \( f_0 \) range are plotted in Fig. 10 for three of these representative traits.

VTL scaling had the greatest effect on the scores for deepness and body size while \( f_0 \) height slightly affected them. The width of the \( f_0 \) range did not affect the scores for any of the three traits. Thus, the \( f_0 \) range effect was peculiar to the listener’s impression of voice pitch.

4. DISCUSSION

4.1. Reversal Phenomenon for Voice Pitch Impression and \( f_0 \) Height

The results of Experiment 1 revealed that reducing the VTL caused the participants to have the impression of a higher voice pitch and that increasing the VTL caused them to have the impression of a lower pitch. This cognitive bias affected the impressions of voice pitch in addition to the effects of \( f_0 \) height. In fact, the biases were strong enough to reverse the relationship between the impression of voice pitch and height of \( f_0 \) pattern. That is, the cognitive bias that appeared in this experiment caused voice-pitch illusion [19,20].

This phenomenon means that the perceived pitch of speech could be inconsistent with the \( f_0 \) height. Tonal quality evidently affected the pitch impressions systematically.

4.2. Disappearance of Reversal Phenomenon

The results of Experiment 2 showed the reversal relationship between the pitch impression and the \( f_0 \) height when 1) the \( f_0 \) height and VTL scaling affected pitch impression in opposite directions and 2) the \( f_0 \) range was within that of natural speech (i.e., general intonation).

However, as the \( f_0 \) contour became flat, this phenomenon disappeared. When intonation was flat, the perceived pitch impression reflected the \( f_0 \) height. This tendency is similar to that of singing voices in which the fundamental frequency plays an important role in terms of perceived musical pitch.
4.3. Spectrum Centroid and Manipulation of VTL

This study principally focused on the structure and function of speech articulation when manipulating the assumed length of the vocal tract to systematically alter voice quality.

However, if we consider human voices as complex tones that dynamically change over time, VTL manipulation can be redefined as manipulation of the spectral centroid. This perspective of the human voice allows a more general theorization of the phenomenon.

In general, the shape of the long-term average speech spectra is practically flat below 800 Hz and inclined −10 dB/oct over 800 Hz [21]. Therefore, when we contract or expand the spectral-frequency scale to model manipulation of the VTL, we are moving the location of the spectral centroid. Reducing the VTL means raising the spectral centroid while increasing the VTL means lowering it.

4.4. Alternating Dominance between Spectrum Centroid and Fundamental Frequency Caused by Characteristics of $f_o$ Dynamics

For general intonation (in a wide $f_o$ range), the spectral centroid is the dominant factor in voice pitch evaluation. When the intonation is flat, however, $f_o$ is the dominant factor. In this condition, the evaluation of voice pitch follows the fundamental frequency.

In sum, the width of the $f_o$ range, which can be regarded as representing the stability of $f_o$ over time, determines whether the spectrum centroid or the fundamental frequency is the dominant factor in pitch impression.

4.5. Directions for Further Research

Smith et al. investigated the effects of the glottal-pulse rate (i.e., $f_o$) and the VTL on the judgment of speaker size, gender, and age. The judgment of speaker size showed that VTL had a strong effect on perceived speaker size [1,2]. In the current experiments, a similar relationship for speaker size was observed (see Figs. 5, 6).

Moreover, these experiments revealed that changing the VTL also affects the perception of voice pitch. In other words, a change in tonal quality affects pitch impression. However, this is not a new discovery. For example, for Shepard tones, tonal quality can affect perceived pitch [22,23].

The most intriguing finding of this study is that the dominant factor in pitch impression swings between the fundamental frequency and the spectrum centroid, seemingly irregularly. How long is the duration needed to observe this phenomenon? Does a consistent spectral shape on the time axis affect pitch impression? Further research is required to clarify this phenomenon and the causes of this cognitive tendency.

ACKNOWLEDGMENTS

This work was partially supported by a grant (JP15K04103) from the Japanese Society for the Promotion of Science (KAKENHI). This paper includes the contents of a previous article [4] published as an acoustical letter.
REFERENCES

[1] D. R. R. Smith, R. D. Patterson, R. Turner, H. Kawahara and T. Irino, “The processing and perception of size information in speech sounds,” J. Acoust. Soc. Am., 117, 305–318 (2005).

[2] D. R. R. Smith and R. D. Patterson, “The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age,” J. Acoust. Soc. Am., 118, 3177–3186 (2005).

[3] T. Uchida, “Subjective evaluations of voice conversion procedures modeled after manipulation of vocal tract length and speaker de-identification,” J. Acoust. Soc. Jpn. (J), 73, 151–162 (2017) (in Japanese).

[4] T. Uchida, “Reversal of the relation between impressions of voice pitch and height of fundamental frequency: Cognitive biases caused by conversion of tone quality,” Acoust. Sci. & Tech., 39, 143–146 (2018).

[5] S. Itahashi, M. Yamamoto, T. Takezawa and T. Kobayashi, “Development of ASJ Japanese newspaper article sentences corpus,” Proc. Autumn Meeting Acoust. Soc. Jpn., 2-Q-36, pp. 187–188 (1997) (in Japanese).

[6] T. Ohsuga, Y. Ishimoto and S. Itahashi, “A revision of the ASJ Japanese newspaper article sentences read speech corpus,” Proc. Autumn Meeting Acoust. Soc. Jpn., 3-Q-19, pp. 419–420 (2014) (in Japanese).

[7] D. O’Shaughnessy, Speech Communication: Human and Machine, 2nd ed. (Wiley-IEEE Press, New York, 1999), pp. 109–139.

[8] M. Morise, “Toward speech synthesis for manipulating perceived information,” J. Acoust. Soc. Jpn. (J), 74, 608–612 (2018) (in Japanese).

[9] M. Morise, F. Yokomori and K. Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Trans. Inf. Syst., E99-D, 1877–1884 (2016).

[10] M. Morise, “WORLD,” http://www.iki.yamanashi.ac.jp/~mmorise/world/english/links.html (accessed 20 Apr. 2018).

[11] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, “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(3–4), pp. 187–207 (1999).

[12] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno, “Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation,” Proc. ICASSP 2008, pp. 3933–3936 (2008).

[13] H. Banno, M. Morise and H. Kawahara, “Implementation of TANDEM-STRAIGHT in C and its GPGPU acceleration,” Proc. Autumn Meet. Acoust. Soc. Jpn., 3-7-4, pp. 1469–1470 (2013) (in Japanese).

[14] H. Kawahara, T. Kitamura, H. Takemoto, R. Nisimura and T. Irino. “Vocal tract length estimation based on vowels using a database consisting of 385 speakers and a database with MRI-based vocal tract shape information,” Proc. Interspeech 2014, pp. 870–874 (2014).

[15] H. Kido and H. Kasuya, “Everyday expressions associated with voice quality of normal utterance: Extraction by perceptual evaluation,” J. Acoust. Soc. Jpn. (J), 57, 337–344 (2001) (in Japanese).

[16] H. Kido, Y. Minowa and H. Kasuya, “Non-linear analysis of acoustic correlates associated with the expressions of voice quality: Decision-tree based approach,” J. Acoust. Soc. Jpn. (J), 58, 586–588 (2002) (in Japanese).

[17] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski and M. Ekelid, “Speech recognition with primarily temporal cues,” Science, 270(5234), 303–304 (1995).

[18] T. Kishida, Y. Nakajima, K. Ueda and G. B. Remijn, “Three factors are critical in order to synthesize intelligible noise-vocoded Japanese speech,” Front. Psychol., 7(517), pp. 1–9 (2016).

[19] The Japanese Psychonomic Society, “Illusion Contest in Japan,” http://www psy.ritsumei.ac.jp/~akitaoka/sakkon/sakkon2016.html (accessed 20 Apr. 2018).

[20] Neuroinformatics Unit, RIKEN Center for Brain Science, “Voice pitch illusion—Visiome Platform,” https://visiome.neuroinf.jp/modules/xoonips/detail.php?item_id=7243 (accessed 20 Apr. 2018).

[21] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbrabham, R. Cox, B. Hagerman, R. Hetu, J. Kei, C. Lui, J. Kiessling, M. N. Koby, N. H. A. Nasser, W. A. H. El Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, R. Meredith, T. Sirimanna, G. Tavarkiladze, G. I. Frenkonv. S. Westerman and C. Ludvigsen, “An international comparison of long-term average speech spectra,” J. Acoust. Soc. Am., 96, 2108–2120 (1994).

[22] R. N. Shepard, “Circularity in judgements of relative pitch,” J. Acoust. Soc. Am., 36, 2346–2353 (1964).

[23] R. N. Shepard, “Geometrical approximations to the structure of musical pitch,” Psychol. Rev., 89, pp. 305–333 (1982).

Teruhisa Uchida received B.A., M.A., and Ph.D. degrees in educational psychology from Nagoya University, Nagoya, Japan, in 1988, 1990, and 1996, respectively. From 1994 to 2002, he was a research associate in the Research Division of the National Center for University Entrance Examinations. He became an associate professor in 2002 and a professor in 2017 at the National Center. His work includes developing listening comprehension tests for the annual National Center Test for University Admissions and analyzing trends and characteristics of candidates who have taken the National Center Test. He belongs to the Acoustical Society of Japan, the Acoustical Society of America, the Japanese Psychological Association, the Japanese Association of Educational Psychology, the Phonetic Society of Japan, and the Japan Association for Research on Testing. He received the 2003 Japanese Psychological Association Research Award, was the winner of the 2016 Special Jury Prize and the 2018 Auditory Illusion Prize of the Visual and Auditory Illusion Contest in Japan, and received the Best Paper Award from the Japan Association for Research on Testing (JART) in 2014 and 2018.