Title:

Acoustic features correlated to perceived urgency in evacuation announcements

Authors:
Maori Kobayashi, Yasuhiro Hamada, and Masato Akagi

Affiliation:
School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahi-dai, Nomi, Ishikawa, Japan

Corresponding Author:
Maori Kobayashi, Ph.D.
E-mail: maori-k@jaist.ac.jp

Abstract:

To encourage prompt evacuation behavior during disasters, evacuation announcement systems are required to adjust the perceived urgency of announcements to the danger of a situation while maintaining voice clarity. In this study, we aimed to understand acoustic features correlated to the perceived urgency in speech when clarity is maintained. For this purpose, we used a speech synthesis tool to manipulate the key acoustic features: the time duration, F0 (instantaneous fundamental frequency), and spectrum. Specifically, we replaced these features in five real evacuation announcements that were spoken clearly by a TV announcer during an actual disaster and had different magnitudes of urgency. We found quantitatively and qualitatively that the perceived urgency was mostly influenced by the F0 of speech. Furthermore, by manipulating the F0 time average and variation, we compared the influence of the F0's static constant feature and its dynamic fluctuation pattern on the magnitude of perceived urgency. The results indicated that both types of F0 information influenced the magnitude of perceived urgency in real Japanese speech. Our results suggest that the sense of urgency in evacuation announcements can be controlled by adjusting the F0 while maintaining voice clarity.

Keywords:

vocoded speech, evacuation announcement speech, perceived urgency

1. Introduction

Japan faces the persistent threat of natural disasters and takes pains to be prepared for the worst. An example of such preparation is a disaster administration wireless communication system called *bousai-musen*, which consists of a local warning system network. Many municipalities have banks of loudspeakers mounted on poles as part of a broadcast system that stands ready to alert residents to impending natural disasters or other large-scale civil emergencies. The *bousai-musen* system uses speech and alarm sounds to raise warnings of earthquakes or provide other vital information in an emergency, giving residents valuable seconds to find a safe place.

Although the *bousai-musen* system has proven invaluable in emergencies, problems have also been pointed out. For example, in the 2011 Great Tohoku Earthquake in Japan, it was reported that some citizens ignored warnings and did not evacuate, despite hearing announcements calling for evacuation. This phenomenon is considered to be due to normalcy bias, which refers to a mental state that causes some people facing a disaster to underestimate both the possibility of a disaster occurring and its possible effects (Drabek, 1986; Omer and Alon, 1994).

In the case of the 2011 Great Tohoku Earthquake, it is assumed that some people did not recognize the situation's danger because the *bousai-musen* announcements were calm and polite. In contrast, fewer people fell victim to the resulting tsunami in municipalities where the announcements sounded urgent (Reports on tsunami evacuation measures, 2012). These reports indicate the need to examine speech styles for conveying danger to people appropriately. It was also reported that some people who read the evacuation announcements were late to escape the tsunami (Reports on tsunami evacuation measures, 2012). Accordingly, an automatic evacuation announcement system is necessary, and it should control the magnitude of an announcement's urgency according to the situation.

Previous studies have examined the speech duration, including the speech rate (syllables per minute) and time intervals between sentences, the speaker's gender, and the fundamental frequency (F0) as factors affecting the perceived urgency in speech (Jang, 2007; Hellier et al., 2002; Park and Jang, 1999).

Notably, it was demonstrated that the time duration is essential for determining the magnitude of perceived urgency: a reduction in duration increases the magnitude (Hellier et al., 2002; Park and Jang, 1999; Jang, 2007). It was also reported that women's speech gives the impression of higher urgency than men's speech, which might result from differences in acoustic effects due to the differing vocal-tract characteristics between genders (Edworthy et al., 1991; Edworthy, 1994). Consistently with those findings, some reports showed that the magnitude of perceived urgency is also influenced by the F0 (F0 average, F0 contour) of a sentence (Park and Jang, 1999). Also, in general, the louder a warning signal, the higher the estimation of urgency (Jang, 2007). A relation between loudness and a speaking style of "shouting" with high urgency has also been reported (Mittal and Yegnanarayana, 2013; Seshadri and Yegnanarayana, 2009). Finally, it has been pointed out that the overall intensity of a voice plays a critical factor in urgency, but this has not been studied systemically (Jang, 2007).

In the above studies, researchers arbitrarily manipulated the types and amounts of acoustic features in synthesized speech produced by a simple text-to-speech (TTS) system, in order to examine each acoustic feature's influence on the magnitude of perceived urgency in speech. Therefore, it is still unclear how humans actually perceive a sense of urgency in real speech and what strategies they use to inform others of danger via speech. Previous studies reported that prosody-related features have an important role on affective speech perception (Grinkovtsova et al., 2012; Hammerschmidt and Juregens, 2007; Lieberman and Michaels, 1962). For instance, by using human voice conversion techniques, Xue et al. (2019) showed that the F0 and spectrum are important for emotional speech perception. It has been reported that the F0 is very important for rich voice expression, e.g., for artificial voices (Kawahara, Fujimura, and Konpaku, 2006; Mittal and Yegnanarayana, 2015; Sudarsana and Yegnanarayana, 2019). Jang (2007) reported that a decrease in speech duration to 70% increases the magnitude of perceived urgency by 150%; however, it is challenging to actually speak at such a fast rate. It is also hard to hear sentences spoken quickly in public spaces, especially for the elderly (Konkle et al., 1977; Wingfield et al., 1985). Therefore, to develop automatic evacuation announcement systems, it is necessary to examine the appropriate acoustic factors for conveying danger while maintaining high intelligibility.

Accordingly, we focus here on evacuation announcements spoken by professional announcers in a real disaster. Professional speakers are trained to speak in an easy-to-understand speech style (Kuwabara and Ogushi, 1984; Kashimada et al., 2009; Bele, 2007; Hazlett et al., 2011). After the 2011 Great Tohoku Earthquake, professional announcers on TV and radio urgently warned of the need for evacuation. In addition, they consciously or unconsciously changed the magnitude of perceived urgency in evacuation announcements. We assume that evacuation announcements by professional announcers are both intelligible and urgent, with the urgency changing gradually.

In this article, we describe a study that sought to examine the acoustic features determining the perceived urgency of speech, in order to adjust the magnitude of perceived urgency to the situation while maintaining intelligibility. To focus on the speaking style, we manipulated the prosody-related features—namely, the time duration, F0, and spectrum (added to the power envelope)—of speech by professional announcers. We conducted a series of quantitative and qualitative experiments to determine the significant acoustic features and the associations among them when human participants perceived a sense of urgency in real speech. Many studies have been conducted using only quantitative measures to examine acoustic features' influence on the magnitude of perceived urgency (e.g., Hellier et al., 2002; Jang, 2007). However, a sense of urgency is considered a higher-order concept composed of multiple impressions. In our previous study on auditory impressions of speech, which used the semantic differential (SD) method and factor analysis (Kobayashi and Akagi, 2018), the perceived urgency of speech was described with various adjectives. Thus, by using psychological methods, we measured quantitative changes and qualitative influences to determine acoustic features' effect on the magnitude of perceived urgency.

In the present study, we performed two preliminary experiments and three main experiments to examine the acoustic features of perceived urgency. Through the preliminary experiments, we selected speech stimuli that conveyed different magnitudes of perceived urgency for use in the main experiments. We used the selected stimuli to synthesize new stimuli that were modulated to have specific acoustic characteristics. Then, we presented these synthesized stimuli as experimental stimuli in the main experiments. We conducted two experiments (Experiments 1 and 2) to examine the acoustic features related to the perceived urgency of speech, qualitatively or quantitively. Then, we conducted Experiment 3 to examine the effects of the static and dynamic components of the F0, which was the most influential acoustic feature in the first two experiments, on the perceived urgency of speech. We used a magnitude estimation method in Experiments 1 and 3 to enable participants to quantitatively evaluate the perceived urgency of the speech stimuli, and we applied analysis of variance (ANOVA) to these data. In Experiment 2, we asked participants to qualitatively evaluate the perceived urgency by the SD method, and we examined these data through factor analysis.

The rest of the paper is structured as follows. Section 2 details the original speech data and the method of selecting them for speech synthesis. Section 3 describes the details of the signal processing used for synthesis and the acoustic features that were examined in each experiment. Sections 4 and 5 describe the common methodology among the experiments and the specific procedures for each experiment, respectively. Finally, the results of each experiment are given in section 6, and the results for the acoustic features of perceived urgency are discussed in section 7.

2. Speech data

To experimentally examine the acoustic features related to the perceived urgency of speech, we had to present speech stimuli that were manipulated to highlight each feature. To this end, we used a speech synthesis tool, STRAIGHT (Legacy-STRAIGHT: Kawahara, et al., 1999), to produce converted speech by replacing three acoustic features in the original speech—namely, the duration, F0, and spectrum—with different magnitudes of urgency. It was also necessary to keep the linguistic content constant. Thus, we used five speech stimuli with the same linguistic information but different speaking styles. Through two preliminary experiments, these five stimuli were chosen from 57 evacuation announcements spoken by professional announcers during an actual disaster. In this section, we describe the preliminary experiments to select the original speech stimuli and the acoustic feature replacement procedure to produce the speech stimuli for the main experiments.

2.1. Evacuation speech data from actual disaster

We conducted two perceptual listening experiments to select speech stimuli that conveyed different magnitudes of perceived urgency and could be used for synthesis. These experiments were reported previously in a published work (Kobayashi and Akagi, 2018), but we repeat the findings here because the published paper was written in Japanese.

In these experiments, we used recorded speech data from actual evacuation announcements when a tsunami was predicted after an earthquake off the Fukushima Prefecture coast on November 22, 2016. The announcements were spoken by multiple professional announcers with various speaking styles on different TV channels. From the 16 hours of recorded data, we selected 57 evacuation announcements that had similar content (e.g., "Get away now") and were spoken by 14 announcers (nine men, five women; age range: 20s–50s) with different speech styles. In preliminary tests before the preliminary experiments, we confirmed that participants could distinguish the speech styles of these 57 stimuli.

2.2. Selection of speech data with different magnitudes of perceived urgency for synthesis

The first experiment was conducted to examine the speech style's influence on the evacuation behavior. We asked 50 participants (25 men, 25 women; age range: 20s–60s, with 10 participants for each decade) to evaluate each speech stimulus in terms of four evaluation items. Participants selected the closest opinion to their own from four options provided for each evaluation item. Two of the four items and their answer options were as follows: Q1: "*Would you follow the voice instruction?*" (A1: "*yes, I would*"; A2: "*probably*"; A3: "*probably not*"; A4: "*no, I would not*"); Q2: "*How dangerous did you think the situation was from the voice*?" (A1: "*very dangerous*"; A2: "*pretty dangerous*"; A3: "*somewhat dangerous*"; A4: "*not dangerous*"). The percentage of participants who chose A1 was used as an index of the instruction's effectiveness (Q1) and its indication of danger (Q2) for each announcement. Figure 1 shows a scatter plot of the indexes for Q1 and Q2 for the 57 voice announcements. There was a high correlation between these indexes (Pearson product-moment correlation coefficient: $r = 0.74$, $p < 0.01$), which implies that the speaking styles of the evacuation announcements significantly conveyed the danger and encouraged evacuation.

In the second experiment, we examined the auditory impressions of these 57 announcements by using the SD method and factor analysis. We asked the same 50 participants to rate each voice in terms of 16 adjective pairs by using a 7-point scale. As summarized in Table 1, three factors were extracted via each adjective pair's factor loading: "clarity," "urgency," and "vocal quality." As a measure of each impression's strength, we calculated each factor's score for each announcement; a positive score indicated greater strength. Figure 2 shows a scatter plot of the "urgency" score and the Q2 index from Preliminary Experiment 1 for the 57 evacuation announcements. The high correlation ($r = 0.76$, $p < 0.01$) suggests that the chosen announcements were appropriate for investigating the magnitude of perceived urgency. In addition, we found a high correlation ($r = 0.76$, $p < 0.01$) between the "clarity" and "urgency" scores, which suggests that the perceived clarity of the announcements depended on the strength of the impression of urgency. Furthermore, the average evaluation values for the eight adjective pairs involving clarity (see Table 1) were more than 3 for all 57 announcements, which means that all of them received a positive rating for clarity. From these results, we concluded that the clarity of the announcements was preserved regardless of the speech style in terms of the perceived urgency.

From these preliminary results, we selected five announcements with different magnitudes of urgency for conversion in the main experiments. As mentioned above, several speech stimuli had the same content but different prosody when spoken by the same announcer on a TV broadcast during the actual disaster. Among such stimuli, the five chosen stimuli were ranked by an appropriate procedure, and we thus regarded them as reasonable for examining acoustic features correlated to the perceived urgency of speech. Here, we denote their magnitudes as high (H), moderately high (MH), moderate (M), moderately low (ML), and low (L) according to the "urgency" score. Figure 2 represents these five stimuli with black circles. Hereafter, when we refer to speech stimuli, we strictly mean these five stimuli. They were spoken by the same male announcer, who spoke the same Japanese sentence in each case (「今すぐ逃げてください」 /i/ma/su/gu/ni/ge/te/ku/da/sa/i/, meaning "Get away now" in Japanese).

Some might consider our dataset too small for this kind of validation. However, as mentioned above, our speech stimuli were taken from actual broadcast announcements, and there were few announcements in which the same announcer uttered exactly the same words. From these voice announcements, we selected 57 speech stimuli that were sufficient to distinguish different speaking styles in the preliminary tests before the preliminary experiments. Accordingly, we consider these five speech stimuli to be a fair dataset for this analysis.

============ here are Figures 1, 2 & Table 1 for 1-column =============

3. Signal processing methods and acoustic features

The experimental speech stimuli were generated according to speech parameters obtained by the STRAIGHT analysis and synthesis procedure (Kawahara et al., 1999). STRAIGHT combines pitch-adaptive spectral analysis with a surface reconstruction method in the time-frequency region. The procedure also included F0 extraction through an instantaneous frequency calculation based on the concept of "fundamentalness." Successive refinements of the F0 and spectral parameter extraction procedures enabled the total system to resynthesize high-quality speech. For more details, refer to the cited study (Kawahara et al., 1999).

The time duration, F0 and spectrum of the speech stimuli were extracted for synthesis. The time duration of each phoneme was measured manually. The F0 and spectrum (added power spectrum) were estimated using STRAIGHT. In this analysis, the FFT length was 1024 points, and the frame rate was 1 ms.

3.1. Acoustic feature replacement procedure

Figure 3 illustrates the procedure for replacing the F0. The time information was modified to keep the speech durations of the source and target stimuli the same; this had to be done before converting the F0 or the spectrum. To modify the time duration, first, the speech signal was manually segmented at the phoneme level for the target and source stimuli. Then, each phoneme duration in the source stimulus was modified to match that in the target stimulus, according to a linear ratio of their durations in each stimulus. By applying STRAIGHT, an initial synthesized stimulus (source speech 2) was obtained by modifying only the time duration as described above. Next, the F0s, spectrum, and aperiodic component (Ap) of the source stimulus (source speech 2) were extracted at 1-ms intervals by using STRAIGHT. Similarly, these features were also extracted from the target stimulus. Because the time duration of source speech 2 was the same as that of the target stimulus, the source stimulus' F0s could be directly replaced with those of the target stimulus. The Ap and spectrum of source speech 2 and the F0s of the target stimulus were then combined for synthesis by STRAIGHT. Finally, the synthesized speech with F0 replacement for conversion was obtained. By following this procedure, the source stimulus' spectrum and Ap information were kept, but its F0s were changed to those of the target stimulus.

Replacement of the spectrum mostly followed the same procedure as for F0 replacement; the exception was the last step, in which we used the F0 and Ap from the source stimulus, so that the spectrum of the target stimulus could be synthesized. In this case the source stimulus' spectrum was changed to that of the target stimulus, but the other information was kept. Note that the spectrum here refers to the upper envelope of a speech signal's power spectrum, which is related to the power of speech.

============== here are Figures 3 for 2-column ==================

3.2. Acoustic feature conditions for each experiment

In our main experiments described below, the source stimulus was the L or H stimulus, and the other stimuli were used as the targets for synthesis. We replaced the specified features of the L, ML, M, or MH target stimulus with those of the H source stimulus; similarly, we replaced the specified features of the ML, M, MH, or H target stimulus with those of the L source stimulus.

3.2.1. Experiments 1 and 2

The experimental speech stimuli consisted of the five original stimuli and the stimuli synthesized by the above procedure. We used two types of experimental conditions, which involved four magnitudes of the perceived urgency (L, ML, M, MH, H) and five acoustic features. The components of the acoustic feature conditions were as follows.

Duration (Dur): The length of the source stimulus was stretched to match that of each target stimulus. Under this condition, the speech rate and pause duration between phrases were also manipulated.

Fundamental frequency (F0s): The F0s of the source stimulus were replaced with those of each target stimulus.

Spectrum (Spec): The spectrum of the source stimulus was replaced with that of each target stimulus. Also, the RMS of the synthesized speech was confirmed to be equivalent to that of the target.

All (All): For the source stimulus, all three of the above features were replaced with those of each target stimulus. In theory, the converted speech under this condition was the resynthesized target stimulus.

Original: The original five speech stimuli selected from Preliminary Experiment 1 (L, ML, M, MH, H) were used as is.

Because of the synthesis process, the duration of each stimulus also varied depending on the target stimulus, under all conditions. The speech stimuli were presented at 58–62 dBA.

3.2.2. Experiment 3

As shown in Figure 4, the experimental stimuli were created by using STRAIGHT to synthesize speech stimuli with the same F0 time average but different F0 time variation, or with the same F0 time variation but a different F0 time average. First, we extracted the F0s of each stimulus by the same procedure used in Experiment 1. Then, the overall time average was calculated from the F0s for each stimulus. To obtain the fluctuation in F0, the frequency differences between the F0s and the overall time average were calculated every 1 ms. The mean F0 of the source stimulus was modified to be the same as that of the target stimulus, and the stimulus was then resynthesized using the modified F0s. We refer to this case as the "F0 average" condition (Figure 4A). Alternatively, the mean F0 of the target stimulus was modified to be the same as that of the source stimulus, the F0s of the source stimulus were replaced with those of the target stimulus, and the mean F0 was modified by the same replacement procedure used in Experiment 1. We refer to this case as the "F0 fluctuation" condition (Figure 4B). We confirmed in advance that the mean F0 and the F0 fluctuation of the synthesized speech were equal to those of the target.

We used the L and H stimuli as the sources for speech synthesis. For the L source stimulus, the target was the ML, M, MH, or H stimulus; similarly, the targets for the H source stimulus were the other four stimuli. The sound pressure level was fixed at 62 dBA.

In addition to the acoustic feature conditions described above, we also used a condition referred to as the "F0s" condition. Under this condition, both the F0 averages and the F0 fluctuation were varied among those of the target stimuli (Figure 4C). Accordingly, this condition was the same as the F0 component of the acoustic feature conditions used in Experiments 1 and 2.

=========== here is Figure 4 for single column ============

4. Experimental methodology

4.1. Ethical statement

Informed written consent was obtained from each participant before the experiments were conducted. The Japan Advanced Institute of Science and Technology (JAIST) ethics committee approved all procedures.

4.2. Participants

A total of 48 men and women participated in the experiment as listeners. In Experiment 1, 10 native Japanese speakers participated (five men, five women; between 22 and 24 years old; normal hearing). In Experiment 2, 28 native Japanese speakers participated (18 men, 10 women; between 22 and 24 years old; normal hearing). In Experiment 3, 10 native Japanese speakers participated (five men, five women; between 24 and 26 years old; normal hearing). Ten listeners participated in all of the experiments. The interval between experiments 1 and 3 was about nine months.

4.3. Apparatus

The experiment was controlled using MATLAB 2015 (Mathworks) on a computer (Windows 10, Intel Core i5). The stimuli were presented to both ears of the listeners through a digital-to-analog (D/A) converter (RME, Fireface UCX), a driver unit (STAX, SRM-1/MK-2), and electrostatic headphones (STAX, SR-404).

5. Experimental procedures

5.1. Experiment 1: Quantitative measurement of acoustic features related to perceived urgency

Experiment 1 was conducted to examine the influence of the acoustic features of speech on the magnitude of perceived urgency through quantitative psychological measurement. We converted the five original speech stimuli for synthesis, as described above, by replacing the speech duration, spectrum, and F0s with those of other stimuli among the five. If changing only the F0s caused the synthesized speech to be evaluated as similar to the target stimulus in terms of the perceived urgency, then this would mean that the F0s greatly contributed to the perceived urgency. On the other hand, if this kind of change was evaluated as similar to the source stimulus, then this would mean that the F0s were not related much to the perceived urgency.

The tests were conducted in a soundproof room. The magnitude of perceived urgency of each stimulus was measured using the magnitude estimation method (Stevens, 1957). The listeners were asked to quantify the magnitude of the urgency of each stimulus as compared to that of a control stimulus with a base score of 100. For example, they were asked to give a score of 200 when they perceived twice the magnitude of urgency as compared to the control; similarly, they were asked to give a score of 50 when they perceived half the magnitude. The original L stimulus was used as the control. Ten trials were repeated for each stimulus, and the order was randomized for each listener.

5.2 Experiment 2: Qualitative measurement of acoustic features related to perceived urgency

Experiment 2 was conducted to examine the qualitative change in perceived urgency in terms of the acoustic features of speech by using the semantic differential (SD) method and factor analysis, for the same speech stimuli used in Experiment 1. The participants' auditory impressions of each stimulus were evaluated using the SD method on a 7-point scale. As summarized in Table 2, we used the same 16 adjective pairs as in Preliminary Experiment 2. The orders of the speech stimuli and adjectives were randomized. The listeners were allowed to listen to each stimulus repeatedly. The alignment of each adjective pair was fixed for each participant in order to maintain balance of the alignments across all listeners.

5.3. Experiment 3: Effects of static or dynamic component of F0 on perceived urgency

Experiment 3 was conducted to determine the essential factor for perceived urgency by manipulating the F0 time average and F0 time variation. In this study, we assumed that the F0 was the sum of the F0 fluctuation and the mean F0 (in terms of the log frequency). The procedure was the same as that in Experiment 1.

6. Results

6.1. Experiment 1: Quantitative measurement of acoustic features related to perceived urgency

The geometric means of the five scores were calculated for each stimulus as the magnitude of perceived urgency for each listener. Figure 5 shows these mean magnitudes for each converted stimulus with respect to each target stimulus. Under the "All" condition, the magnitudes for the converted stimuli were almost the same as those of the target stimuli. We thus conclude that the synthesis process did not affect the evaluation of perceived urgency. When the time duration or spectrum of the resynthesized speech was replaced with that of a target stimulus, however, the magnitudes of perceived urgency differed significantly. In contrast, the magnitudes of perceived urgency under the "F0s" condition were almost the same as for the target stimuli.

We performed two-way repeated analysis of variance (ANOVA) for the acoustic feature conditions (Dur, F0s, Spec, All, Original) and magnitude conditions (ML, M, MH, and L or H). The results showed that the interaction between the acoustic features and the magnitude was significant for both source speech stimuli (L: $F_{9,159}$ =12.98, $p < .001$, η^2 =0.06; H: $F_{9,159}$ =9.94, $p < .001$, η^2 =0.05). In addition, for both source stimuli, post hoc tests (Ryan's method) revealed significant differences in the magnitude of perceived urgency among the "All," "Dur" (p = 0.0001), and "Spec" (p = 0.0001) conditions, and among the "F0s," "Dur" (p = 0.0001), and "Spec" (p = 0.0001) conditions. However, these tests revealed no significant differences between the "All" and "F0s" conditions (p = 0.93) and between the "Dur" and "Spec" conditions (p = 0.43). These results indicated the same tendency as in our pilot study with other listeners. Accordingly, they suggest that changes in the F0s of speech have an essential effect on the magnitude of perceived urgency.

================ here is Figure 5 for single column ==================

6.2. Experiment 2: Qualitative measurement of acoustic features related to perceived urgency

For each evaluation score, factor analysis was conducted for the 16 adjective pairs with the main factor method and promax rotation. Because the factor loading of one pair ("*stiff-soft*") was less than 0.4 for all factors, the factor analysis was repeated with that pair excluded. Two factors were extracted with an eigenvalue of at least 1 (see Table 2). The adjective pairs "*busy–tranquil*," "*tense–relaxed*," and "*fervent–detached*" each had a high factor load on the first factor. We considered this factor the same as "urgency" in Preliminary Experiment 2 and regarded it to be related to the perceived urgency of speech. The adjective pairs "*forceful–not forceful*," "*aware–unaware*," and "*well projected–poorly projected*" each had a high factor load on the second factor. We considered this factor the same as "clarity" in Preliminary Experiment 2 and regarded it to be related to the clarity of speech.

Figure 6 shows the "urgency" score for each converted stimulus with respect to each target stimulus. We found that this score under the "All" condition was almost the same as for each target stimulus. The results were consistent with those of Experiment 1, indicating that the conversion process did not affect the impression of perceived urgency in speech. Moreover, the "urgency" score under the "F0s" condition was close to the score for each target stimulus. In contrast, the scores under the "Spec" and "Dur" conditions significantly differed from those for the target stimuli. These results indicate that the F0s are essential for perceiving urgency, which matches the results of Experiment 1.

Figure 7 shows the semantic profiles of each converted stimulus under each acoustic feature condition. The scores for the "urgency" adjective pairs ("*busy-tranquil*" to "*light-heavy*") varied with the F0s irrespective of the source stimulus; this result indicates that the F0s affect all aspects of perceived urgency. In contrast, the scores under the "Spec" and "Dur" conditions changed depending on the source stimulus (L or H) irrespective of the target stimulus. The scores under the "Dur" condition changed depending on the target stimulus for adjective pairs related to the speech rate, such as "*busy-tranquil*" or "*fast-slow*," although they changed depending on the original stimulus for other adjective pairs with a high "urgency" factor load. This result indicates that the speech rate only somewhat affects the magnitude of perceived urgency. Overall, these results suggest that the auditory impression of perceived urgency is strongly affected by the F0s of speech.

=============== here is table 2 for 2-column ====================

======= here is Figure 6 for 1-column and Figure 7 for 2-column =======

6.3. Experiment 3: Effects of static or dynamic components of F0 on perceived urgency

As in Experiment 1, we calculated the geometric means of the five scores for each stimulus as the magnitudes of perceived urgency for each listener. Figure 8 shows these mean magnitudes. Under the "F0s" condition, the magnitude increased or decreased depending on the magnitude of the target stimulus. The results supported the hypothesis that the magnitude of perceived urgency is affected by the F0 information.

Two-way repeated ANOVA for the acoustic feature conditions ("F0 average," "F0 fluctuation," and "F0s") and magnitude conditions (ML, M, MH, and L or H) revealed that the magnitude condition had a significant main effect for both source stimuli (L: $F_{3,27} = 6.41$, $p < 0.01$, $\eta^2 = 0.16$; H: $F_{3,27} = 13.38$, $p < 0.01$, $\eta^2 = 0.14$). For the H source stimulus, the two-way repeated ANOVA showed a significant interaction between the three acoustic feature conditions and the magnitude condition ($F_{6,54} = 3.21$, $p < 0.01$, $\eta^2 = 0.01$). In a post hoc test, the simple main effect of the acoustic feature condition was significant for each target stimulus other than MH (L: $F_{2,72} = 5.54$, $p < 0.01$; ML: $F_{2,72} = 9.86$, $p < 0.01$; M: $F_{2,72} = 10.54$, $p < 0.01$; MH: $F_{2,72} = 1.94$, $p = 0.15$). These results ($p < 0.05$) revealed that the magnitude was different among the "F0s," "F0 average," and "F0 fluctuation" conditions, but for all target speech stimuli, there was no significant difference between the "F0 average" and "F0 fluctuation" conditions. The post hoc test also revealed that the magnitudes of perceived urgency differed between the magnitude condition and all three acoustic feature conditions. For the L source stimulus, there was no significant interaction ($F_{6,54} = 1.21$, $p = 0.31$, $\eta^2 = 0.02$), nor was there a significant effect of the acoustic feature conditions ($F_{2,18} = 2.42$, $p = 0.12$, $\eta^2 = 0.03$). Also, the magnitude of perceived urgency under the "F0 fluctuation" condition was consistent with the results under the "F0s" condition in Experiment 1. This agreement suggests that the results were reliable.

Finally, these results indicate that both the F0 time average and F0 time variation influence the magnitude of perceived urgency, although each effect is weak.

============== here is Figure 8 for single column ============

7. Discussion

Influences of F0 on perceived urgency

This study aimed to elucidate the acoustic features determining the perceived urgency of speech in order to control the urgency while maintaining clarity in an automatic evacuation announcement system. By replacing the F0s and other acoustic features in actual speech stimuli, we performed experiments using evacuation announcements spoken by professional announcers. From the results of these experiments, we argue that the F0 is the most critical acoustic feature for determining the magnitude of the perceived urgency of speech. We showed quantitatively and qualitatively that the F0 affects the sense of urgency. In particular, the semantic profiles obtained in Experiment 2 revealed that the F0 affected the participants' evaluation of the adjective pair "*tense-relaxed*." In contrast, the speech duration changed the evaluation of the "*busyness*" of speech but had little influence on the magnitude of perceived urgency. These results indicate that the impression of tension in evacuation announcements influences the magnitude of perceived urgency.

When people announce emergencies, they feel both physical tension (tightening of the muscles around the vocal cords) and mental tension. In this study, we assumed that such muscle tension is reflected in the F0 information of speech. Previous studies reported that the respiratory system is affected by excited autonomic nervous system activities in emergencies, and glottal pressure changes the F0 and sound level of speech (Scherer, 1986). We may perceive the F0 as necessary for determining the magnitude of perceived urgency in speech as a result of learning associations between F0 changes and the activities of the autonomic nervous system in daily life. However, there is little established knowledge concerning the effect of physiological arousal on speech production and consequent changes in acoustic speech signals (Banse and Scherer, 1996). It will thus be necessary to examine the relationships among physiological arousal, speech production, and acoustic features with perceived urgency.

Our study results were consistent with those of previous studies showing the influence of F0 modulation, but our results indicated a difference concerning the dynamic components of the F0. Specifically, our study showed that the magnitude of perceived urgency increases when the dynamic F0 fluctuation is significantly large (see Figure 7). This is inconsistent with a previous study, which found that the magnitude of perceived urgency increases when the F0 contour is flat rather than fluctuating (Park and Jang, 1999). Such differences in F0 fluctuation, which are useful for the perceived urgency, may be due to language differences. Alternatively, we assume that both flat and large fluctuations are essential. The H voice has both flat parts and large-fluctuation parts in the same sentence. The contrasts may be essential or may depend on the part of a sentence. We consider both local and overall enhancement to occur when people try to convey information desperately.

Because of limits in our method of feature replacement accompanied by changes in the time duration, we could not strictly distinguish the effect of the F0s from the effect of the time duration. We will thus need to further examine prosody effects, especially the F0 fluctuation.

Evacuation announcements in terms of affective speech and F0

It is possible that our results can be regarded as a case of affective speech, especially speech by a charismatic or persuasive voice, because the voices in our stimuli were intended to make people obey an evacuation order. Conventional research has reported that the F0 or F0 range is correlated with a voice's characteristic of persuasion or leadership (Mayew et al., 2013; Neibuhr et al., 2016; Signorello et al., 2020; Zoghaib, 2019). Previous studies have shown that a lower pitch is positively associated with dominance or leadership (Mayew et al., 2013; Zoghaib, 2019), which is inconsistent with our results. On the other hand, it was recently reported that charismatic leaders convey messages by using a voice with a higher overall F0, a wider F0 range, and an SPL range (Signorello et al., 2020; Niebuhr, 2016); those findings are similar to our results. We cannot address the differences here, but we conclude that our results were at least related to the dangerous situation being conveyed in the experimental speech data. The importance of the F0 in persuasive speech will thus need further examination.

Benefits of F0 modulation to control perceived urgency of speech

We also point out the benefits of F0 modulation for controlling the magnitude of perceived urgency in speech. In this study, we demonstrated that the magnitude of perceived urgency is influenced by F0 fluctuation: a higher mean F0 and larger F0 fluctuation increased the magnitude of perceived urgency. Such F0 modulation does not affect the intelligibility of evacuation announcements and can encourage specific evacuation behaviors. Our findings also show the possibility of arbitrarily controlling the magnitude of perceived urgency in evacuation announcements according to the emergency. We expect that appropriate expression of urgency will lead people to recognize the danger of a situation and reduce their normalcy bias during disasters. For future work, we will study the relationship between temporal F0 fluctuation and the magnitude of perceived urgency in more detail and develop a guideline for using this relationship in synthesized speech.

Differences in results from previous studies

Finally, we describe the inconsistencies from previous findings. Our study revealed that the sound pressure level and time duration are not especially crucial for the perceived urgency of speech. These results differ from those of previous reports (Jang, 2007). A decreased speech duration changed the impression of busyness (Experiment 2) but did not change the perceived urgency of speech (Experiment 1). Moreover, we found that the perceived urgency was not influenced by spectrum modulation with the power envelope in Experiments 1 and 2. This result indicates that the effect of voice intensity is smaller than has been reported. We thus suggest two possible explanations.

The first is that the five speech stimuli used in our study were spoken by a professional TV announcer. Professional announcers are trained to speak clearly and rationally at all times. The results of Preliminary Experiment 2 showed that these speech stimuli were heard clearly regardless of the magnitude of perceived urgency. It has been reported that clear speech has specific spectrum features such as formant contours or formant space (Uchasaki, 2008; Smiljanic and Bradlow, 2005; Amano-Kusumoto et al., 2014; Furui, 1985) and a speech rate that is not too fast. The five stimuli used for conversion in this study had sufficient spectral features to sound clear, so there were no significant differences between them. As a result, the replacement of spectral information had no influence on the magnitude of perceived urgency. Moreover, actual speech is constrained by the articulatory organs, so there was little difference in the amount of each acoustic feature among the five stimuli. For example, a speech rate change may be effective: it is difficult to speak clearly at 1.5 times the standard speech rate.

The second possible explanation is that the five speech stimuli were in Japanese, which distinguishes words by using pitch to accent particular mora, i.e., by using F0 differences (Kawahara, 2015). Accordingly, the listeners emphasized or were sensitive to pitch differences in the stimuli. In the future, we will examine this influence of language in detail.

8. Conclusions

In this study, we examined the acoustic features controlling perceived urgency in speech with clarity maintained. Through preliminary experiments, we selected five speech stimuli with different magnitudes of urgency but the same clarity from 57 evacuation announcements spoken by professional announcers during a real disaster. In our main experiments, we used converted speech stimuli that replaced three acoustic features—namely, the speech duration, fundamental frequency (F0), and spectrum (voice intensity)—with the features of other stimuli among the five. Both quantitatively and qualitatively, the results indicated that the perceived urgency was most influenced by the F0. Moreover, we examined the influences of dynamic and static F0 information on the magnitude of perceived urgency. These results indicated that both types of F0 information influenced the magnitude of perceived urgency in real Japanese speech. Overall, our results suggest the possibility that the perceived urgency of evacuation announcements is controlled by the F0, depending on the emergency, when speech clarity is maintained. We will need to further examine the role of F0 time variation on the perceived urgency.

Acknowledgments

This study was supported by the SECOM Science and Technology Foundation, the JST-Mirai Program of the Japan Science and Technology Agency (Grant Number: JPMJMI18D1), and the SCOPE Program of the Ministry of Internal Affairs and Communications (Grant Number: 201605002).

References

Amano-Kusumoto, A., Hosom, J., Kain, A., Aronoff, J. M. 2014. Determining the relevance of different aspects of formant contours to intelligibility. Speech Communication, 59, 1–9.

Banse, R., Scherer, K. R. 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70, 3, 614-636.

Bele, I. V. 2007. Dimensionality in voice quality. Journal of Voice, 21, 3, 257-272.

Cabinet Office Japan, Disaster Management in Japan, Tsunami Evacuation Countermeasures Working Group 2012. Reports on tsunami evacuation measures (in Japanese). http://www.bousai.go.jp/jishin/tsunami/hinan/pdf/report.pdf (accessed July 2, 2020).

Drabek, T. E. 1986. Human system responses to disaster: an inventory of sociological findings. New York: Springer Verlag.

Edworthy, J., Loxley, S., Dennis, I. 1991. Improving auditory warning design: relationship between warning sound parameters and perceived urgency. Human Factors, 33, 2, 205–231.

Edworthy, J. 1994. The design and implementation of non-verbal auditory warnings. Applied Ergonomics, 25, 4, 202–210.

Furui, S. 1986. On the role of spectral transition for speech perception. Journal of Acoustical Society of America, 80, 4, 1016–1025.

Grichkovtsova, I., Morel, M., Lacheret, A. 2012. The role of voice quality and prosodic contour in affective speech perception. Speech Communication, 54, 3, 414-429.

Hammerschmidt, K., Jurgens, U. 2007. Acoustical correlates of affective prosody. Journal of Voice, 21, 5, 531-540.

Hazlett, D. E., Duffy, O. M., Moorhead, S. A. 2011. Review of the impact of voice training on the vocal quality of professional voice users: implications for vocal health and recommendations for further research. Journal of Voice, 25, 2, 181-191.

Hellier, E., Edworthy, J., Weedon, B., Walters, K., Adams, S. 2002. The perceived urgency of speech warnings: semantics versus acoustics. Human Factors: The Journal of the Human Factors and Ergonomics Society, 44, 1, 1–17.

Jang, P. S. 2007. Designing acoustic and non-acoustic parameters of synthesized speech warnings to control perceived urgency. International Journal of Industrial Ergonomics, 37, 3, 213–223.

Kashimada, D., Ogita, K., Ishikawa, T., Hasegawa, K., Ayama, M. 2009. Effects of voice training on subjective evaluation of voice quality. The Journal of The Institute of Image Information and Television Engineers, 63, 12, 1818–1823 (in Japanese).

Kawahara, S. 2015. The phonology of Japanese accent. In Kubozono, H. (ed.), The handbook of Japanese language and linguistics: phonetics and phonology. Berlin: Mouton, pp. 444–492.

Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A. 1999. Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Communication, 27, 3-4, 187–207.

Kawahara, H., Fujimura, O., Konpaku, Y. 2006. Voice quality of artistic expression in noh: an analysis synthesis study on source-related parameters. Journal of Acoustical Society of America, 120, 3028.

Kobayashi, M., Akagi, M. 2018. Psychological evaluation of evacuation calling voice. Journal of the Acoustical Society of Japan, 74, 12, 633–640 (in Japanese).

Konkle, D. F., Beasley, D. S., Bess, F. H. 1977. Intelligibility of time-altered speech in relation to chronological aging. Journal of Speech and Hearing Research, 20, 1, 108–115.

Kuwabara, H, Ohgushi, K. 1984. Acoustic characteristics of professional male announcers' speech sounds. Acta Acustica united with Acustica, 55, 233–240.

Lieberman, P., Michaels, S. B. 1962. Some aspects of fundamental frequency and envelope amplitudes as related to the emotional content of speech. Journal of the Acoustical Society of America, 34, 7, 922–927.

Mayew, W. J., Parsons, C. A., Venkatachalam, M. 2013. Voice pitch and the labor market success of male chief executive officers. Evolution and Human Behavior, 34, 4, 243-248.

Mittal, V. K., Yegnanarayana, B. 2013. Effect of glottal dynamics in the production of shouted speech. Journal of Acoustical Society of America, 130, 5, 3050-3061.

Mittal, V. K., Yegnanarayana, B. 2015. Study of characteristics of aperiodicity in noh voice. Journal of Acoustical Society of America, 137, 3411.

Niebuhr, O., Voße, J., Brem, A. 2016. What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice. Computers in Human Behavior, 64, 366-382.

Omer, H., Alon, N. 1994. The continuity principle: a united approach to disaster and trauma. American Journal of Community Psychology, 22, 2, 273–287.

Park, K. S., Jang, P. S. 1999. Effects of synthesized voice warning parameters on perceived urgency. International Journal of Occupational Safety and Ergonomics, 5, 1, 73–95.

Scherer, K. R. 1986. Vocal affect expression: a review and a model for future research. Psychological Bulletin, 99, 2, 143–165.

Seshadri, G., Yegnanarayana, B. 2009. Perceived loudness of speech based on the characteristics of glottal excitation source. Journal of Acoustical Society of America, 126, 4, 2061-2071.

Signorello, R., Demolin, D., Bernardoni, N. H., Gerratt, B. R., Zhang, Z., Kreiman, J. 2020. Vocal fundamental frequency and sound pressure level in charismatic speech: a cross-gender and –language study. Journal of Voice, 34, 5, 808.e1-808.e13.

Smiljanić, R., Bradlow, A. R. 2005. Production and perception of clear speech in Croatian and English. Journal of Acoustical Society of America, 118, 1677–1688.

Stevens, S. S. 1957. On the psychological law. Psychological Review, 64, 3, 153-181.

Sudarsana, R. K., Yegnanarayana, B. 2019. Analysis of aperiodicity in artistic noh singing voice using an impulse sequence representation of excitation source. Journal of Acoustical Society of America, 146, 4446.

Uchasaki, R. M. 2008. Clear Speech. In: Pisoni, D. B., Remez, R. E. (eds.), The Handbook of Speech Perception, Blackwell Publishing, pp. 207–235.

Wingfield, A., Poon, L. W., Lombardi, L., Lowe, D. 1985. Speed of processing in normal aging: effects of speech rate, linguistic structure, and processing time. Journal of Gerontology, 40, 5, 579–585.

Xue, Y., Hamada, Y., Akagi, M. 2018. Voice conversion for emotional speech: rule-based synthesis with degree of emotion controllable in dimensional space. Speech Communication, 102, 54-67.

Zoghaib, A. 2019. Persuasion voices: the effects of speaker's voice characteristics and gender on consumers' responses. Recherche et Applications en Marketing (English edition), 34, 383-110.

Captions:

Figure 1:

Scatter plot of the results of Preliminary Experiment 1 to examine the influence of speech style on the evacuation behavior for 57 evacuation speech announcements. The vertical axis indicates the percentage of participants who selected "*yes*" for the evaluation item Q1: "*Would you follow the voice instruction?*". The horizontal axis indicates the percentage who selected "*very dangerous*" for the evaluation item Q2: "*How dangerous did you think the situation was from the voice?*". The black circles represent the five speech stimuli used in the main experiments, while the gray circles represent speech stimuli used only in the preliminary experiments.

Figure 2:

Scatter plot of the results of Preliminary Experiments 1 and 2. The vertical axis indicates the "urgency" score in Preliminary Experiment 2, which examined the auditory impressions of the 57 evacuation speech announcements. The horizontal axis indicates the percentage of participants who selected "*very dangerous*" in Preliminary Experiment 1. A positive score indicates a high magnitude of perceived urgency. The black circles represent the five speech stimuli used in the main experiments, while the gray circles represent speech stimuli used only in the preliminary experiments. The black circles are labeled according to the "urgency" scores for those stimuli: high (H), moderately high (MH), moderate (M), moderately low (ML), and low (L).

Figure 3:

Synthesis process using STRAIGHT to produce the converted speech stimuli for the experiments. This figure illustrates the process for the F0s, in which the F0s of the L source stimulus replaced those of the H target stimulus.

Figure 4:

F0 contours for an example of a stimulus used in Experiment 3, under the (A) F0 average, (B) F0 fluctuation, and (C) F0s conditions.

Figure 5:
Mean magnitudes of perceived urgency for the converted stimuli with respect to the original stimuli, across all listeners by the magnitude estimation method. The gridlines show that the magnitudes of perceived urgency were the same between the original and converted stimuli. The original speech used for synthesis was (A) the L source stimulus or (B) the H stimulus.

Figure 6:
"Urgency" factor scores obtained by factor analysis and the SD method for the converted stimuli. The gridlines show that the magnitudes of perceived urgency were the same between the target and converted stimuli. The original speech used for synthesis was (A) the L source stimulus or (B) the H stimulus.

Figure 7:
Semantic profiles of stimuli under the "Dur," "F0s," and "Spec" conditions.

Figure 8:
Mean magnitudes of perceived urgency for each converted stimulus. The horizontal axis indicates the target stimulus with urgency. The original speech used for synthesis was (A) the L source stimulus or (B) the H stimulus.

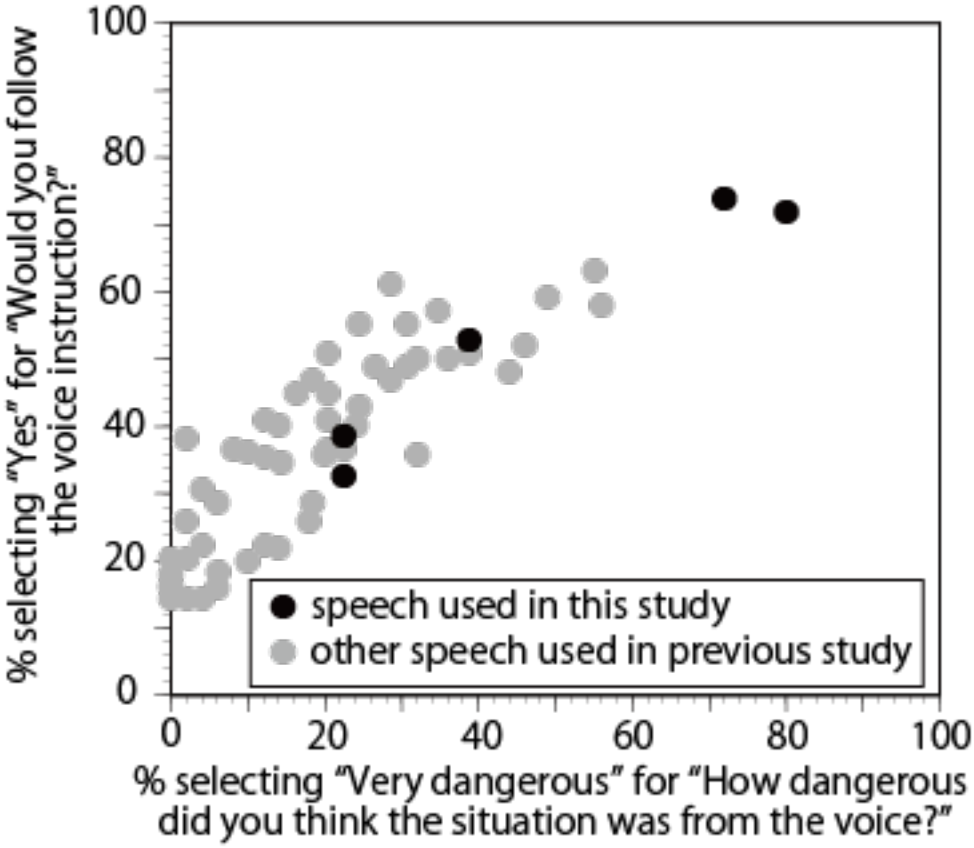
Table 1: Factor analysis results in Preliminary Experiment 2 with the SD method for evacuation announcements.

			1st factor (clarity)	2nd factor (urgency)	3rd factor (voice quality)
1. Clarity					
Well projected	-	Poorly projected	.83	.37	.11
Powerful	-	Weak	.82	.49	-.06
Aware	-	Unaware	.81	.55	.14
Distinct	-	Vague	.80	.40	.08
Forceful	-	Not forceful	.80	.45	.08
Loud	-	Quiet	.78	.47	-.00
Pleasant	-	Unpleasant	.62	.30	-.13
Bright	-	Dark	.56	.45	.42
2. Urgency					
Fervent	-	Detached	.54	.81	.26
Tense	-	Relaxed	.44	.80	.25
Busy	-	Tranquil	.41	.80	.42
Emotional	-	Rational	.30	.75	.32
Fast	-	Slow	.44	.69	.46
3. Voice quality					
Stiff	-	Soft	.45	.46	.17
High	-	Low	.37	.43	.61
Light	-	Heavy	-.28	.09	.58
		Eigenvalue	6.57	2.30	65.0
		Cumulative contribution ratio	43.5	57.9	1.14

Table 2: Factor analysis results in Experiment 2 with the SD method for synthesized speech.

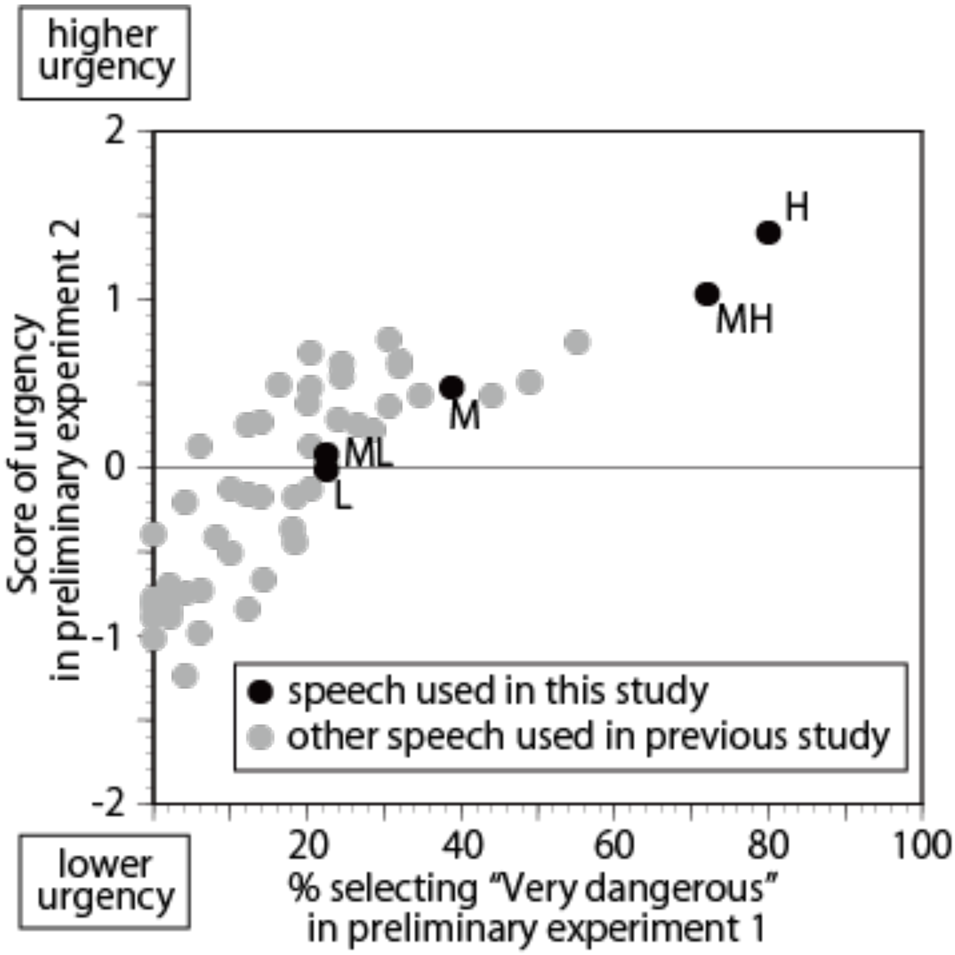
			1st factor (urgency)	2nd factor (clarity)
1. Urgency				
Busy	-	Tranquil	.88	-.11
Tense	-	Relaxed	.83	.07
Fervent	-	Detached	.81	.11
Emotional	-	Rational	.79	.00
Fast	-	Slow	.73	-.14
High	-	Low	.69	.04
Bright	-	Dark	.54	.31
Light	-	Heavy	.51	-.36
2. Clarity				
Forceful	-	Not forceful	.02	.77
Aware	-	Unaware	.12	.72
Pleasant	-	Unpleasant	-.19	.63
Well projected	-	Poorly projected	-.09	.62
Powerful	-	Weak	.19	.60
Distinct	-	Vague	-.10	.57
Loud	-	Quiet	.74	.53
		Eigenvalue	4.91	3.6
		Cumulative contribution ratio	32.71	56.7

Figure 1:



100
80
60
40
20
0
% selecting "Yes" for "Would you follow the voice instruction?"
speech used in this study
other speech used in previous study
0
20
40
60
80
100
% selecting "Very dangerous" for "How dangerous did you think the situation was from the voice?"

Figure 2:



higher urgency
lower urgency
Score of urgency in preliminary experiment 2
% selecting "Very dangerous" in preliminary experiment 1
2
1
0
-1
-2
20
40
60
80
100
H
MH
M
ML
L
speech used in this study
other speech used in previous study

Figure 3:

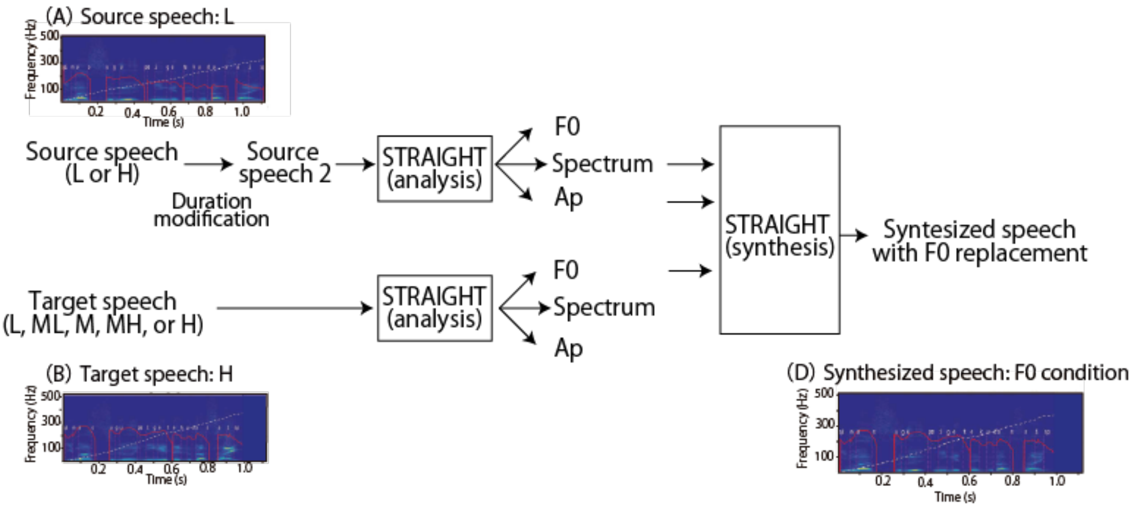

(A) Source speech: L
Source speech
(L or H)
Duration
modification
Source
speech 2
STRAIGHT
(analysis)
F0
Spectrum
Ap
Target speech
(L, ML, M, MH, or H)
STRAIGHT
(analysis)
F0
Spectrum
Ap
STRAIGHT
(synthesis)
Syntesized speech
with F0 replacement
(B) Target speech: H
(D) Synthesized speech: F0 condition
Frequency (Hz)
Time (s)

Figure 4:

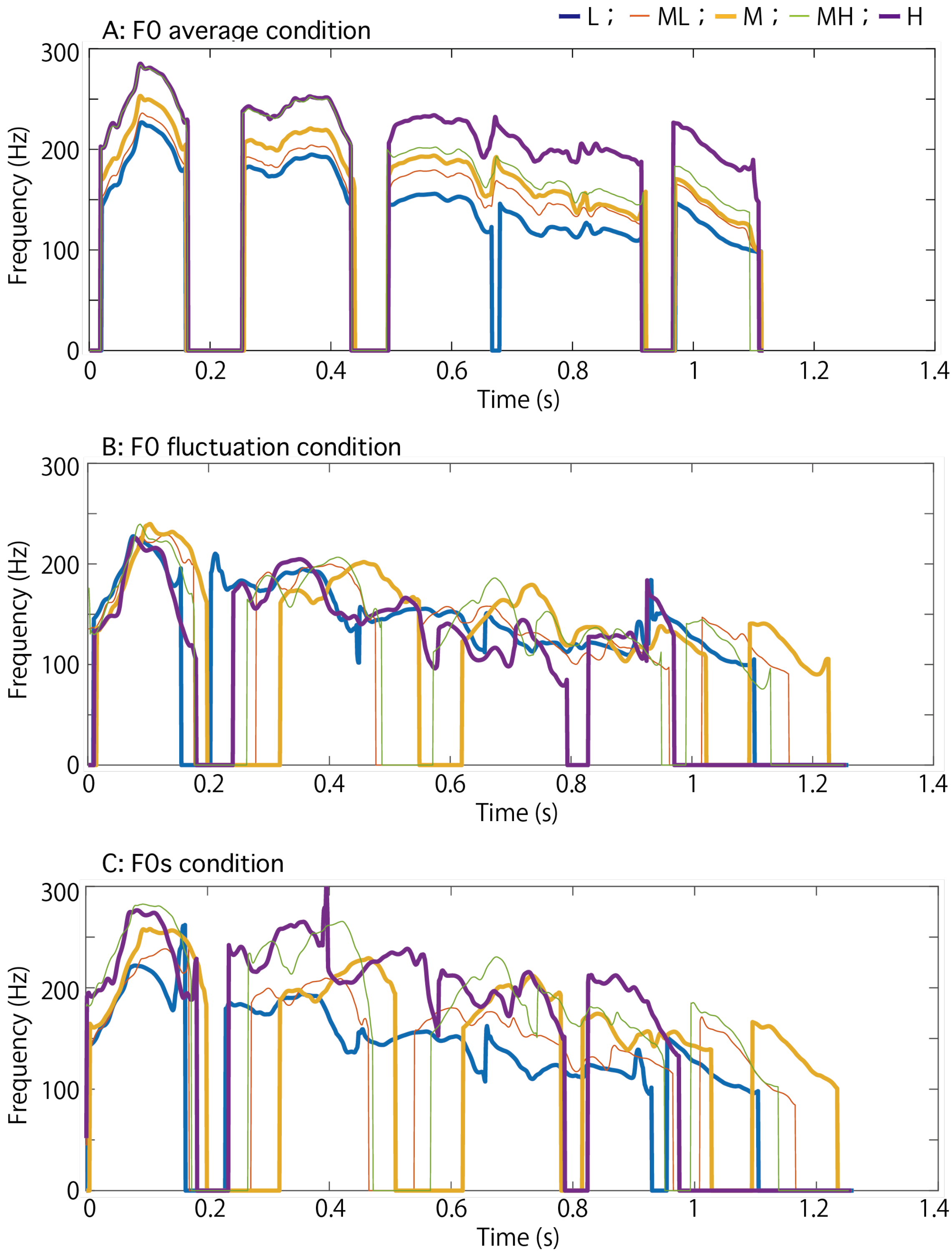
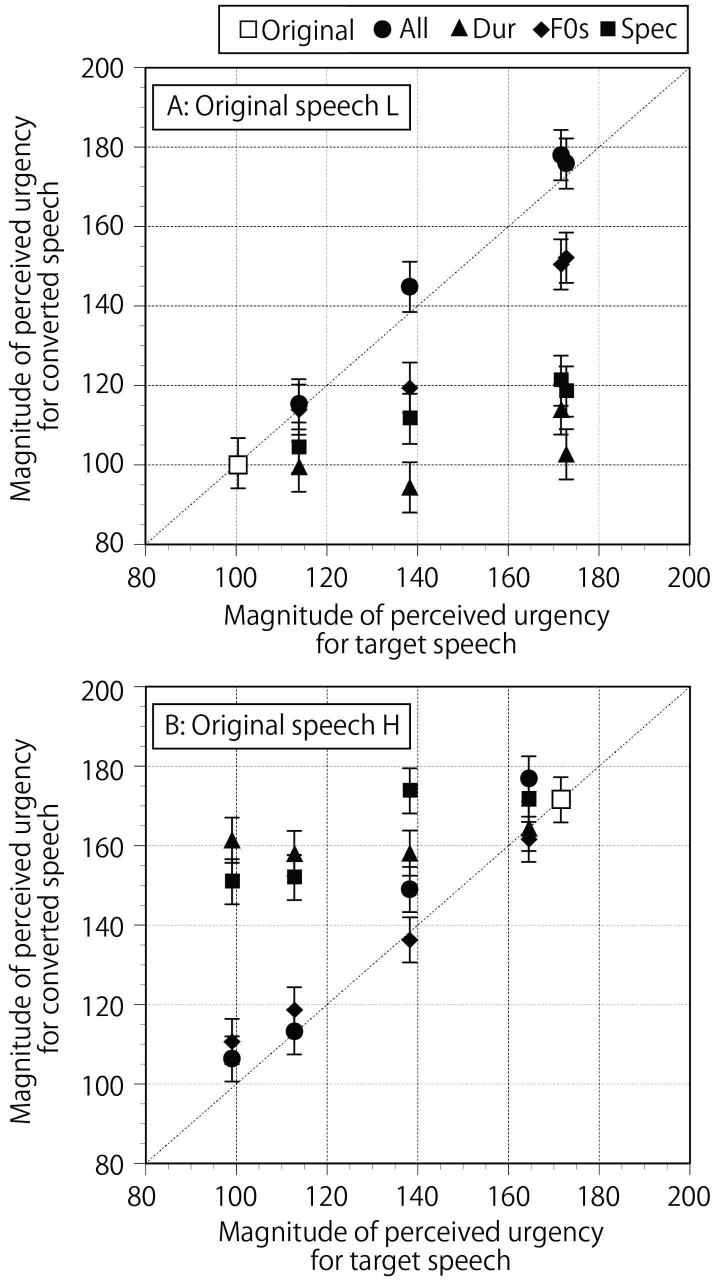

A: F0 average condition
L ; ML ; M ; MH ; H
Frequency (Hz)
Time (s)
B: F0 fluctuation condition
Frequency (Hz)
Time (s)
C: F0s condition
Frequency (Hz)
Time (s)

Figure 5:


□Original ● All ▲Dur ◆F0s ■Spec
A: Original speech L
Magnitude of perceived urgency for converted speech
Magnitude of perceived urgency for target speech
200
180
160
140
120
100
80
80
100
120
140
160
180
200
B: Original speech H
Magnitude of perceived urgency for converted speech
Magnitude of perceived urgency for target speech
200
180
160
140
120
100
80
80
100
120
140
160
180
200

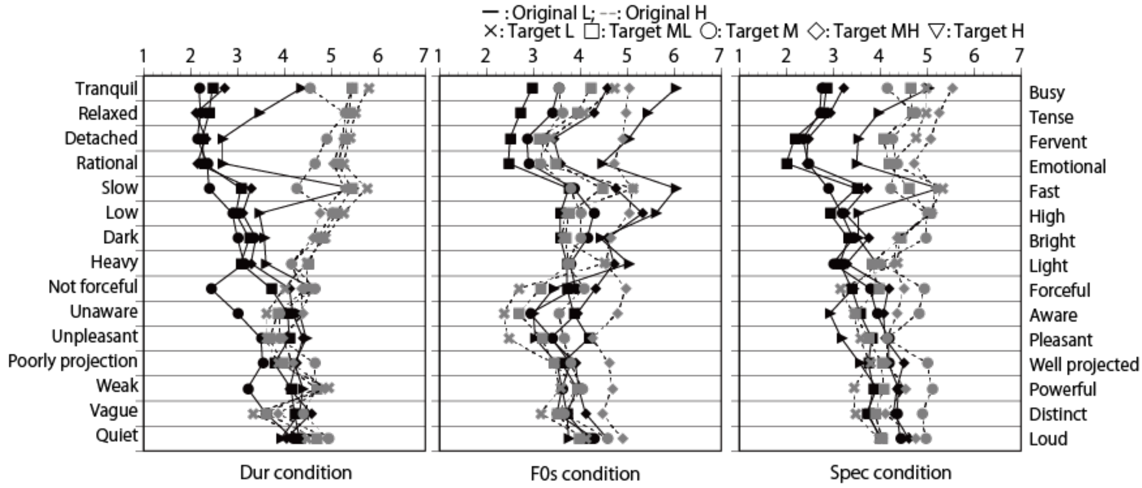
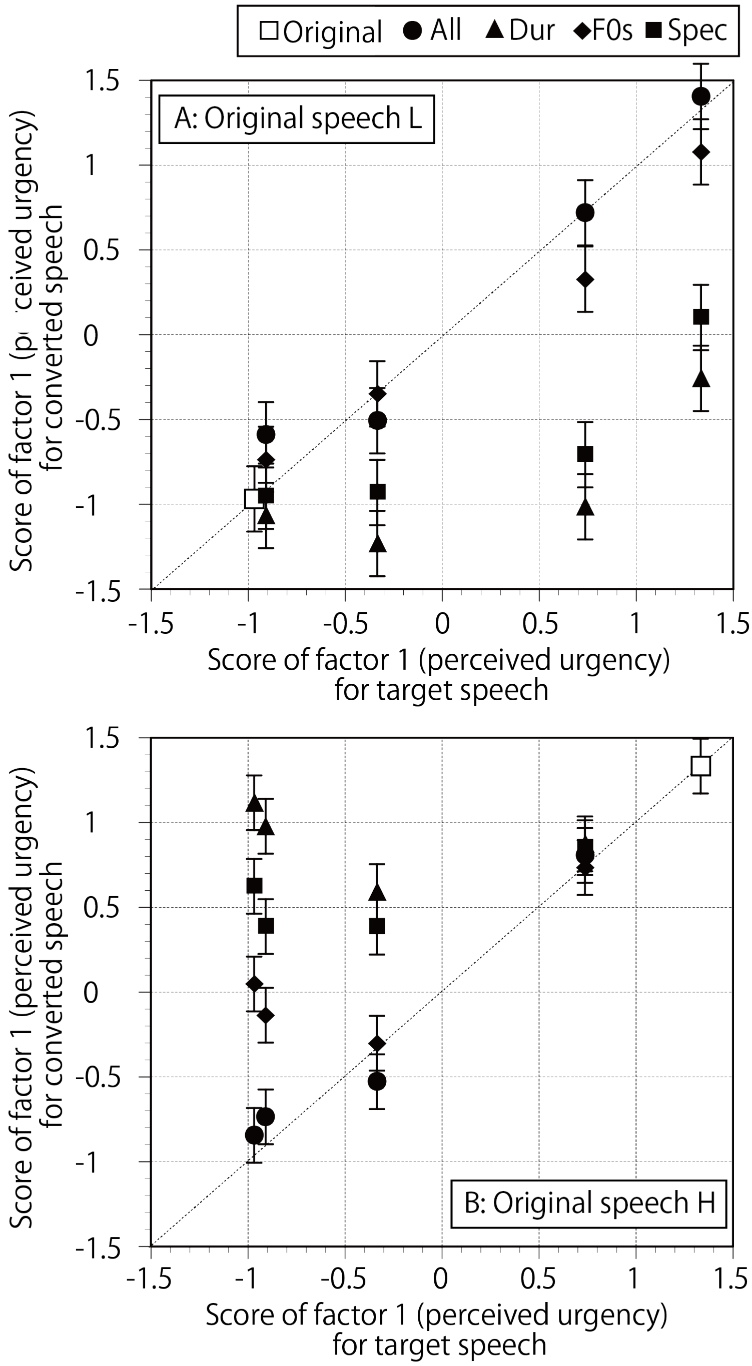

—: Original L; - -: Original H
×: Target L □: Target ML ○: Target M ◇: Target MH ▽: Target H
1 2 3 4 5 6 7
Tranquil
Relaxed
Detached
Rational
Slow
Low
Dark
Heavy
Not forceful
Unaware
Unpleasant
Poorly projection
Weak
Vague
Quiet
Busy
Tense
Fervent
Emotional
Fast
High
Bright
Light
Forceful
Aware
Pleasant
Well projected
Powerful
Distinct
Loud
Dur condition
F0s condition
Spec condition

Figure 6:



□Original ●All ▲Dur ◆F0s ■Spec
A: Original speech L
Score of factor 1 (perceived urgency) for converted speech
Score of factor 1 (perceived urgency) for target speech
B: Original speech H
Score of factor 1 (perceived urgency) for converted speech
Score of factor 1 (perceived urgency) for target speech
1.5
1
0.5
0
-0.5
-1
-1.5

Figure 7:

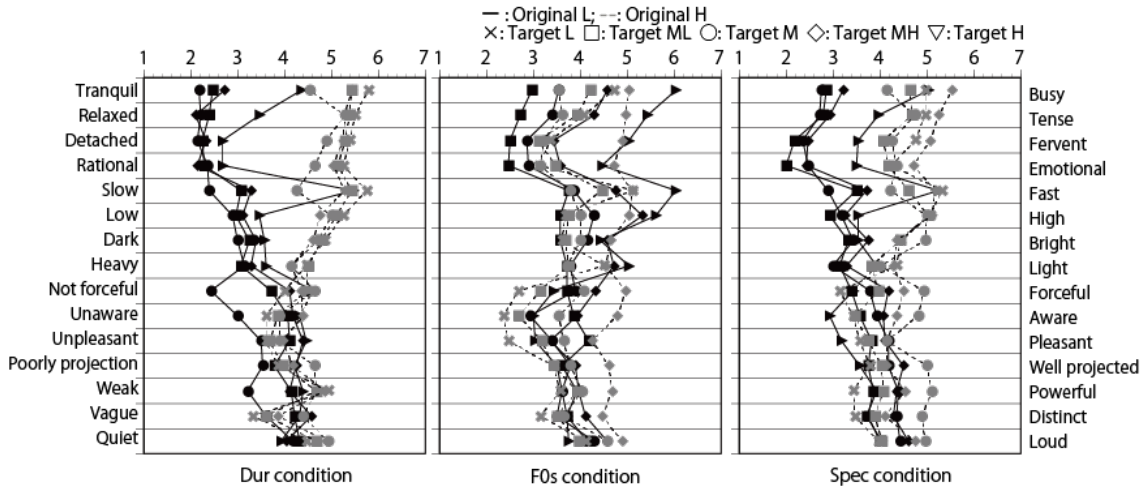
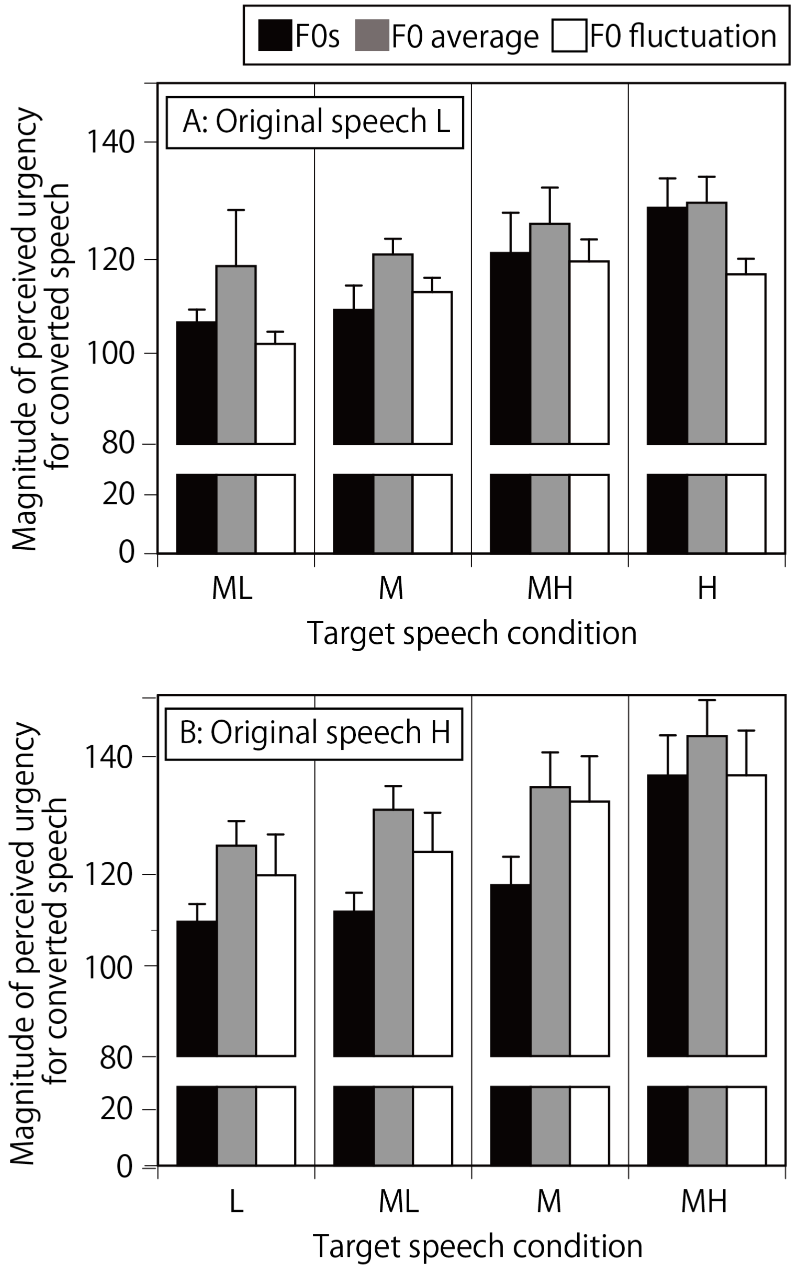

—: Original L; - -: Original H
×: Target L □: Target ML ○: Target M ◇: Target MH ▽: Target H
1 2 3 4 5 6 7
Tranquil
Relaxed
Detached
Rational
Slow
Low
Dark
Heavy
Not forceful
Unaware
Unpleasant
Poorly projection
Weak
Vague
Quiet
Busy
Tense
Fervent
Emotional
Fast
High
Bright
Light
Forceful
Aware
Pleasant
Well projected
Powerful
Distinct
Loud
Dur condition
F0s condition
Spec condition

Figure 8:


F0s
F0 average
F0 fluctuation
A: Original speech L
Magnitude of perceived urgency for converted speech
140
120
100
80
20
0
ML
M
MH
H
Target speech condition
B: Original speech H
Magnitude of perceived urgency for converted speech
140
120
100
80
20
0
L
ML
M
MH
Target speech condition