Speech Communication 182 (2026) 103431

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Acoustic cues in Japanese pitch-accent identification: Perception of sine-wave and noise-vocoded sine-wave speech[1]

Yasuaki Shinohara

Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo, 169-8050, Japan

ARTICLE INFO

Keywords:
Japanese pitch accent
Production and perception
Sine-wave speech
Noise-vocoded sine-wave speech

ABSTRACT

Fundamental frequency (F0) is a highly correlated acoustic cue for Japanese pitch accent. This study examined whether Japanese listeners make use of acoustic cues other than F0 to identify Japanese pitch-accent minimal-pair words when F0 is unavailable. After demonstrating that two-mora pitch-accent contrasts are correlated with F0, F1, duration, and intensity in production, I transformed the natural recordings into sine-wave speech (SWS) and noise-vocoded SWS (NzVocSWS) to investigate any secondary acoustic correlates of pitch-accent identification. SWS typically consists of three time-varying sinusoids following the frequencies and amplitudes of the first, second, and third formants (F1, F2, and F3) of natural recordings. I predicted that listeners would rely on the lowest sinusoidal component (F1) as a substitute for F0 when hearing SWS, because the F1 lies within the dominance region (approximately 400–1000 Hz), which affects pitch perception. In contrast, the noise-vocoding technique, which modulates the frequency and amplitude patterns of the three sinusoids with noise, was expected to reduce the influence of F1. For the NzVocSWS stimuli, Japanese listeners would rely more on the temporal envelope, leading to significant effects of duration and intensity cues on the pitch-accent identification. As a result, the identification accuracy was predicted to increase from the SWS to the NzVocSWS condition. The results supported all of these predictions and demonstrated that duration and intensity are secondary acoustic cues for Japanese pitch-accent perception.

1. Introduction

1.1. Pitch perception

Pitch is an auditory sensation in which a sound is perceived as high or low on a musical scale, reflecting how the auditory system interprets the physical properties of acoustic signals (Yost, 2009). Pitch perception relies on both spectral and temporal coding mechanisms. In the spectral domain, each location on the basilar membrane within the cochlea acts as an auditory filter and specifically responds to a certain frequency range with a center frequency. Listeners use those overlapping auditory filters and infer pitch from the resulting excitation pattern (Moore, 2008). The fundamental frequency (F0) is a highly correlated acoustic cue with the perceived pitch, but it is not the only relevant acoustic property. For example, even when the F0 component is missing, pitch can still be perceived based on harmonics for low-frequency sound (e.g., approximately 50 Hz) where the auditory periphery could resolve the harmonics. However, for high-frequency sound (e.g., approximately 1000 Hz) where the auditory periphery is unable to resolve the harmonics, listeners might use temporal analyses of auditory nerve fibers to account for the pitch of harmonic sounds (Schouten, 1938, 1940; Yost, 2009).

According to Yost (2009), in the temporal domain, auditory nerve fibers can synchronize their discharges with rapid fluctuations in the acoustic waveform, known as phase locking. This preserves information about the temporal fine structure of the signal, and this temporal representation can support the autocorrelation-like pitch extraction mechanism. Such coding is effective primarily below approximately 5000 Hz. In addition to fine-structure information, slower modulation in neural firing rate tracks the amplitude envelope of the waveform. When this envelope exhibits regular periodicity, the reciprocal of its period can contribute to the perceived pitch of the stimulus, particularly in situations where fine-structure cues are degraded or unavailable (Yost, 2009).

[1] A portion of this work was presented in "Japanese pitch-accent perception of noise-vocoded sine-wave speech" at the 183rd meeting of Acoustical Society of America, Nashville TN, United States, 5–9 December 2022. The abstract of the presentation was published in *The Journal of the Acoustical Society of America, 152*(4), A175. https://doi.org/10.1121/10.0015940

E-mail address: y.shinohara@waseda.jp.

https://doi.org/10.1016/j.specom.2026.103431
Received 19 March 2025; Received in revised form 17 April 2026; Accepted 9 June 2026
Available online 10 June 2026

Taken together, these findings support the idea that pitch perception reflects contributions from spectral and temporal coding, which will be modulated by the available information in signal. Spectral coding is most reliable at low frequencies, where auditory filters provide high resolution and individual auditory nerve fibers respond to a small number of resolved harmonics. Temporal coding based on phase locking to the temporal fine structure is also robust in the frequency region below about 5000 Hz (Yost, 2009). At higher frequencies, because of the limitation of phase locking, temporal fine-structure coding becomes less reliable (Oxenham, 2013). In this region, pitch perception may rely more on temporal envelope cues. Thus, pitch perception is shaped by both acoustic properties of the stimulus and their transformation in the auditory system.

1.2. Japanese pitch accents

Japanese is a pitch-accent language having lexical contrasts signaled by the presence and location of accents (Beckman and Pierrehumbert, 1986; Kawahara, 2015; Kitahara, 2001). For example, when [haɕi] is produced with the accented first mora [ha] using a high fundamental frequency (F0) followed by a sharp fall toward the following mora [ɕi] (i.e., high–low pitch), it means *chopsticks* in Tokyo Japanese. In contrast, when [haɕi] is produced with a low F0 for [ha] and high F0 for [ɕi] (i.e., low–high pitch), it can mean either *bridge* or *edge*. When the accented second mora [ɕi] with a high F0 is followed by a sharp fall toward a particle (e.g., [haɕio] with a low–high–low (LHL) pitch), the word means *bridge*; when the second mora [ɕi] has a high F0 followed by a particle retaining a high F0 with no abrupt pitch fall (e.g., [haɕio] with low–high–high (LHH) pitch, known as an "unaccented" word), it means *edge*. Thus, F0-related cues (e.g., F0 peak/fall locations and F0 fall rate) are typically used as the primary acoustic cues to identify accented syllables (Hasegawa and Hata, 1992; Kawahara, 2015; Kitahara, 2001).

Many studies have investigated the presence of secondary acoustic cues in addition to the primary F0 cue in Japanese pitch accents. Compared to tone (e.g., Mandarin), intonation (e.g., English), and other pitch-accent languages (e.g., Swedish, Serbo-Croatian, Turkish), Japanese lexical pitch accents rely heavily on F0 cues (Beckman, 1986; Hualde et al., 2002; Ladd, 2008; Levi, 2005). Although intensity is an acoustic correlate of pitch accent, it has been debated whether duration is a secondary acoustic cue for lexical pitch accents (Cutler and Otake, 1999; Igarashi, 2015; Kawahara, 2015; Sugiyama, 2017, 2022; Weitzman, 1970). Sugiyama (2017, 2022), for example, demonstrated that Japanese listeners could identify final-accented words (e.g., [haɕio] with LHL) or unaccented words (e.g., [haɕio] with LHH) to some extent without F0 information, suggesting that secondary acoustic cues exist in pitch-accent perception. However, as there was no strong correlation between word identification accuracy and the use of duration, intensity, or formant frequency cues, they concluded that it was unclear whether duration and intensity were secondary acoustic cues for Japanese pitch-accent perception.

1.3. Properties of sine-wave speech and noise-vocoded sine-wave speech

Sine-wave speech (SWS) is a type of degraded speech comprising time-varying sinusoids typically following the frequencies and amplitudes of the first three formants in a speech signal (Remez et al., 1981, 1994). It is often generated via a multistep process that begins with tracking the energy peaks of formants in natural recordings. Next, a time-varying sinusoid is generated representing each formant, the amplitude of which is replicated. Finally, the three sinusoids representing F1, F2, and F3 of the original natural recordings are combined together to generate SWS (Remez et al., 1981). As it only includes three time-varying sinusoids, there are no harmonics and it does not contain fundamental frequency. However, as long as the dynamic spectral variation is even vaguely preserved, people can perceive it as speech and identify words and sentences in SWS (Benson et al., 2006; Feng et al., 2012; Liebenthal et al., 2003; Nittrouer et al., 2014; Remez et al., 2011, 2018; Rosen and Hui, 2015).

Noise vocoding is another signal-processing technique, generally dividing a speech signal into a certain number of frequency bands, taking the amplitude envelope from each band, and modulating bands of noise based on the amplitude envelope (Newman and Chatterjee, 2013). All amplitude-modulated noise bands are combined to generate noise-vocoded speech. A noise vocoder with a smaller number of filters smear the spectral detail, whereas a vocoder with a greater number of filters loses less informational content of the signal (Rosen and Hui, 2015). Studies with noise vocoding are often conducted to simulate the perception pattern seen in listeners with cochlear implants and the effect of background noise on their performance (Friesen et al., 2001; Fu et al., 1998; Ihlefeld et al., 2010; Newman and Chatterjee, 2013; Schnupp et al., 2011; Shannon et al., 1995). However, when this noise-vocoding technique is applied to SWS, it generates three bands of noise replicating the original three time-varying sinusoids in terms of frequency and amplitude. Here, I used it to minimize the influence of three separate time-varying sinusoids on speech perception, creating noise-vocoded sine-wave speech (NzVocSWS).

The acoustic cues that distinguish speech sounds are often redundant, and listeners adjust their use of these cues depending on the language and the listening situation. Neither F0 nor harmonics are available in SWS, and as a result, perception of pitch may be poor. For example, tone perception of SWS is poor in Mandarin Chinese even by native listeners because SWS only consists of three time-varying sinusoids (Feng et al., 2012). Studies have shown that English listeners rely on an irrelevant, misleading acoustic cue—the periodicity of the lowest sinusoid replicating F1—to identify English intonation in SWS (Remez and Rubin, 1984, 1993). This strategy may be attributed to the dominance region of approximately 400–1000 Hz, which affects the pitch perception (Remez and Rubin, 1984, 1993).

Rosen and Hui (2015) hypothesized that noise vocoding would minimize the effect of focusing on the misleading cue of F1 for voice pitch contour. Their noise-vocoding technique divided the SWS into 33 frequency bands allocated via the Greenwood function (Greenwood, 1990) and resulted in three bands of noise varying in frequency and amplitudes. Their experimental results demonstrated that the intelligibility of Cantonese SWS sentences improved when they were noise-vocoded. They suggested that noise vocoding reduced the effect of F1 in SWS and, thus, led to higher intelligibility (Rosen and Hui, 2015). Rosen and Hui (2015) argued that the spectrotemporal dynamics did not differ significantly between SWS and NzVocSWS, and that the intelligibility of the English stimuli was not significantly different between the two conditions. However, they argued that it was different for Cantonese, a tonal language in which pitch itself can change word meaning. Their findings suggested that the impact of degradation or transformation of speech varies between languages. That is, when formants became less available due to the noise transformation, the slower temporal envelope, which was still available in NzVocSWS, increased intelligibility in Cantonese (Rosen and Hui, 2015).

Based on these findings, Shinohara (2022) examined Japanese listeners' pitch-accent identification in three stimulus conditions: natural recordings (NatRec), SWS, and NzVocSWS. The author predicted that identification accuracy would be at chance level in the SWS condition, as was the case with Cantonese listeners for tone identification in Feng et al. (2012). In SWS, neither F0 nor harmonics exists and, thus, participants would rely on F1 for pitch-accent perception (Remez and Rubin, 1984, 1993). Since F1 is also related to jaw lowering for vowel quality, F1 was predicted to be a misleading voice pitch contour. Additionally, listeners' identification accuracy was expected to increase significantly from the SWS to the NzVocSWS condition due to the reduced salience of the periodicity of the lowest sinusoidal component in NzVocSWS. Degradation of formants through noise vocoding would lead listeners to rely on the temporal envelope. However, no significant difference was found between the identification accuracies for the SWS and

NzVocSWS conditions, and their identification accuracies were barely above the chance level.

Shinohara (2022) suggested that noise vocoding was ineffective because the particular noise-vocoding method used did not sufficiently distort the SWS stimuli. The conventional noise vocoder used by Shinohara (2022) had 33 rectangular synthesis filters allocated between 70 Hz and 10 kHz via the Greenwood function (Greenwood, 1990). This transformation involved extracting the amplitude envelope from each narrow frequency band and modulating the signal with noise within each band. Since this was applied to SWS, the original formants were replicated with noise and there was little spectral information lost. Each formant might still have remained separate and trackable in NzVocSWS, and Japanese listeners may have still relied on noise-vocoded tonal replica of F1 for pitch-accent perception.

The present study used another noise-vocoding technique to reduce the F1 effect and examined whether Japanese speakers relied on the temporal envelope, particularly intensity and duration of vowels for identification. This should lead to a significant increase in the accuracy from the SWS condition to the NzVocSWS condition. Fig. 1 displays the narrowband spectrograms of an example Japanese word [ame] with an HL pitch and an LH pitch as recorded by a native Tokyo Japanese speaker, transformed into SWS, and then noise-vocoded. The NatRec and SWS stimuli were the same in both Shinohara (2022) and the present study; however, a different NzVocSWS approach was applied. As described above, the conventional noise vocoder used by Shinohara (2022) did not sufficiently distort the spectral information of SWS so that separate formants are visible in Fig. 1. In contrast, this study employed a peak-picking noise vocoder to select the eight highest-energy channels out of the 33 analysis filters (Winn, 2021). The peak-picking noise vocoder had triangle-shaped channels, and noise was modulated in a wider frequency band, making the separate formants less visible in Fig. 1. The prediction was that the periodicity of each sinusoid representing a formant would be more difficult to track during the perception of NzVocSWS.

1.4. The present study and hypotheses

The present study examined whether the temporal envelope, as reflected in vowel duration and intensity, correlates to the identification of Japanese pitch accent. The initial objective of this study was to analyze the two-mora Japanese HL vs. LH pitch-accent minimal pairs produced by Tokyo Japanese speakers to confirm that the Japanese pitch-accent minimal pairs are correlated with F0, F1, duration, and intensity of vowels. F1 has not been currently discussed as a secondary acoustic cue for Japanese pitch accent. However, it was included in the analysis to examine potential correlations. For example, F1 was predicted to be significantly correlated with intensity and duration. Across languages, emphasized syllables are often produced with a higher degree of jaw lowering, which often leads to higher intensity and longer duration. Jaw lowering also involves lowering the volume of back cavity in the vocal tract, resulting in higher F1 (Chiba and Kajiyama, 1958; Erickson, 2002; Fant, 1965). This suggests a potential correlation between F1 and Japanese pitch-accent production, as well as positive correlations between F1 and intensity and between F1 and duration (Kawahara et al., 2017; Plag et al., 2011; Schulman, 1989).

However, the F0 and F1 were not expected to show a significant positive or negative correlation because of the following contradictions. First, jaw lowering for articulation is potentially related to F0. Lowering a jaw moves the hyoid bone and thyroid cartilage backwards, slackening the vocal folds, which lowers F0 (Chen et al., 2021; Erickson et al., 2017; Lim et al., 2006; Whalen and Levitt, 1995). Jaw protrusion, on the other hand, pulls forward the hyoid bone and thyroid cartilage, which stretches the vocal folds, producing higher F0 (Chen et al., 2021; Erickson et al., 2017; Lim et al., 2006; Whalen and Levitt, 1995). Thus, open vowels (e.g., /a/) often have lower F0 and higher F1, whereas close vowels (e.g., /i/ and /u/) tend to have higher F0 and lower F1 (Chen et al., 2021; Erickson et al., 2017). Nevertheless, this potential negative correlation between F0 and F1 was expected be offset by the stronger correlation of F0 with the HL vs. LH pitch accents (i.e., high-pitched morae are produced with higher F0 and higher F1). This study thus analyzed the acoustic correlates of Japanese pitch accents and correlations among acoustic cues in production.

After confirming that duration, intensity, and F1 of vowels are acoustic correlates of Japanese pitch-accent production, SWS and NzVocSWS were used to examine Japanese listeners' pitch-accent identification. If listeners rely on F1 for pitch perception in SWS, as shown in previous studies testing the perception of English intonation (Remez and Rubin, 1984, 1993), the F1 effect related to jaw lowering for segments (e.g., open vs. close vowels) may mislead listeners regarding pitch perception. I tested whether Japanese listeners relied on the lowest sinusoidal component (i.e., F1) when perceiving SWS in which F0 was unavailable, and whether their ability to identify pitch-accent minimal pairs was poor. In addition, I examined whether Japanese listeners made use of the secondary acoustic cues for pitch-accent identification, when F0 and F1 were unavailable in NzVocSWS. I predicted that each formant would be more difficult to track because of noise vocoding so that Japanese listeners would not be able to focus on F1. Instead, they would rely on the temporal envelope reflected in the duration and intensity of vowels for identification. If duration and intensity are the acoustic correlates of pitch accents, this change in the acoustic cue availability from SWS to NzVocSWS would lead to a higher identification accuracy. Finally, regardless of the intelligibility of Japanese pitch accent, listeners would be able to discriminate the auditory inputs in the SWS and NzVocSWS conditions to some extent, using any available acoustic properties.

Hence, this study recorded and analyzed two-mora pitch-accent minimal pairs produced by Tokyo Japanese speakers in Experiment 1, and the acoustic cues correlated with the HL and LH pitch accents were investigated. In Experiment 2, identification and discrimination tests were conducted to analyze Japanese listeners' pitch-accent perception. The identification test assessed whether Japanese speakers could judge each target stimulus as an HL or LH pitch-accent minimal-pair word in the three stimulus conditions (NatRec, SWS, and NzVocSWS). Discrimination was subsequently evaluated to ensure that participants perceive auditory differences between the HL and LH words. In the discrimination test, participants were not required to identify HL or LH based on specific acoustic cues; rather, they needed to demonstrate their auditory sensitivity to any acoustic properties. Finally, the acoustic correlates of participants' identification responses were analyzed.

2. Experiment 1: production

2.1. Methods

2.1.1. Participants

Eight Tokyo Japanese speakers (four female and four male) aged 21.9 years on average (standard deviation (SD) = 1.95) were recruited at Waseda University, Tokyo, Japan. They all spent most of their lives in areas where Tokyo Japanese is spoken (Akinaga and Kindaichi, 2014). None of the speakers had a history of language or speech impairment, and their parents were native Japanese speakers. They did not speak any other language in their daily lives except in university classes.

2.1.2. Stimuli, apparatus, and procedure

The 22 minimal pairs (44 words) listed in Appendix A were recorded by the eight Tokyo Japanese speakers. Since the two-mora words were in isolation, final-accented and unaccented words were recorded without distinction, serving as the LH counterparts of the HL pitch accent (Poser, 1984; Vance, 1995; Warner, 1997). Each word was randomly displayed on a screen using the ProRec 2.4 software (Huckvale, 2020), and its production was recorded using a Rode NT2-A microphone with 44,100 16-bit samples per second. Words with a close vowel (/i/ or /u/)

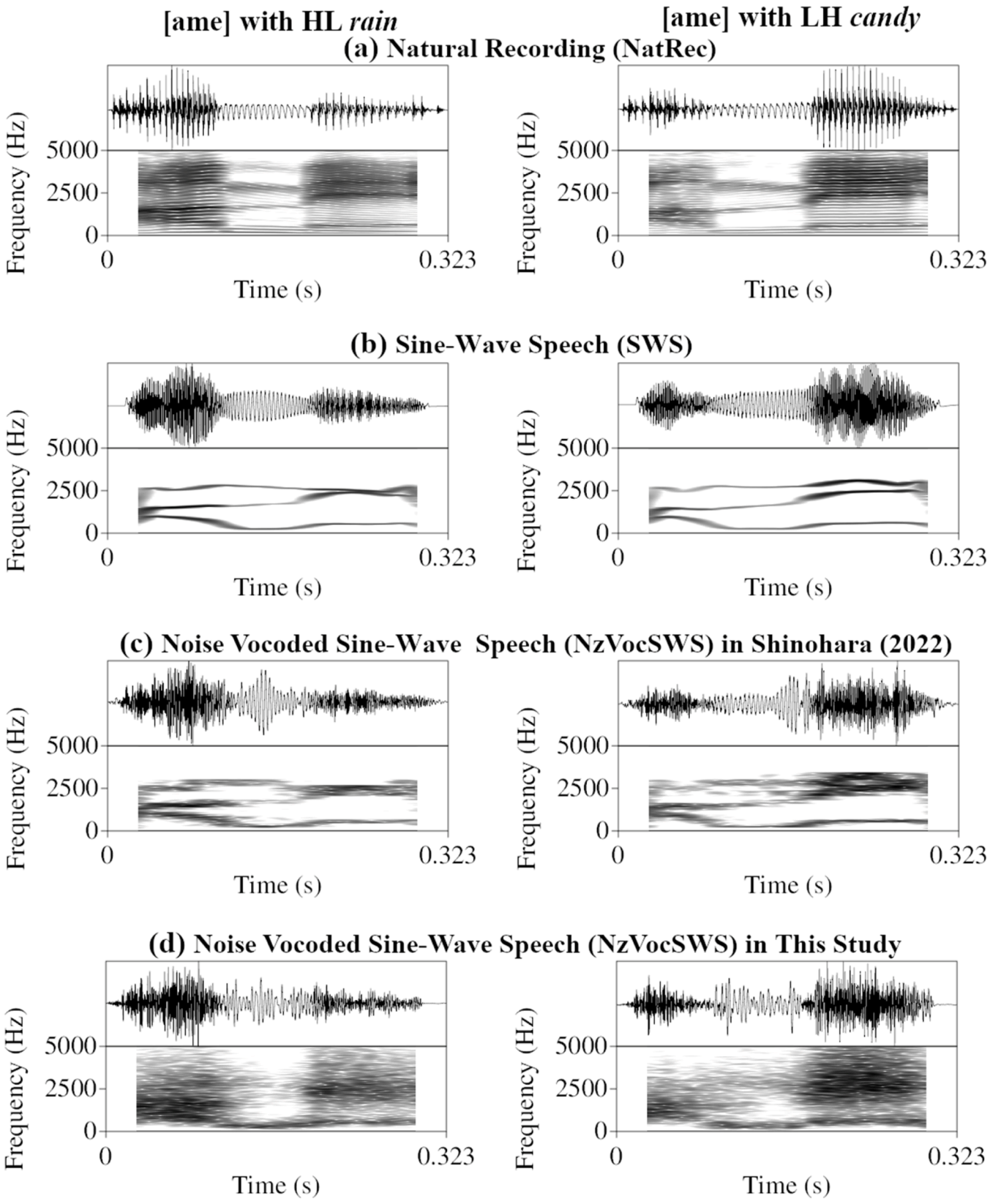

Fig. 1. Waveforms and narrowband spectrograms of the Japanese pitch-accent minimal pair [ame], with high–low (HL) pitch (left) and low–high (LH) pitch (right) in (a) natural recordings (NatRec), (b) sine-wave speech (SWS), (c) noise-vocoded SWS (NzVocSWS) from Shinohara (2022), and (d) the NzVocSWS used in this study.

between voiceless obstruents were not included in the list to avoid devoicing (Kitahara, 2001; Kitahara and Amano, 2001). The naturally recorded word stimuli were the same as those used by Shinohara (2022).

Before conducting acoustic analyses of the NatRec stimuli, word durations were manipulated to be the average value between minimal-pair words produced by each of the eight speakers, and the intensities were normalized across all tokens using the root-mean-squared method in Praat (Boersma and Weenink, 2022). After normalization, each vowel of the 44 two-mora Japanese words was manually annotated as the first or second vowel (V1 or V2, respectively), using TextGrid objects in Praat (Boersma and Weenink, 2022). Next, the acoustic properties, such as F0 (Hz), F1 (Hz), duration (ms), and intensity (dB) of each vowel were measured and extracted automatically using a Praat script. Formant analysis was conducted based on linear predictive coding (LPC) with the respective sex settings, F0 and F1 were measured at the middle point of each vowel in Hertz, and they were converted to Bark. Duration was defined as the time interval between the onset and offset of vowels, and intensity was measured as the mean amplitude of each vowel. For some phonetic contexts, it was difficult to judge a segment boundary (e.g., between vowels in [kau] *keep a pet* vs. *buy*). However, since identical criteria (e.g., drawing a segment boundary in the middle of formant transition from one vowel to another) were used for both the HL and LH words, the criteria did not have a strong effect on the HL and LH word difference. In Praat, the F0 of 12 tokens and F1 of two tokens out of 704 vowels (2 vowels × 44 words × 8 speakers) were unidentified. No outliers were present either in F0 or F1, but one in duration and six in intensity exceeded ±3 SDs from the sample mean. These unidentified tokens and outliers were accordingly excluded from the acoustic analyses.

2.2. Results

Fig. 2 shows the results of the acoustic analysis of the 44 pitch-accent minimal-pair words produced by the eight Tokyo Japanese speakers. There were clear acoustic differences in F0, F1, duration, and intensity between the HL and LH pitch-accent words. As two vowels were included in each word, the difference between the first and second vowels was calculated for each acoustic property by subtracting the acoustic value of the second vowel from that of the first vowel (i.e., V1 – V2). For words with an HL pitch, the first vowel was predicted to have higher F0, F1, and intensities, as well as longer durations than the second vowel (Igarashi, 2015; Kawahara, 2015); thus, the V1 – V2 differences were expected to have positive values. In contrast, the V1 – V2 differences for words with an LH pitch were predicted to be negative, because the second vowel was expected to be produced with higher F0, F1, and intensities and longer durations than the first one (Igarashi, 2015; Kawahara, 2015). After the subtraction, each acoustic value was normalized to each property for statistical analyses. The sample mean of the V1 – V2 difference was subtracted from the individual V1 – V2 difference, and the result was divided by 1 SD. The HL words had more positive values in each parameter than the LH words.

Four linear mixed effects models were constructed including each acoustic cue (F0, F1, duration, and intensity) as the dependent variable and the pitch accent (HL or LH) as the fixed effect for all models. By-speaker and by-minimal pair random intercepts were also included in each model. The linear mixed effects models demonstrated that the F0 differed significantly between Japanese HL and LH pitch-accent words, $\beta = 0.57$, $SE = 0.01$, $t = 54.77$, $p < .001$, as did F1, $\beta = 0.08$, $SE = 0.02$, $t = 4.42$, $p < .001$, duration, $\beta = 0.30$, $SE = 0.04$, $t = 8.26$, $p < .001$, and intensity, $\beta = 0.55$, $SE = 0.03$, $t = 18.65$, $p < .001$. All the results of the statistical analyses are available in Appendix B.

Table 1 displays the results of correlation tests between the normalized (V1 – V2) values in F0, F1, duration, and intensity. Since the Shapiro-Wilk tests demonstrated that the normalized values were not normally distributed in F0, $W = 0.90$, $p < .001$, or F1, $W = 0.91$, $p < .001$; Spearman rank-order correlation tests were performed for all correlations except for the duration and intensity pair. The assumption of normal distribution was not rejected in either duration or intensity ($p > .05$). All other assumptions for the Pearson product-moment correlation test were met for the duration and intensity pair. The results of Spearman's and Pearson's correlation tests demonstrated that all combinations of the four acoustic properties were significantly correlated except for the F0 and F1 pair.

Table 1
Results of Spearman rank-order correlation and Pearson product-moment correlation tests between acoustic cues of F0, F1, duration, and intensity.

	F1	Duration	Intensity
F0 (Bark)	rho = 0.05 $p > .05$	rho = 0.14 $p = .014$	rho = 0.56 $p < .001$
F1 (Bark)		rho = 0.20 $p < .001$	rho = 0.34 $p < .001$
Duration (ms)			$r = 0.22$ $p < .001$

Note: p values were adjusted using the false discovery rate method (Benjamini and Hochberg, 1995). Since F0 and F1 were not normally distributed, Spearman's rank-order correlation tests were conducted for all acoustic pairs except for the duration and intensity pair, for which the Pearson's correlation test was performed.

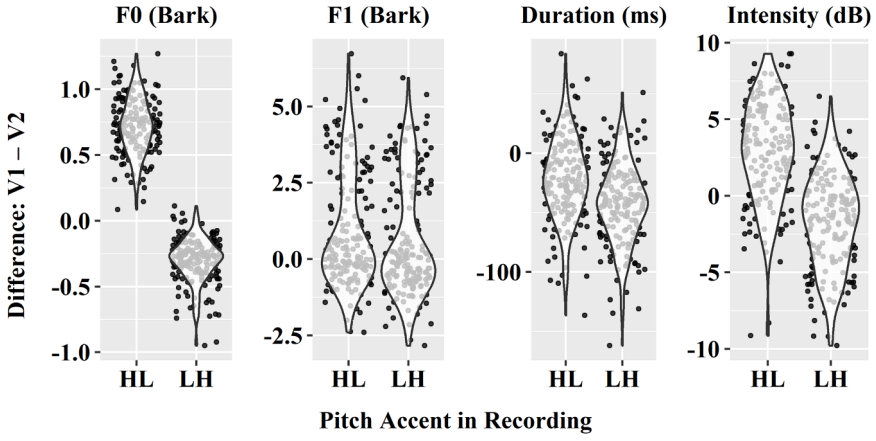

Fig. 2. Acoustic differences between Japanese two-mora pitch-accent minimal-pair words (HL vs. LH) in F0 (Bark), F1 (Bark), duration (ms), and intensity (dB). The difference between the first and second vowels in each word was calculated by subtracting the acoustic value of the second vowel from that of the corresponding first vowel (V1 – V2).

3. Experiment 2: perception

3.1. Methods

3.1.1. Participants

Twenty Tokyo Japanese speakers (10 female and 10 male) aged 20.7 years on average (SD = 1.93) participated in the identification and discrimination tests at Waseda University, Tokyo, Japan. None of them had participated in the previous Japanese pitch-accent perception study (Shinohara, 2022) or had a history of language, speech, or hearing impairments. They all spent most of their lives in areas where the Tokyo dialect was spoken (Akinaga and Kindaichi, 2014), and their parents were native Japanese speakers. They did not speak any other language in their daily lives except in university classes.

3.1.2. Stimulus conditions

NatRec. The 22 two-mora Japanese pitch-accent minimal pairs (44 words; see Appendix A) recorded by the eight Tokyo Japanese speakers in Experiment 1 were used for the perception tests. As described in Experiment 1, the duration was manipulated to be the mean duration of each minimal pair, and the intensity was normalized across all tokens using the root-mean-squared method in Praat (Boersma and Weenink, 2022). Of the 22 minimal pairs (44 words), two were used for a practice session and the remaining 20 were used for the identification and discrimination tests.

SWS. The SWS stimuli were the same as those used in Shinohara (2022). All 44 naturally recorded words produced by the eight speakers were transformed into SWS using a Praat script (Darwin, 2005) with default settings. The F1, F2, and F3 frequencies were tracked every 10 ms with the respective sex settings, and the amplitude tracks were low-pass filtered at 50 Hz. After individual sinusoids representing F1, F2, and F3 of the NatRec stimuli were generated, the three sinusoids were combined to construct the SWS version at the final stage of the Praat script (Darwin, 2005).

NzVocSWS. This study used the Praat script written by Winn (2021), which was the same as that used in Shinohara (2022) except that the SWS was noise-vocoded differently. Shinohara (2022) used 33 flat-spectrum rectangular channels of a conventional noise vocoder with a frequency range of 70 Hz to 10 kHz, whereas this study used a peak-picking noise vocoder with the same frequency range. The eight highest-energy channels were selected from the 33 analysis filters after pre-emphasis (increasing spectral tilt approximately by +6 dB/octave), and these pre-emphasized filters were de-emphasized after selection. The width of the peaked synthesis filter, which simulates the spread of the cochlear excitation, was set to 6 dB/mm using the Praat script (Winn, 2021). Thus, the formants in this NzVocSWS condition were more distorted than those used in Shinohara (2022). Indeed, compared to Shinohara (2022), each formant could hardly be observed in the spectrogram (Fig. 1). Note that the default setting of the 600 Hz envelope cutoff filter was unintentionally used in this noise vocoder, but the amplitude tracks were low-pass filtered at 50 Hz when the NatRec stimuli were transformed into SWS. The other parameters for noise vocoding were the same as those used by Shinohara (2022).

3.1.3. Procedure

Identification test. All participants completed the identification and discrimination tests for all three stimulus conditions (NatRec, SWS, and NzVocSWS) in a soundproof booth at Waseda University, Tokyo, Japan. During the identification test, participants randomly heard a Japanese word (e.g., [ame] with HL) through a pair of headphones (Sennheiser HD 280 Pro). A minimal pair was displayed on the screen with both Japanese kanji (Japanese writing system with Chinese characters) and hiragana (Japanese mora-based letters) (e.g., 雨(あめ) vs. 飴(あめ)). Participants were instructed to click on the word they thought they had heard (e.g., 雨(あめ)). They did not receive any feedback on their responses.

Before each of the NatRec, SWS, and NzVocSWS identification tests, the participants underwent a three-trial practice session using two pitch-accent minimal pairs. After this practice session, they participated in the identification test with the remaining 20 minimal pairs (40 words). Half took the NatRec test first, followed by the SWS and NzVocSWS tests, and the other half took the NatRec test first, followed by the NzVocSWS and SWS tests. Owing to the large number of stimuli, only two pairs (four words) produced by the eight speakers were used for each participant. Each token was randomly repeated three times, resulting in 288 trials per participant (4 words × 8 speakers × 3 repetitions × 3 conditions). Since each of the 20 participants had two minimal pairs, all 20 minimal pairs were used twice each for the identification test.

Discrimination test. During the discrimination test, participants heard three stimuli (e.g., [ame] with HL, HL, and LH). Three numbers (1, 2, and 3) were displayed on the screen, and the participants clicked on the number representing the sound that was different from the other two. The participants were given a practice session with two pairs and took the discrimination test with the remaining 20 pairs (40 words). Owing to the large number of stimuli, two pairs (four words) produced by eight speakers with three odd stimulus positions and three conditions were used per participant, resulting in 288 tokens. As in the identification task, two minimal pairs were used per participant, such that all 20 minimal pairs were covered twice by all 20 participants. Half of the participants completed the NatRec condition first, followed by the SWS and NzVocSWS conditions, and the other half completed the NatRec first, followed by the NzVocSWS and SWS conditions.

Acoustic correlates in identification. To analyze the effect of each acoustic property on Japanese listeners' pitch-accent identification, the F1 (Hz), duration (ms), and intensity (dB) were measured for each stimulus in the NatRec, SWS, and NzVocSWS conditions. Because F0 was absent in the SWS and NzVocSWS conditions, it was excluded from this analysis. Using the same annotation files as Experiment 1 (i.e., TextGrid files in Praat; Boersma and Weenink, 2022), the acoustic values were extracted from the first and second vowels (V1 and V2) in each word. Formant analysis was conducted with the respective sex settings as performed in Experiment 1, F1 was measured from the middle point of each vowel in Hertz, and it was converted to Bark. Duration was measured as the time interval between the onset and offset of vowels. Since the same TextGrid files were used in Praat, duration was identical for each word across stimulus conditions. Intensity was considered as the mean amplitude of each vowel. Table 2 shows the number of vowels with unidentified acoustic values and outliers over ±3 SD from the sample mean. Note that though F1 was supposed to be less audible in the NzVocSWS condition, it was still possible to track and measure for each vowel in Praat (Boersma and Weenink, 2022). Those acoustic values that were unidentified or outliers were excluded from the analyses. Finally, as in Experiment 1, each acoustic value was normalized to each property for statistical analyses. The sample mean of the V1 – V2 difference was subtracted from the individual V1 – V2 difference, and the result was divided by 1 SD.

Table 2
The numbers of vowels with unidentified acoustic values of F1 (Hz), duration (ms), and intensity (dB) in Praat and outliers deviating over ±3 SDs from the sample mean.

	F1	Duration	Intensity
NatRec (N = 704)	2, 0	0, 1	0, 6
SWS (N = 704)	2, 0	0, 1	0, 8
NzVocSWS (N = 704)	2, 0	0, 1	0, 7

Note: The first and second values in each cell respectively represent the number of vowels with unidentified acoustic values and the number of outliers. Since the same TextGrid was used for all three conditions, duration of each vowel was the same across conditions.

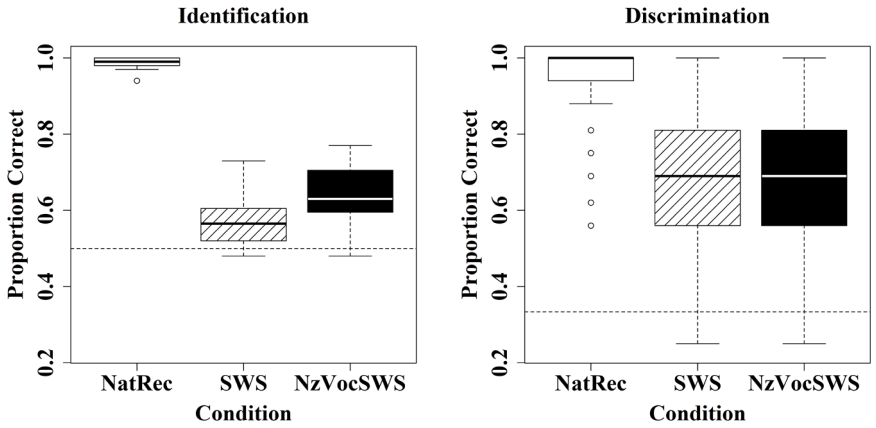

Fig. 3. Identification and discrimination accuracies of Japanese pitch-accent minimal pairs in three conditions: natural recording (NatRec), sine-wave speech (SWS), and noise-vocoded sine-wave speech (NzVocSWS). The horizontal dashed lines represent the chance level in the identification and discrimination tests.

3.2. Results

3.2.1. Identification and discrimination accuracies

Fig. 3 shows the identification and discrimination accuracies of Japanese pitch-accent minimal pairs in the NatRec, SWS, and NzVocSWS conditions. Identification accuracy in the NatRec condition was at the ceiling, whereas accuracies in the degraded speech conditions (SWS and NzVocSWS) were lower than in the NatRec condition. Accuracy in the SWS condition was only slightly above the chance level (i.e., 50%), while accuracy in the NzVocSWS condition was higher than SWS. Regarding discrimination, accuracy in the NatRec condition was also at the ceiling, and those in the degraded speech conditions (SWS and NzVocSWS) were lower than in NatRec. Accuracy in both SWS and NzVocSWS was higher than the chance level (i.e., 33.3%), with no significant difference between the two conditions.

To confirm these results, a logistic mixed effects model was constructed for each of the identification and discrimination analyses, using the correct or incorrect binomial responses as the dependent variable and the stimulus condition (NatRec, SWS, and NzVocSWS) as the fixed effect. Effective coding was used for each analysis, and the orthogonal contrasts were set for the stimulus condition (NatRec vs. SWS & NzVocSWS; SWS vs. NzVocSWS). Random effects were carefully constructed by checking convergence failure, singular fits, and variance for random slopes. Model comparisons were also conducted by referring to the Akaike Information Criteria (AIC). For the logistic mixed effects model of the identification test, by-minimal pair random intercepts were included, but other random effects did not significantly improve the model fit. For the model of discrimination, by-participant, by-speaker, by-minimal pair (e.g., [ame], [ɕiro]), and by-stimulus pair (e.g., HL-HL-LH, LH-LH-HL) random intercepts were included. Random slopes for condition were included for participants.

The results of the identification analysis demonstrated that the accuracy in degraded speech conditions (SWS & NzVocSWS) was significantly lower than in the NatRec condition, $\beta = 1.29$, $SE = 0.07$, $z = 19.31$, $p < .001$. However, identification accuracy in the NzVocSWS condition was significantly higher than in the SWS condition, $\beta = -0.15$, $SE = 0.03$, $z = -4.35$, $p < .001$, suggesting that noise vocoding increased the intelligibility of the SWS pitch-accent minimal pairs. The logistic mixed effects model for the discrimination analysis demonstrated that the accuracy of degraded speech conditions (SWS & NzVocSWS) was significantly lower than in the NatRec condition, $\beta = 0.88$, $SE = 0.08$, $z = 10.90$, $p < .001$, and there was no significant difference in the accuracy between the SWS and NzVocSWS conditions, $\beta = 0.01$, $SE = 0.05$, $z = 0.24$, $p > .05$.

Together, these results suggest that participants were able to hear the acoustic difference between HL and LH pitch accents in both SWS and NzVocSWS conditions to some extent, but they could not identify the words in the SWS condition. The identification accuracy in the NzVocSWS condition increased from that in the SWS condition. All the results of the statistical analyses are available in Appendix C (Table C1–2).

3.2.2. Acoustic correlates in pitch-accent identification

Fig. 4 displays Japanese listeners' pitch-accent identification responses (HL vs. LH words) and the acoustic differences between the two vowels in each word (V1 – V2) in terms of F1, intensity, and duration for three stimulus conditions (NatRec, SWS, and NzVocSWS). The figure illustrates how the identification responses are associated with acoustic changes from the first to the second vowel in each condition. The subtraction (V1 – V2) and normalization procedures were the same as those used in Experiment 1.

F1 showed a stronger association with identification responses in the degraded speech conditions (SWS & NzVocSWS) than in the NatRec condition. In addition, the F1 effect on participants' identification responses was higher in the SWS condition than in NzVocSWS. This means that the identification responses were most clearly separated by the F1 difference in the SWS condition among the three conditions. On the contrary, duration and intensity showed stronger associations with identification responses in the NatRec condition than in the degraded speech conditions (SWS & NzVocSWS). The effects of duration and intensity on identification responses were higher in the NzVocSWS condition than in the SWS condition, suggesting that Japanese listeners rely more on the duration and intensity cues in the NzVocSWS condition than in the SWS condition.

A logistic mixed effects model was constructed for each acoustic property with the participants' binomial identification responses (HL vs. LH words) as the dependent variable, regardless of their correctness. For example, the logistic mixed effects model used to test the F1 effect included the stimulus condition, normalized F1 (V1 – V2), and their interaction as fixed effects. The logistic mixed effects model testing the duration effect included the stimulus condition, normalized duration (V1 – V2), and their interaction as fixed effects. Similarly, the intensity model included the stimulus condition, normalized intensity (V1 – V2), and their interaction as fixed effects. Orthogonal contrasts were set for

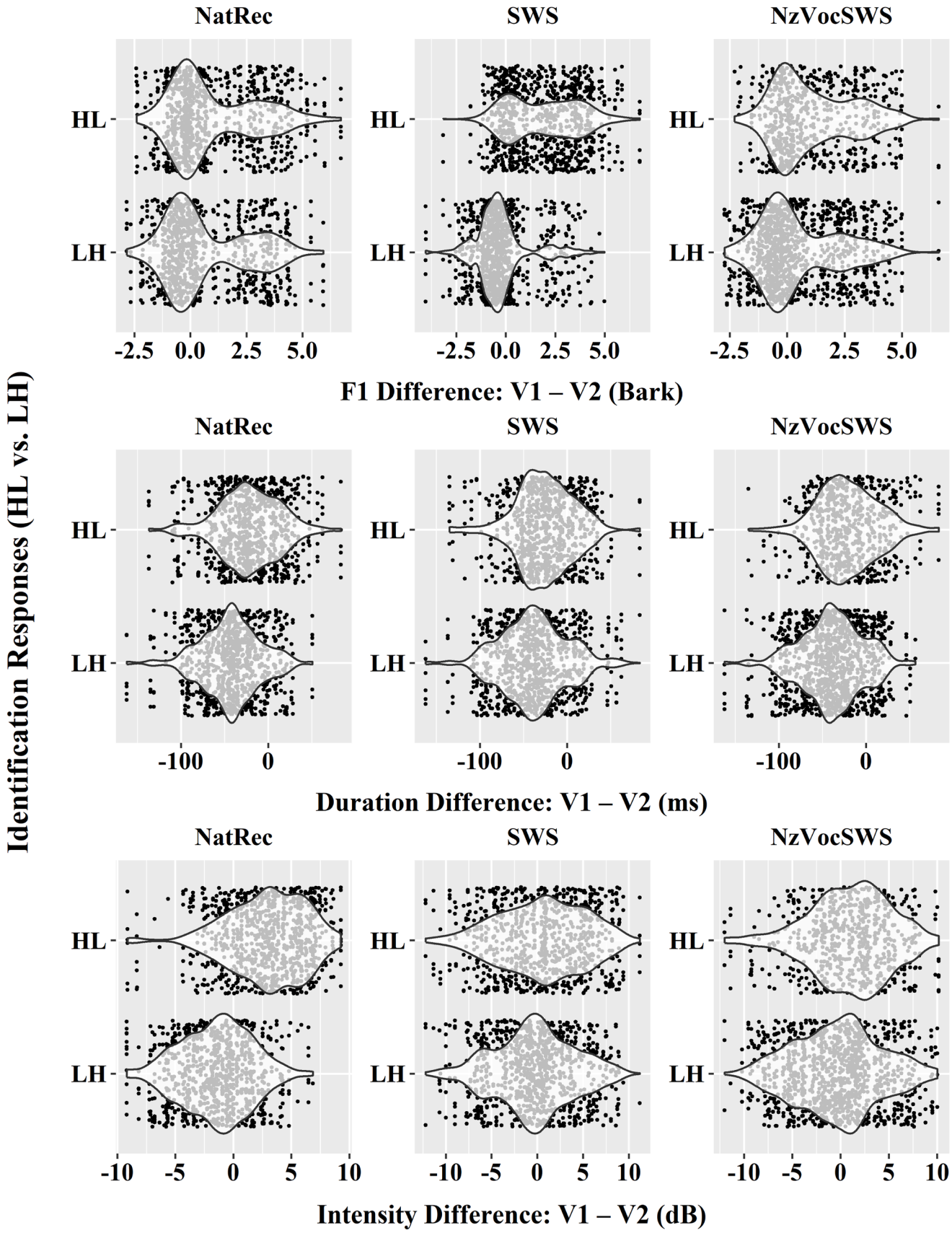

Fig. 4. Japanese listeners' HL vs. LH pitch-accent identification responses in three stimulus conditions (NatRec, SWS, and NzVocSWS) and the difference in F1, duration, and intensity between the first and second vowels of each minimal-pair stimulus. Each acoustic value was calculated by subtracting the value of the second vowel (V2) from that of the first vowel (V1) for F1 (Bark), duration (ms), and intensity (dB).

the stimulus condition to test the effect differences between NatRec and degraded speech (SWS & NzVocSWS) and between the SWS and NzVocSWS conditions. By-participant, by-speaker and by-minimal pair random intercepts were included in all mixed effects models. By-minimal pair random slopes for condition were included only for the intensity model. Model comparison based on AIC strongly favored a model including by-minimal pair random slopes for condition over a model including by-participant slopes. The selected model showed substantial variability in condition effects across minimal pairs.

The logistic mixed effects model including the fixed effect of F1 demonstrated a significant effect of F1 (V1 – V2), $\beta = 1.42$, $SE = 0.09$, $z = 15.16$, $p < .001$, suggesting that Japanese listeners' identification responses (HL vs. LH) were associated with the F1 (V1 – V2) (e.g., Japanese listeners tend to identify HL words when perceiving stimuli having a more positive F1 difference). The significant effects of condition demonstrated that listeners identified fewer HL words in the degraded speech conditions (SWS & NzVocSWS) than in the NatRec condition regardless of its F1 values, $\beta = 0.11$ $SE = 0.02$, $z = 5.49$, $p < .001$, and fewer HL words in the NzVocSWS condition than in the SWS condition, $\beta = 0.24$, $SE = 0.04$, $z = 6.32$, $p < .001$. The interaction between the F1 (V1 – V2) and stimulus condition demonstrated that the F1 effect was larger in the degraded speech conditions (SWS & NzVocSWS) than in the NatRec condition, $\beta = -0.29$, $SE = 0.02$, $z = -13.31$, $p < .001$. In addition, the F1 effect on identification responses was significantly larger in the SWS condition than in the NzVocSWS condition, $\beta = 0.60$, $SE = 0.05$, $z = 13.19$, $p < .001$. These results suggest that Japanese listeners relied more on the F1 cue when identifying words in the SWS condition than in the NatRec and the NzVocSWS conditions. Note that F1 had a relatively large effect on Japanese listeners' identification responses in the SWS condition, but the identification accuracy in the same condition was barely above chance (Fig. 3).

The logistic mixed effects model including the fixed effect of duration demonstrated that there was a significant effect of duration (V1 – V2) on the identification responses across stimulus conditions, $\beta = 0.91$, $SE = 0.05$, $z = 19.90$, $p < .001$. The condition effects were both significant between NatRec and the degraded speech conditions (SWS & NzVocSWS), $\beta = 0.13$, $SE = 0.02$, $z = 6.26$, $p < .001$, and between SWS and NzVocSWS, $\beta = 0.29$, $SE = 0.04$, $z = 7.90$, $p < .001$. The significant interaction between duration (V1 – V2) and stimulus condition demonstrated that the effect of the duration cue in the degraded speech conditions (SWS & NzVocSWS) was significantly smaller than in the NatRec condition, $\beta = 0.07$, $SE = 0.02$, $z = 3.11$, $p < .01$. In addition, the effect of duration was significantly larger in the NzVocSWS condition than in the SWS condition, $\beta = -0.10$, $SE = 0.04$, $z = -2.68$, $p < .01$. These results suggest that Japanese listeners rely more on duration cue when perceiving stimuli in the NzVocSWS condition than in the SWS condition.

Similarly, the logistic mixed effects model including the intensity effect demonstrated that intensity (V1 – V2) had a significant effect on the participants' identification responses across all conditions, $\beta = 1.25$, $SE = 0.05$, $z = 24.16$, $p < .001$. The condition effects were not significant in either the contrasts between NatRec and the degraded speech conditions (SWS & NzVocSWS), $\beta = 0.03$, $SE = 0.15$, $z = 0.20$, $p > .05$, or between SWS and NzVocSWS, $\beta = 0.17$, $SE = 0.22$, $z = 0.76$, $p > .05$. The significant interaction between the intensity and stimulus condition demonstrated that Japanese listeners' identification responses were less associated with the intensity cue in degraded speech conditions (SWS & NzVocSWS) than in the NatRec condition, $\beta = 0.71$, $SE = 0.04$, $z = 17.10$, $p < .001$, and the identification responses were more associated in the NzVocSWS condition than in the SWS condition, $\beta = -0.12$, $SE = 0.05$, $z = -2.77$, $p < .01$. Note that the intensity model without by-minimal pair random slopes for condition (i.e., the same random-effects structure as the F1 and the duration models) showed the only difference in the effect of condition (SWS vs. NzVocSWS) which was significant ($p < .001$).

Together, these results suggest that Japanese listeners' reliance on F1 in the SWS condition was significantly reduced by noise vocoding and that they relied more on the secondary acoustic cues of intensity and duration of vowels in the NzVocSWS condition than in the SWS condition. All the results of the statistical analyses for the effects of F1, duration, and intensity on the participants' identification responses are available in Appendix C (Table C3–5).

4. Discussion

This study examined three hypotheses related to Japanese-speaking participants' production and perception of pitch-accent minimal-pair words. First, I predicted that the production of HL and LH pitch accents would be correlated with F0, F1, duration, and intensity of vowels. Second, noise vocoding was expected to significantly increase Japanese listeners' pitch-accent identification of SWS because it would reduce the effect of a misleading cue for voice pitch contour (i.e., the lowest sinusoidal component, F1) (Feng et al., 2012; Remez and Rubin, 1984, 1993; Rosen and Hui, 2015). Finally, I predicted that when both F0 and formant information were unavailable in NzVocSWS, Japanese listeners would rely on the temporal envelope, as reflected in changes in intensity and duration from one vowel to another for identification responses. The results of this study supported all of these hypotheses.

4.1. Production

The primary finding of this study was that F0, F1, duration, and intensity were acoustic correlates of Japanese pitch-accent production. The high-pitched morae (i.e., the H mora) in HL and LH pitch-accent minimal-pair words were produced with high F0. The higher F1, greater intensity, and longer duration of the H morae were likely due to jaw lowering for emphasis (Kawahara et al., 2017; Plag et al., 2011; Schulman, 1989). The results also demonstrated that these four acoustic properties were correlated with each other, except for the F0 and F1 pair. The correlation between F0 and F1 was not significant, possibly due to the conflicting effects. While jaw lowering for articulation (e.g., for open vowels) primarily raises F1 but may lower F0, jaw lowering for emphasis on the H mora in HL and LH pitch accents primarily raises F0 and also F1.

4.2. Perception of sine-wave speech and noise-vocoded sine-wave speech

In previous studies, although vowel intensity was correlated with pitch accents in Japanese (Cutler and Otake, 1999; Weitzman, 1970), there has been considerable discussion about whether duration serves as a secondary acoustic cue for pitch accents (Cutler and Otake, 1999; Igarashi, 2015; Kawahara, 2015; Levi, 2005; Sugiyama, 2017, 2022). This study confirmed that the temporal envelope functions as a secondary acoustic cue for pitch-accent identification.

In SWS, neither F0 nor harmonics are available, but listeners can still perceive three sinusoids and their temporal envelope. Japanese listeners were expected to focus on the lowest sinusoidal component (F1) for pitch perception, as English listeners did for identifying intonation (Feng et al., 2012; Remez and Rubin, 1984, 1993; Rosen and Hui, 2015). Since F1 is strongly associated with jaw lowering (i.e., open vowels have

higher F1, while close vowels have lower F1), Japanese listeners' pitch-accent identification was predicted to be barely above chance level in SWS due to this misleading cue. The results demonstrated that Japanese speakers relied on F1 changes from the first to the second vowel to identify Japanese HL vs. LH pitch-accent words in SWS, and their identification accuracy was only slightly above chance, as predicted.

In contrast, noise vocoding was expected to reduce the F1 effect on pitch-accent perception in SWS. As a result, Japanese listeners would rely on the temporal envelope reflected in the changes in duration and intensity from one vowel to another. If duration and intensity are the secondary acoustic cues in Japanese pitch-accent perception, identification accuracy was predicted to increase, which was demonstrated by the results. The periodicities of the lowest and other sinusoidal components were weakened, and Japanese listeners relied more on the temporal envelope in the NzVocSWS condition than in the SWS condition, leading to higher identification accuracy. Rosen and Hui (2015) found a similar result for the Cantonese language. They explained that noise vocoding weakened the periodicity of the lowest sinusoidal component and cohered the spectral components. As a result, the lowest component was difficult to perceive as a distinct perceptual feature signaling a melodic contour. This study supports these explanations and confirms that Japanese speakers rely on the temporal envelope in pitch-accent perception.

One interesting finding is that the signal degradation effect was not consistent across the identification and discrimination tests. The identification accuracy in SWS was barely above chance (50%), but the discrimination accuracy was higher than the chance level (33.3%). Furthermore, the identification accuracy significantly increased from SWS to NzVocSWS, but such an increase did not occur in the discrimination accuracy. These results suggest that Japanese listeners were able to hear acoustic differences between the HL and LH words in SWS but they were not able to identify the words correctly. In addition, Japanese listeners were still able to hear acoustic differences in the NzVocSWS condition although the spectral information was more degraded. Because the periodicity of the lowest sinusoidal component (i.e., F1) affecting pitch perception was weakened, Japanese listeners were able to identify the HL and LH words to some extent in NzVocSWS, relying on the acoustic cues related to the temporal envelope.

4.3. Limitations

A remaining question is the cause of the difference in the results obtained by Shinohara (2022) and in this study. Shinohara (2022) found no significant difference in identification accuracy between the SWS and NzVocSWS conditions, whereas this study demonstrated a significant difference. Notably, the previous study used a more conventional noise vocoder with 33 square-shaped analysis and synthesis filters, whereas this study used a peak-picking noise vocoder with 33 analysis and eight synthesis filters (Winn, 2021). Comparing the narrowband spectrograms of Shinohara (2022) with those of this study in Fig. 1, the formants in the NzVocSWS condition in this study were less visible. The 33 rectangular filters used in the previous study modulated noise in a narrow frequency range, whereas each of the eight channels of the peak-picking noise vocoder used in this study was wide and overlapped with the other channels. Given the wider frequency range of noise, participants may not have been able to clearly track each formant during perception. This may have forced Japanese listeners to rely on the temporal envelope instead. Because both the number and shape of the synthesis filters used by Shinohara (2022) were different from those used in this study, it was impossible to investigate the effect of each factor on the identification result. In order to make a direct comparison with the results of Shinohara (2022), it may be useful to simply reduce the number of analysis and synthesis filters in the conventional noise vocoder used in the previous study. Thus, further research is necessary to investigate the effects of the number and shape of noise vocoding filters.

In addition, this study was subject to notable limitations. First, the use of linear predictive coding (LPC) analysis can introduce bias when measuring formant frequencies in the presence of F0 variability. In many cases, strong harmonics near the resonance peaks are measured as formants, and F0 can influence formant tracking in LPC analysis (Whalen et al., 2022). Second, although F1 was found to be an acoustic correlate of Japanese HL vs. LH pitch-accent minimal pairs in production, the present study did not fully examine the role of F1 in perception. As described in the Introduction, since people rely on F1 to identify vowels (e.g., open vowels have high F1; close vowels have low F1), the F1 correlate for pitch accent (e.g., jaw lowering related to emphasis) may have been offset in perception. However, F1 may still serve as a substitute cue even in perception. It may be observed under conditions where only one minimal pair is used with the same segments and with the HL vs. LH pitch-accent contrast. In future research, investigation of all acoustic properties' effects for Japanese-pitch accent production and perception is needed.

4.4. Conclusion

In conclusion, the fundamental frequency is the primary acoustic cue for Japanese pitch-accent production and perception, whereas intensity and duration are secondary acoustic cues. When F0 is unavailable in SWS, Japanese listeners tend to perceive F1 as a substitute for F0 in pitch-accent recognition. When both F0 and F1 are unavailable in NzVocSWS, they rely upon secondary acoustic cues (duration and intensity), resulting in a significant increase in identification accuracy from SWS to NzVocSWS. This identification recovery may apply to the examination of cross-linguistic perception and identification of listeners' difficulties in pitch-accent perception.

Ethics approval

This study was approved by the Ethics Review Committee at Waseda University (Tokyo, Japan). Informed consent was obtained from all participants in this study.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by JSPS KAKENHI Grant No 23K00490, Waseda University Grant for Special Research Projects No 2022C-074, and WASEDA Research Acceleration Program for Early-Stage Principal Investigators. I am grateful to Mr. Satsuki Kurokawa, who helped collect the data, and to Dr. Douglas Whalen, Dr. Valerie Shafer, Dr. Kevin Roon, Dr. Matthew Winn, Dr. Katarina Antolovic, and Prof. Kate Elwood, for their helpful comments.

Appendix A. Tokyo Japanese pitch-accent minimal-pair list

Minimal Pairs		Minimal Pairs	
HL	**LH**	**HL**	**LH**
雨 (あめ) [ame] *rain*	飴 (あめ) [ame] *candy*	二時 (にじ) [niʑi] *two o'clock*	虹 (にじ) [niʑi] *rainbow*
腿 (もも) [momo] *thigh*	桃 (もも) [momo] *peach*	朝 (あさ) [asa] *morning*	麻 (あさ) [asa] *hemp*
降る (ふる) [huru] *fall*	振る (ふる) [huru] *shake*	組む (くむ) [kumu] *team up*	汲む (くむ) [kumu] *scoop out*
飼う (かう) [kau] *keep (a pet)*	買う (かう) [kau] *buy*	春 (はる) [haru] *spring (season)*	張る (はる) [haru] *stretch*
神 (かみ) [kami] *God*	紙 (かみ) [kami] *paper*	病む (やむ) [jamu] *fall ill*	止む (やむ) [jamu] *stop (raining)*
白 (しろ) [ɕiro] *white*	城 (しろ) [ɕiro] *castle*	練る (ねる) [neru] *knead*	寝る (ねる) [neru] *sleep*
夜 (よる) [joru] *night*	寄る (よる) [joru] *drop by*	隅 (すみ) [sumi] *corner*	墨 (すみ) [sumi] *ink*
狩り (かり) [kari] *hunting*	借り (かり) [kari] *debt*	膿む (うむ) [umu] *fester*	産む (うむ) [umu] *give birth*
赤 (あか) [aka] *red*	垢 (あか) [aka] *dirt*	切る (きる) [kiru] *cut*	着る (きる) [kiru] *wear*
汁 (しる) [ɕiru] *soup*	知る (しる) [ɕiru] *know*	維持 (いじ) [iʑi] *maintenance*	意地 (いじ) [iʑi] *pride*
* 琴 (こと) [koto] *Japanese harp*	* 事 (こと) [koto] *thing*	* 経る (へる) [heru] *pass*	*減る (へる) [heru] *decrease*

* minimal-pair words used in the practice sessions.
Note: HL indicates high–low pitch and LH indicates low–high pitch for two-mora words. This list is the same as that used in Shinohara (2022).

Appendix B. Statistical Analysis Reports for Experiment 1

Tables B1, B2, B3, B4

Table B1
Linear mixed effects model results for normalized F0 (V1 minus V2) in the Japanese HL vs. LH pitch-accent word production.

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*	
(Intercept)	0.17	0.03	6.22	< .001	***
Pitch accent (HL vs. LH)	0.57	0.01	54.77	< .001	***

*$p < .05$, **$p < .01$, ***$p < .001$.

Table B2
Linear mixed effects model results for normalized F1 (V1 minus V2) in the Japanese HL vs. LH pitch-accent word production.

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*	
(Intercept)	0.03	0.21	0.12	> .05	
Pitch accent (HL vs. LH)	0.08	0.02	4.42	< .001	***

*$p < .05$, **$p < .01$, ***$p < .001$.

Table B3
Linear mixed effects model results for normalized duration (V1 minus V2) in the Japanese HL vs. LH pitch-accent word production.

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*	
(Intercept)	−0.003	0.20	−0.02	> .05	
Pitch accent (HL vs. LH)	0.30	0.04	8.26	< .001	***

*$p < .05$, **$p < .01$, ***$p < .001$.

Table B4
Linear mixed effects model results for normalized intensity (V1 minus V2) in the Japanese HL vs. LH pitch-accent word production.

Fixed effect (Contrast)	Estimate	*SE*	*t*	*p*	
(Intercept)	0.10	0.14	0.73	> .05	
Pitch accent (HL vs. LH)	0.55	0.03	18.65	< .001	***

*$p < .05$, **$p < .01$, ***$p < .001$.

Appendix C. Statistical Analysis Reports for Experiment 2

Tables C1, C2, C3, C4, C5

Table C1
Logistic mixed effects model results for the HL vs. LH word identification accuracy in the three stimulus conditions (NatRec, SWS, and NzVocSWS).

Fixed effect (Contrast)	Estimate	*SE*	*z*	*p*	
(Intercept)	1.74	0.09	19.98	< .001	***
Condition (NatRec vs. SWS & NzVocSWS)	1.29	0.07	19.31	< .001	***
Condition (SWS vs. NzVocSWS)	−0.15	0.03	−4.35	< .001	***

*$p < .05$, **$p < .01$, ***$p < .001$.

Table C2
Logistic mixed effects model results for the HL vs. LH word discrimination accuracy in the three stimulus conditions (NatRec, SWS, and NzVocSWS).

Fixed effect (Contrast)	Estimate	*SE*	*z*	*p*	
(Intercept)	1.70	0.30	5.71	< .001	***
Condition (NatRec vs. SWS & NzVocSWS)	0.88	0.08	10.90	< .001	***
Condition (SWS vs. NzVocSWS)	0.01	0.05	0.24	> .05	

*$p < .05$, **$p < .01$, ***$p < .001$.

Table C3
Logistic mixed effects model results for the HL vs. LH word identification responses, with the fixed effects of normalized F1 (V1 minus V2), stimulus condition (NatRec, SWS, and NzVocSWS), and their interaction.

Fixed effect (Contrast)	Estimate	*SE*	*z*	*p*	
(Intercept)	−0.27	0.22	−1.26	> .05	
F1	1.42	0.09	15.16	< .001	***
Condition (NatRec vs. SWS & NzVocSWS)	0.11	0.02	5.49	< .001	***
Condition (SWS vs. NzVocSWS)	0.24	0.04	6.32	< .001	***
F1 & Condition (NatRec vs. SWS & NzVocSWS)	−0.29	0.02	−13.31	< .001	***
F1 & Condition (SWS vs. NzVocSWS)	0.60	0.05	13.19	< .001	***

*$p < .05$, **$p < .01$, ***$p < .001$.

Table C4
Logistic mixed effects model results for the HL vs. LH identification responses, with the fixed effects of normalized duration (V1 minus V2), stimulus condition (NatRec, SWS, and NzVocSWS), and their interaction.

Fixed effect (Contrast)	Estimate	*SE*	*z*	*p*	
(Intercept)	−0.27	0.22	−1.19	> .05	
Duration	0.91	0.05	19.90	< .001	***
Condition (NatRec vs. SWS & NzVocSWS)	0.13	0.02	6.26	< .001	***
Condition (SWS vs. NzVocSWS)	0.29	0.04	7.90	< .001	***
Duration & Condition (NatRec vs. SWS & NzVocSWS)	0.07	0.02	3.11	< .01	**
Duration & Condition (SWS vs. NzVocSWS)	−0.10	0.04	−2.68	< .01	**

*$p < .05$, **$p < .01$, ***$p < .001$.

Table C5
Logistic mixed effects model results for the HL vs. LH identification responses, with the fixed effects of normalized intensity (V1 minus V2), stimulus condition (NatRec, SWS, and NzVocSWS), and their interaction.

Fixed effect (Contrast)	Estimate	*SE*	*z*	*p*	
(Intercept)	−0.46	0.26	−1.79	.074	
Intensity	1.25	0.05	24.16	< .001	***
Condition (NatRec vs. SWS & NzVocSWS)	0.03	0.15	0.20	> .05	
Condition (SWS vs. NzVocSWS)	0.17	0.22	0.76	> .05	
Intensity & Condition (NatRec vs. SWS & NzVocSWS)	0.71	0.04	17.10	< .001	***
Intensity & Condition (SWS vs. NzVocSWS)	−0.12	0.05	−2.77	< .01	**

*$p < .05$, **$p < .01$, ***$p < .001$.

Data availability

Data will be made available on request.

References

Akinaga, K., Kindaichi, H., 2014. Shin Meikai Nihongo Akusento Jiten, 2nd Edition. Sanseido.

Beckman, M.E., 1986. Acoustic correlates of accent in English and Japanese. Stress and Non-Stress Accent. De Gruyter. https://doi.org/10.1515/9783110874020.145.

Beckman, M.E., Pierrehumbert, J.B., 1986. Intonational structure in Japanese and English. Phonol. 3, 255–309. https://doi.org/10.1017/S095267570000066X.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57 (1), 289–300.

Benson, R.R., Richardson, M., Whalen, D.H., Lai, S., 2006. Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech. NeuroImage 31 (1), 342–353. https://doi.org/10.1016/j.neuroimage.2005.11.029.

Boersma, P., Weenink, D., 2022. *Praat: doing phonetics by computer [Computer program]. Version 6.2.09* (6.2.09). http://www.praat.org/.

Chen, W.-R., Whalen, D.H., Tiede, M.K., 2021. A dual mechanism for intrinsic f0. J. Phon. 87, 101063. https://doi.org/10.1016/j.wocn.2021.101063.

Chiba, T., Kajiyama, M., 1958. The Vowel, Its Nature and Structure. The Phonetic Society of Japan.

Cutler, A., Otake, T., 1999. Pitch accent in spoken-word recognition in Japanese. J. Acoust. Soc. Am 105 (3), 1877–1888. https://doi.org/10.1121/1.426724.

Darwin, C., 2005. SWS [Praat Script]. https://users.sussex.ac.uk/~cjd/Praatscripts/SWS.

Erickson, D., 2002. Articulation of extreme formant patterns for emphasized vowels. Phonetica 59 (2–3), 134–149. https://doi.org/10.1159/000066067.

Erickson, D., Honda, K., Kawahara, S., 2017. Interaction of jaw displacement and F0 peak in syllables produced with contrastive emphasis. Acoust. Sci. Technol. 38 (3), 137–146. https://doi.org/10.1250/ast.38.137.

Fant, G., 1965. Formants and cavities. In: Bethge, W., Zwirner, E. (Eds.), Proceedings of the 5th International Congress of Phonetic Sciences. S. Karger AG, pp. 120–141. https://doi.org/10.1159/000426938.

Feng, Y.-M., Xu, L., Zhou, N., Yang, G., Yin, S.-K., 2012. Sine-wave speech recognition in a tonal language. J. Acoust. Soc. Am 131 (2), EL133–EL138. https://doi.org/10.1121/1.3670594.

Friesen, L.M., Shannon, R.V., Baskent, D., Wang, X., 2001. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. J. Acoust. Soc. Am 110 (2), 1150–1163. https://doi.org/10.1121/1.1381538.

Fu, Q.-J., Shannon, R.V., Wang, X., 1998. Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing. J. Acoust. Soc. Am 104 (6), 3586–3596. https://doi.org/10.1121/1.423941.

Greenwood, D.D., 1990. A cochlear frequency-position function for several species—29 years later. J. Acoust. Soc. Am 87 (6), 2592–2605. https://doi.org/10.1121/1.399052.

Hasegawa, Y., Hata, K., 1992. Fundamental frequency as an acoustic cue to accent perception. Lang Speech 35 (1–2), 87–98. https://doi.org/10.1177/002383099203500208.

Hualde, J.I., Elordieta, G., Gaminde, I., Smiljanić, R., 2002. From pitch-accent to stress-accent in Basque. In: Gussenhoven, C., Warner, N. (Eds.), Laboratory Phonology, Laboratory Phonology, 7. De Gruyter Mouton, pp. 547–584. https://doi.org/10.1515/9783110197105.2.547.

Huckvale, M., 2020. *ProRec: speech prompt & record system* (2.4). https://www.phon.ucl.ac.uk/resource/prorec/.

Igarashi, Y., 2015. Intonation. In: Kubozono, H. (Ed.), Handbook of Japanese Phonetics and Phonology. De Gruyter.

Ihlefeld, A., Deeks, J.M., Axon, P.R., Carlyon, R.P., 2010. Simulations of cochlear-implant speech perception in modulated and unmodulated noise. J. Acoust. Soc. Am 128 (2), 870–880. https://doi.org/10.1121/1.3458817.

Kawahara, S., 2015. The phonology of Japanese accent. In: Kubozono, H. (Ed.), Handbook of Japanese Phonetics and Phonology. De Gruyter, pp. 445–492. https://doi.org/10.1515/9781614511984.445.

Kawahara, S., Erickson, D., Suemitsu, A., 2017. The phonetics of jaw displacement in Japanese vowels. Acoust. Sci. Technol. 38 (2), 99–107. https://doi.org/10.1250/ast.38.99.

Kitahara, M., 2001. Category Structure and Function of Pitch Accent in Tokyo Japanese. Indiana University. Dissertation.

Kitahara, M., Amano, S., 2001. Perception of pitch accent categories in Tokyo Japanese. Gengo Kenkyu Lang. Res. 120, 1–34.

Ladd, D.R., 2008. *Intonational Phonology* (2nd Edition). Cambridge University Press. https://doi.org/10.1017/CBO9780511808814.

Levi, S.V., 2005. Acoustic correlates of lexical accent in Turkish. J. Int. Phon. Assoc. 35 (1), 73–97. https://doi.org/10.1017/S0025100305001921.

Liebenthal, E., Binder, J.R., Piorkowski, R.L., Remez, R.E., 2003. Short-term reorganization of auditory analysis induced by phonetic experience. J. Cogn. Neurosci. 15 (4), 549–558. https://doi.org/10.1162/089892903321662930.

Lim, M., Lin, E., Bones, P., 2006. Vowel effect on glottal parameters and the magnitude of jaw opening. J. Voice 20 (1), 46–54. https://doi.org/10.1016/j.jvoice.2004.09.003.

Moore, B.C.J., 2008. Basic auditory processes involved in the analysis of speech sounds. Philos. Trans. R. Soc. B: Biol. Sci. 363 (1493), 947–963. https://doi.org/10.1098/rstb.2007.2152.

Newman, R., Chatterjee, M., 2013. Toddlers' recognition of noise-vocoded speech. J. Acoust. Soc. Am 133 (1), 483–494. https://doi.org/10.1121/1.4770241.

Nittrouer, S., Lowenstein, J.H., Wucinich, T., Tarr, E., 2014. Benefits of preserving stationary and time-varying formant structure in alternative representations of speech: implications for cochlear implants. J. Acoust. Soc. Am 136 (4), 1845–1856. https://doi.org/10.1121/1.4895698.

Oxenham, A.J., 2013. Revisiting place and temporal theories of pitch. Acoust. Sci. Technol. 34 (6), 388–396. https://doi.org/10.1250/ast.34.388.

Plag, I., Kunter, G., Schramm, M., 2011. Acoustic correlates of primary and secondary stress in North American English. J. Phon. 39 (3), 362–374. https://doi.org/10.1016/j.wocn.2011.03.004.

Poser, W.J., 1984. The Phonetics and Phonology of Tone and Intonation in Japanese. Massachusetts Institute of Technology.

Remez, R.E., Dubowski, K.R., Broder, R.S., Davids, M.L., Grossman, Y.S., Moskalenko, M., Pardo, J.S., Hasbun, S.M., 2011. Auditory-phonetic projection and lexical structure in the recognition of sine-wave words. J. Exp. Psychol.: Hum. Percept. Perform. 37 (3), 968–977. https://doi.org/10.1037/a0020734.

Remez, R.E., Rubin, P.E., 1984. On the perception of intonation from sinusoidal sentences. Percept. Psychophys. 35 (5), 429–440. https://doi.org/10.3758/BF03203919.

Remez, R.E., Rubin, P.E., 1993. On the intonation of sinusoidal sentences: contour and pitch height. J. Acoust. Soc. Am. 94 (4), 1983–1988. https://doi.org/10.1121/1.407501.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., Lang, J.M., 1994. On the perceptual organization of speech. Psychol. Rev. 101 (1), 129–156. https://doi.org/10.1037/0033-295X.101.1.129.

Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D., 1981. Speech perception without traditional speech cues. Science 212 (4497), 947–950. https://doi.org/10.1126/science.7233191.

Remez, R.E., Thomas, E.F., Crank, A.T., Kostro, K.B., Cheimets, C.B., Pardo, J.S., 2018. Short-term perceptual tuning to talker characteristics. Lang Cogn. Neurosci. 33 (9), 1083–1091. https://doi.org/10.1080/23273798.2018.1442580.

Rosen, S., Hui, S.N.C., 2015. Sine-wave and noise-vocoded sine-wave speech in a tone language: acoustic details matter. J. Acoust. Soc. Am. 138 (6), 3698–3702. https://doi.org/10.1121/1.4937605.

Schnupp, J., Nelken, I., King, A., 2011. Auditory prostheses: from the lab to the clinic and back again. Audit. Neurosci.: Mak. Sense Sound 295–319.

Schouten, J.F., 1938. The perception of subjective tones. Proceed. Koninklijke Nederl. Akad. van Wetensch. 41, 1086–1093.

Schouten, J.F., 1940. The residue, a new component in subjective sound analysis. Proceed. Koninklijke Nederl. Akad. van Wetensch. 43, 356–365.

Schulman, R., 1989. Articulatory dynamics of loud and normal speech. J. Acoust. Soc. Am. 85 (1), 295–312. https://doi.org/10.1121/1.397737.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270 (5234), 303–304. https://doi.org/10.1126/science.270.5234.303.

Shinohara, Y., 2022. Perception of noise-vocoded sine-wave speech of Japanese pitch-accent words. JASA Express Lett. 2 (8), 085204. https://doi.org/10.1121/10.0013423.

Sugiyama, Y., 2017. Perception of Japanese pitch accent without F0. Phonetica 74 (2), 107–123. https://doi.org/10.1159/000453069.

Sugiyama, Y., 2022. Identification of minimal pairs of Japanese pitch accent in noise-vocoded speech. Front. Psychol. 13. https://doi.org/10.3389/fpsyg.2022.887761.

Vance, T.J., 1995. Final accent vs. no accent: utterance-final neutralization in Tokyo Japanese. J. Phon. 23 (4), 487–499. https://doi.org/10.1006/jpho.1995.0035.

Warner, N., 1997. Japanese final-accented and unaccented phrases. J. Phon. 25 (1), 43–60. https://doi.org/10.1006/jpho.1996.0033.

Weitzman, R.S., 1970. Studies in the Phonology of Asian Languages IX: Word accent in Japanese. Management Information Services.

Whalen, D.H., Chen, W.-R., Shadle, C.H., Fulop, S.A., 2022. Formants are easy to measure; resonances, not so much: lessons from Klatt (1986). J. Acoust. Soc. Am. 152 (2), 933–941. https://doi.org/10.1121/10.0013410.

Whalen, D.H., Levitt, A.G., 1995. The universality of intrinsic F0 of vowels. J. Phon. 23 (3), 349–366. https://doi.org/10.1016/S0095-4470(95)80165-0.

Winn, M.B., 2021. *Vocoder [Praat Script]* (No. 45). http://www.mattwinn.com/praat/vocode_all_selected_v45.txt.

Yost, W.A., 2009. Pitch perception. *Attention*. Percept. Psychophys. 71 (8), 1701–1715. https://doi.org/10.3758/APP.71.8.1701.