The effect of $f_0$ fall, downstep, and secondary cues in perceiving Japanese lexical accent
Yukiko Sugiyama, C. T. Justine Hui and Takayuki Arai

Citation: The Journal of the Acoustical Society of America 150, 2865 (2021); doi: 10.1121/10.0006689
View online: https://doi.org/10.1121/10.0006689
View Table of Contents: https://asa.scitation.org/toc/jas/150/4
Published by the Acoustical Society of America

ARTICLES YOU MAY BE INTERESTED IN

Native listeners rely on rhythmic cues when deciding on the nativeness of speech
The Journal of the Acoustical Society of America 150, 2836 (2021); https://doi.org/10.1121/10.0006537

Intelligibility of dysphonic speech in auralized classrooms
The Journal of the Acoustical Society of America 150, 2912 (2021); https://doi.org/10.1121/10.0006741

Accuracy and cue use in word segmentation for cochlear-implant listeners and normal-hearing listeners presented vocoded speech
The Journal of the Acoustical Society of America 150, 2936 (2021); https://doi.org/10.1121/10.0006448

Overlapped speech detection using phase features
The Journal of the Acoustical Society of America 150, 2770 (2021); https://doi.org/10.1121/10.0006614

Constrained optimal filter design for multi-channel active noise control via convex optimization
The Journal of the Acoustical Society of America 150, 2888 (2021); https://doi.org/10.1121/10.0006738

Estimation of locally reacting surface impedance at modal frequencies using an eigenvalue approximation technique
The Journal of the Acoustical Society of America 150, 2921 (2021); https://doi.org/10.1121/10.0006742
The effect of $f_0$ fall, downstep, and secondary cues in perceiving Japanese lexical accent

Yukiko Sugiyama,1,a) C. T. Justine Hui,2 and Takayuki Arai3
1Department of Foreign Languages and Liberal Arts, Faculty of Science and Technology, Keio University, 4-1-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8521, Japan
2Acoustics Research Centre, Department of Mechanical Engineering, University of Auckland, Auckland 1010, New Zealand
3Department of Information and Communication Sciences, Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo 102-8554, Japan

ABSTRACT:
Lexical pitch accent in Japanese is primarily realized as a steep fall in $f_0$ from an accented syllable into the following one. In addition, when a phrase that contains an accented syllable is followed by another phrase, the following phrase undergoes downstep, a compression of the $f_0$ range. Furthermore, while their acoustic identity is not yet clear, secondary cues to Japanese pitch accent are known to exist. The present study examined how speakers of Tokyo Japanese used acoustic information from these three sources in perceiving lexical pitch accent in Tokyo Japanese. Listeners heard stimuli in which the acoustic cues related to accent were independently manipulated and were asked to identify if a word presented sentence-medially was a final-accented word or its unaccented counterpart. Results found that listeners’ judgments of words were most consistent with the presence or absence of downstep. That is, listeners identified that the preceding phrase contained an accented word when the following phrase was downstepped. Listeners also used the $f_0$ fall to determine if the word in question was a final-accented word or an unaccented word. Secondary cues to pitch accent were most weakly related to listeners’ identification of accent.

(Received 26 May 2021; revised 16 September 2021; accepted 22 September 2021; published online 18 October 2021)
[Editor: Benjamin V Tucker]

I. INTRODUCTION
Lexical prosody of Tokyo Japanese (simply Japanese hereafter) can be described as having a pitch accent (PA) system in which phonological prominence is realized primarily by fundamental frequency ($f_0$). Lexical PA in Japanese manifests itself most clearly as a sudden fall in $f_0$ from an accented syllable into the following one. For example, the word hasi “bridge” has an accent on its final syllable (indicated by an acute symbol) and a pitch pattern of low-high (LH).1 When it is followed by a particle, a grammatical word that usually combines with a preceding word or words into a single accent phrase, the particle has a low pitch. By contrast, when hasi “edge,” which has no accent, is followed by a particle, $f_0$ stays high. This difference in the presence or absence of accent is known to affect the $f_0$ contour of the phrase that follows. When a phrase follows a phrase containing an accented word, it undergoes downstep, or the compression of the $f_0$ range that accompanies a lowering of an $f_0$ peak (KuboZono, 1988; Pierrehumbert and Beckman, 1988; Poser, 1984). Downstep will be explained in more detail in Sec. 1B. When a phrase follows an unaccented word, downstep does not occur, and therefore the lexical accent in the following phrase is fully realized. This means that at least two sources of information are available when a listener tries to identify the accentual property of a word: (1) the $f_0$ contour at the accented syllable into the following syllable and (2) the presence or absence of downstep in the following phrase. Furthermore, although the acoustic identity is not yet fully specified, perception studies that utilized speech containing no $f_0$ or its harmonics suggest that secondary cues to Japanese PA exist (Higashikawa et al., 1996; Sugito et al., 1991; Sugiyama, 2017). In this paper, we investigated how native Japanese listeners use three kinds of acoustic information [(1) the $f_0$ fall that is characteristic of an accented syllable, (2) the presence or absence of downstep in the following phrase, and (3) secondary cues to Japanese PA] in word identification.

This introductory section is organized as follows: Sec. IA gives a brief overview of the nature of Japanese PA, which is followed by an account of downstep in Japanese in Sec. 1B. Then Sec. 1C lays out the research questions addressed in the present study.

A. PA in Japanese
PA in Japanese is used distinctively to distinguish the meaning of words and is phonetically realized as a steep fall in $f_0$ after an accented syllable. For example, when the phoneme sequence hasi has an accent on the initial syllable, it has a pitch pattern of HL (high-low) and means “chopsticks.” When it has an accent on the second syllable, it has a pitch pattern of LH and means “bridge.” Furthermore, words can also have no accent. When hasi has no accent, it has a pitch

---
a)Electronic mail: yukiko@keio.jp, ORCID: 0000-0002-3350-7831.

---
pattern of LH and means “edge.” The most prominent acoustic difference between words that have an accent on their final syllables and words that have no accent (unaccented words) appears when they are followed by a particle (Beckman and Pierrehumbert, 1986; Kubozono, 1988). An accentual phrase in Japanese consists of one or more words, but it can contain at most one accent (Beckman and Pierrehumbert, 1986). In Japanese, nouns often form an accentual phrase with their following particle. When the accent triplet hasi, hasi, and hasi are followed by the nominative particle ga, the location of accent and pitch patterns can be summarized as shown in Table I.

Table I also illustrates some characteristics of Japanese PA. In Japanese, every lexical word can have at most one prominence or accent, known as culminativity (Hyman, 2006). This means that an $f_0$ fall occurs at most once within a lexical word. While accent cannot occur more than once per lexical word, unaccented words are possible in Japanese, meaning that accent is not obligatory. Since the bearer of accent is the syllable in Japanese (McCawley, 1978), for words with $x$ number of syllables, there are $x + 1$ possible accent patterns, including unaccented words. The table also shows that low-pitched syllables are always unaccented, while high-pitched syllables can be accented or unaccented. Phonetically, accented syllables have a greater $f_0$ rise from the preceding syllable to the accented syllable and a higher $f_0$ peak than high-pitched unaccented syllables, known as “accentual boost” (Kubozono, 1988; Pierrehumbert and Beckman, 1988; Poser, 1984). In other words, although accented syllables and high-pitched unaccented syllables are both described as having high pitch phonologically, their precise acoustic properties are not the same at the phonetic level.

When a syllable is accented, the $f_0$ peak typically occurs on the accented syllable. However, Neustupný (1966) found that this is not always the case. That is, the $f_0$ peak sometimes occurs in the syllable following the one that is lexically accented, yet native Japanese listeners perceive the preceding syllable to be accented, a phenomenon known as ososagari (a delayed pitch fall; Sugito, 1972). For a syllable preceding the syllable containing the $f_0$ peak to be perceived as accented, the $f_0$ fall following the peak has to be a steep one. Kitahara (2001) and Hasegawa and Hata (1992) experimentally investigated this issue by shifting the location of the $f_0$ peak and manipulating the degree of the $f_0$ fall to find that there is a trade-off between the location of the $f_0$ peak and the degree of fall in $f_0$ after the peak; the later the peak occurs in the syllable, the greater the $f_0$ fall has to be for the preceding syllable to be perceived as accented. Kitahara’s study included disyllabic minimal pairs of final-accented and unaccented words as test materials. For each pair of words, four different levels were created for the $f_0$ maximum value and the minimum $f_0$ value. In addition, two different $f_0$ fall rates were tested. Kitahara found that the $f_0$ peak values and the amount of fall affected the perception of accent, consistent with Hasegawa and Hata (1992), but unlike Hasegawa and Hata, who found that a steep $f_0$ fall promoted the perception of accent, the fall rates had little effect. Kitamura et al. (2019) made use of a speech database to investigate when pitch delay was likely to occur and found that it was more frequent with female speakers than male speakers, when words had an accent on their initial syllables, in longer words, and when the accented syllable was followed by a syllable containing a back vowel. Hui et al. (2019) examined the effect of $f_0$ contours on the perception of lexical accent by a different group of listeners. While listeners in earlier studies were young adults with normal hearing, Hui et al. focused on elderly listeners aged 60 and older and explored how age related hearing deficits might affect the perception of lexical PA. In their study, a native speaker of Tokyo Japanese produced two minimal pairs of final-accented and unaccented words in the carrier sentence watasi wa __ ga suki “I like/prefer__” in the location indicated by “__.” The $f_0$ in the utterances was manipulated so that the $f_0$ contours of the test words gradually changed from those of final-accented words to unaccented words and vice versa, resulting in four continua of 11 stimuli each. As expected, the listeners’ perception changed from unaccented words to final-accented words as the $f_0$ fall became steeper, although the shift in perception was clearer for normal hearing elderly listeners than for elderly listeners with pitch processing deficits.

In addition to $f_0$ contours, studies suggest that other cues are also available to listeners when they distinguish types of word accents. One such piece of evidence comes from perception studies that used speech that contained no $f_0$ information (Sugito et al., 1991; Sugiyama, 2017). Sugito et al. (1991) examined if listeners could identify words in whispered speech. Since whisper is typically produced without vocal fold vibrations, it contains no $f_0$ or its harmonics. Sugito found that listeners’ accuracy was almost 90%, suggesting that cues other than $f_0$ exist for Japanese PA. However, results from Mandarin Chinese suggest that speakers may exaggerate secondary cues when they know that the primary cue will not be available (Liu and Samuel, 2004). Mandarin Chinese is a tone language where difference in tone is primarily realized by $f_0$ or its harmonics: monosyllabic words produced in whisper or its harmonics. Monosyllabic words produced in whisper and those produced naturally from which the $f_0$ and its harmonics were artificially removed. They found that syllable durations were significantly longer in whispered speech. Based on this finding, they argued that speakers may...
exaggerate secondary cues in whispered speech. Their findings have an important implication that if one intends to investigate secondary cues that help listeners perceive lexical distinction in natural speech, whispered speech is not appropriate as test material. To examine if acoustic information other than \( f_o \) exists for Japanese PA in natural speech, Sugiyama (2017) removed \( f_o \) and its harmonics from speech produced naturally and replaced them by white noise to create stimuli. When minimal pairs between final-accented words and unaccented words embedded in a carrier sentence were presented to Tokyo Japanese speakers, listeners were able to identify words above chance level (mean \( d' = 3.8 \)), suggesting the existence of non-\( f_o \) cues to Japanese PA. When it comes to what acoustic properties other than \( f_o \) carry accent information, however, results were inconclusive, except that duration is generally agreed not to correlate with accent (Beckman, 1986; Cutler and Otake, 1999; Kaiki et al., 1992). One exception is Itahashi (2005), who notes that vowel duration is longer in accented syllables than in unaccented syllables but provides no data or detailed descriptions. Other research that examined duration in a controlled experimental setting (Beckman, 1986; Cutler and Otake, 1999) and corpus-based research (Kaiki et al., 1992) found that duration was not correlated with PA. Given that duration is phonemic in Japanese, where difference in duration results in different words (e.g., \( k_jo ri \) “distance” vs \( kjo ri \) “hometown”), it is not surprising that duration is not affected by accent. For formant frequencies and intensity, results from previous studies show that they tend to be affected by accent, but it is not clear how robust or consistent its effect on these acoustic properties is. Higashikawa et al. (1996) asked Japanese speakers to produce whispered vowels at different intended pitches and found that frequencies of the first formant were higher when higher pitches were intended for some speakers, while frequencies of the third formant were higher for other speakers. Analyzing natural speech, Sugiyama (2017) found that the first formant frequencies tended to be higher when the \( f_o \) was higher. The study also found that intensity tended to be higher for accented words than unaccented words. The results are consistent with earlier findings, which found a tendency toward greater intensity for accented syllables than for unaccented syllables (Beckman, 1986; Weitzman, 1970). However, intensity was only weakly correlated with accent. As far as the authors are aware, no conclusive data exist as to what exactly drives the percept of PA in the absence of \( f_o \) or its harmonics. Nevertheless, findings from these studies support the idea that Japanese PA is realized at least to some extent by cues other than \( f_o \) and its harmonics.

B. Downstep in Japanese

Another important difference between accented words and unaccented words is that while the former trigger downstep, the latter do not (Kubozeno, 1988; Pierrehumbert and Beckman, 1988; Poser, 1984). Downstep is a compression of the \( f_o \) range when an accented phrase follows another accential phrase that contains an accented word (an accented accential phrase). When downstep occurs, the \( f_o \) peak is also lowered. Because it occurs only when a preceding accential phrase contains a word that is lexically accented, it is a phonological process and is differentiated from a phonetic process of declination, an \( f_o \) downtrend that occurs over the course of an utterance irrespective of the presence or absence of accented words. For example, compare a pair of utterances: \( hasi o ari\(\_\)ku \) (bridge ACCUSATIVE walk) and \( hasi o ari\(\_\)ku \) (edge ACCUSATIVE walk). The only difference between the two utterances is that while the first word in the former is an accented word (\( hasi \)), the first word in the latter (\( hasi \)) is an unaccented word. Figure 1 shows \( hasi o ari\_ku \) and \( hasi o ari\_ku \) produced in Tokyo Japanese by a female speaker. In the figure, the blue lines indicate the \( f_o \) movements. In the first utterance, because \( hasi o \) is an accented accentual phrase, \( ari\_ku \) undergoes downstep. By contrast, in the second utterance, \( hasi o \) is an unaccented accentual phrase and therefore downstep does not occur in the following phrase \( ari\_ku \). While the lexical accent in \( ari\_ku \) is fully realized as a steep \( f_o \) fall after the accented syllable when it follows an unaccented accentual phrase (the second utterance), the range of \( f_o \) fall is compressed with its peak lowered when \( ari\_ku \) follows an accented accentual phrase due to downstep (the first utterance). The comparison of the two utterances demonstrates the effects of downstep: the compression of the \( f_o \) range and the lowering of the \( f_o \) peak. The two utterances also illustrate an accentual boost explained in Sec. IA. Although both \( hasi o \) and \( hasi \) have the \( f_o \) pattern of low-high, the amount of \( f_o \) rise is much greater and the \( f_o \) peak is much higher for the high-pitched accented syllable than for the high-pitched unaccented syllable.

A steep \( f_o \) fall at the accented syllable into the following syllable is probably the most studied and well known aspect of lexical PA in Japanese. However, Sugiyama (2012) suggests that downstep may play a more important role in word identification than has been assumed, although no conclusive data were included. Minimal pairs of final-accented words and unaccented words were produced in the carrier sentence \( \_k\_\_re wa__ \) to \( it\_\_u \) (he TOPIC __ CITATION say-PAST) in the sentence-medial position indicated by "__". When accentual phrases consisting of one of the test words and the particle to were excised from the utterances and presented to listeners, listeners’ mean accuracy was roughly 70%. Considering how much the literature on Japanese PA has focused on the \( f_o \) fall at the accented syllable as a manifestation of PA, accuracy was rather low. In a later study, Sugiyama (2017) tested word identification of similar types of words in the sentence \( \_k\_\_re wa__ \_\_\_g\_\_\_i \) (the TOPIC __ NOMINATIVE prefer/like), and the entire sentence was presented to listeners.2 Tokyo Japanese speakers were roughly 96% correct in distinguishing minimal pairs of final-accented and unaccented words. These studies suggest that downstep or its absence also provides relevant information when listeners judge lexical identity of words in the preceding phrase. Put differently, both the \( f_o \) contour of the phrase
that contains the word in question and the $f_o$ contour of the following phrase are needed for reliable word identification.

C. Research questions of the current study

Sections IA and IB show that at least three sources of information are available to Japanese listeners when they judge the lexical identity of words even when they are segmentally identical. One is the $f_o$ contour of the accentual phrase in which the word in question occurs. As seen in Sec. IA, lexical PA in Japanese exhibits a steep $f_o$ fall at the accented syllable into the following one. This $f_o$ fall indicates the location of accent, helping listeners identify the word they heard when there is more than one candidate word for a specific phoneme sequence. When there is no $f_o$ fall, it indicates that the word is unaccented. In addition to the $f_o$ contour, perception studies that used speech containing no $f_o$ or its harmonics suggest that it may not be the only acoustic correlate of Japanese PA, although the acoustic identity of such correlates is not yet certain. If non-$f_o$ information as well as $f_o$ is useful for listeners, their judgment of final-accented words and unaccented words may not totally change by manipulating only the $f_o$ contour. The third source of information is the presence or absence of downstep in the following phrase. As seen in Sec. IB, an $f_o$ contour is affected by whether the preceding phrase contains an accented word or not. When it does, words in the following phrase undergo downstep. By contrast, when it does not, they preserve their original $f_o$ contour. Recent studies suggest that downstep may be a cue in disambiguating the lexical identity of the words in the preceding phrase. The steep $f_o$ fall at an accented syllable and downstep triggered by lexical accent have been well studied in the literature of Japanese PA. However, how they interact to affect the perception of lexical accent has been less understood. The present study aims to investigate how listeners integrate these three kinds of acoustic information related to Japanese PA in word identification by independently manipulating these acoustic properties.

More specifically, disyllabic minimal pairs between final-accented words and unaccented words that differ only in the presence or absence of accent were used as test words. To examine the effect of downstep on the perception of accent in the preceding accentual phrase, test words were presented in a sentence-medial position followed by another accentual phrase. To examine how the steepness of an $f_o$ fall affects the perception of accent, a continuum that differed in the degree of $f_o$ fall was created for each pair of words. Furthermore, by creating test words based on both accented words and unaccented words, the present study also investigated if acoustic properties other than $f_o$ contour affected the perception of words.

II. METHODS

A. Listeners

The participants were 24 students recruited at Keio University (19–24 years old). All were native speakers of Tokyo Japanese who grew up in Tokyo or its neighboring areas (Kanagawa, Saitama, or Chiba). To qualify as a listener in this experiment, at least one of the participants’ parents also had to have grown up mainly in Tokyo or its neighboring areas. None of the listeners reported any history of either speech or hearing disorder. The experiment lasted for roughly 80 min per listener, and the participants were paid for their participation. The experiments were conducted based on approved ethics protocols for the treatment of human subjects.

B. Stimuli

Table II shows a list of test words used in this experiment. The list consists of five minimal pairs of final-accented and unaccented words. Within each pair, the words have the same phoneme sequence and the pitch pattern of low-high. They differed only in the presence or absence of PA in the final syllable. All words had relatively high familiarity with familiarity ratings of 5.0 or higher on a seven-
Because PA is lexical and not predictable in Japanese, it was important that test words were familiar to participants; otherwise, they would not know the accent pattern of the test words. While it was desirable to include more than five pairs of words as test words, considering the total number of stimuli to be presented to each listener, as will be explained later in this section, five pairs was the maximum that could be tested within a reasonable time frame. All test words were produced in the carrier sentence *watasi wa__ ga suk* (I TOPIC__NOMINATIVE like), which means “I like/prefer __,” by a female native Tokyo Japanese speaker. Each test word was produced sentence-medially as indicated by the underscore “__” in the carrier sentence.

Once all the words were produced in the carrier sentence, the utterances were divided into three parts, (1) the initial phrase *watasi wa*, (2) an accentual phrase consisting of a test word and the particle *ga*, and (3) the final phrase *suk*, out of which the second part was used as baseline stimuli to create continua between accented accentual phrases and unaccented accentual phrases.

- **Table II. Minimal pairs used in the experiment.**

<table>
<thead>
<tr>
<th>Phoneme sequence</th>
<th>Final accented (LH)</th>
<th>Unaccented (LH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 hana</td>
<td>“flower”</td>
<td>“nose”</td>
</tr>
<tr>
<td>2 hasi</td>
<td>“bridge”</td>
<td>“edge”</td>
</tr>
<tr>
<td>3 hati</td>
<td>“eight”</td>
<td>“bee”</td>
</tr>
<tr>
<td>4 osu</td>
<td>“male”</td>
<td>“vinaigre”</td>
</tr>
<tr>
<td>5 sita</td>
<td>“tongue”</td>
<td>“below”</td>
</tr>
</tbody>
</table>

The letters in the parentheses indicate that the first syllable has a low pitch and the second syllable a high pitch. The acute sign above the H indicates that the syllable has an accent.

Figure 2 shows the waveform, spectrogram, and $f_o$ contour of the phrase *hana ga suk* taken out of the utterance *watasi wa haná ga suk*. The $f_o$ indicated by the blue line laid over the spectrogram increases toward the accented syllable of the test word and falls steeply after the syllable. Although the utterance-final word is also accented, the $f_o$ stays low because of downstep. Based on this utterance, half the stimuli for the *hana* pair were created. Slope #1 in Fig. 3 corresponds to the $f_o$ contour in Fig. 2. The remaining six slopes down to slope #7 (in the figure, slopes #2 to #6 are not labeled for readability but correspond to the lines between slope #1 and slope #7) were created from slope #1 by manipulating the $f_o$ using Praat (Boersma and Weenink, 2019). Slope #7 was created so as to mimic the $f_o$ contour of the accentual phrase *hana ga* indicated as the blue line in
the spectrogram in Fig. 4. Because the test word is unaccented, the $f_0$ does not fall at the particle. When slope #7 in Fig. 3 and the $f_0$ track in Fig. 4 are compared, one can see that their shapes are very similar.

The slopes of varying degrees of steepness illustrated in Fig. 3 were created as follows. Using the Manipulation function in Praat, the same number of points was selected for each pair of an accented accentual phrase (e.g., hanâ ga) and an unaccented accentual phrase (e.g., hana ga). The number of points was typically three, and they were located manually roughly near the beginning, in the middle, and near the end of the accentual phrase. When a resulting accentual phrase did not sound natural, one or more points were added until it sounded natural. Then $f_0$ differences between an accented accentual phrase and its unaccented counterpart were calculated for each of the corresponding points, and the differences were converted into cents. Once the differences were obtained in cents, they were divided into ten, resulting in 11 slopes that were equi-distanced from one another. Of the 11 slopes created for each accentual phrase, from a practical concern to keep the duration of the experiment to a reasonable time frame, seven slopes were used in the experiment. It was expected that listeners’ judgments would shift from a final-accented word to an unaccented word around the middle of the continuum from slope #1 to slope #7. Therefore, the slopes were closer to one another in the middle (slopes #3 to #5) to observe the change in perception. The slopes are more sparse at the peripheries because it was expected that listeners’ perception would be more or less stable. For slope numbers between 1 and 2, 2 and 3, 5 and 6, and 6 and 7, there was an additional slope in between when they were created, but they were not used as stimuli in the experiment. Figure 5 shows the $f_0$ contours of the slopes created from the utterance in which an unaccented word hana was produced. The slopes were created using the same method by which the slopes for the accented accentual phrase were created. The only difference between the slopes illustrated in Figs. 3 and 5 was that, for the former, the original $f_0$ contour was slope #1, and the other slopes were created by manipulating the contour using Praat, whereas for the latter, the original $f_0$ contour was slope #7, and the other slopes were created by $f_0$ manipulation. The crucial piece to note about the slopes shown in Figs. 3 and 5 is that, as far as the $f_0$ contours of the accentual phrases are concerned, their shapes are very similar. This is because slope #7 in Fig. 3 was created by referring to the $f_0$ contour in Fig. 4, and slope #1 in Fig. 5 was created by referring to the $f_0$ contour in Fig. 2. Since the $f_0$ shows the $f_0$ contours of the slopes created from the utterance in which an unaccented word hana was produced. The slopes were created using the same method by which the slopes for the accented accentual phrase were created. The only difference between the slopes illustrated in Figs. 3 and 5 was that, for the former, the original $f_0$ contour was slope #1, and the other slopes were created by manipulating the contour using Praat, whereas for the latter, the original $f_0$ contour was slope #7, and the other slopes were created by $f_0$ manipulation. The crucial piece to note about the slopes shown in Figs. 3 and 5 is that, as far as the $f_0$ contours of the accentual phrases are concerned, their shapes are very similar. This is because slope #7 in Fig. 3 was created by referring to the $f_0$ contour in Fig. 4, and slope #1 in Fig. 5 was created by referring to the $f_0$ contour in Fig. 2. Since the $f_0$
contours with the same slope numbers should be virtually identical regardless of whether they are created from an accented accentual phrase or unaccented accentual phrase, if listeners’ judgments for the same slope numbers differed within a pair, it would suggest that some acoustic cues other than the $f_o$ were present in the stimuli. The most obvious difference between the utterance in which a final-accented word was the test word and the one in which an unaccented word was the test word appears in the $f_o$ contour in the utterance-final word suki, which is discussed next.

The final part of the carrier sentence suki undergoes downstep when it follows an accented accentual phrase, while it does not when it follows an unaccented accentual phrase. When the $f_o$ contours of suki in Fig. 2 (or Fig. 3) and Fig. 4 (or Fig. 5) are compared, the $f_o$ is overall higher in Fig. 4 (or Fig. 5) because suki does not undergo downstep. To investigate how downstep affects the perception of lexical accent in the preceding phrase, for each accent pair, both a series of seven slopes created from an accented accentual phrase and those created from its corresponding unaccented accentual phrase were spliced onto watasi wa and suki taken from the utterance that contained a final-accented word and watasi wa and suki taken from the utterance that contained its unaccented counterpart. This means that, for each pair of words, 28 stimuli were created: 7 slopes created from an accented accentual phrase and its corresponding unaccented accentual phrase (a total of 14 slopes) × 2 kinds of watasi wa and suki, one in which suki was downstepped and one in which suki was not downstepped. For the sake of clarity, in the following, an accentual phrase created from an utterance in which a final-accented word was produced will be noted as WORD A, and one created from an utterance in which an unaccented word was produced will be noted as WORD U. It should be kept in mind that the particle ga is also part of the notations WORD A and WORD U. Similarly, the initial and final parts of the carrier phrase watasi wa and suki produced with a final-accented word will be noted as CARRIER A, and those produced with an unaccented word will be noted as CARRIER U. Apart from token differences, the initial part of the carrier phrase should basically be the same for CARRIER A and CARRIER U, but suki was subject to downstep for CARRIER A while it was not for CARRIER U.

The notation WORD A + CARRIER A indicates an accented accentual phrase followed by an accentual phrase that is downstepped. This is an $f_o$ contour typically observed when a test word is final-accented in natural speech. WORD U + CARRIER U indicates that an unaccented accentual phrase is followed by an accentual phrase that is not downstepped. This is an $f_o$ contour typically observed when a test word is unaccented. Theoretically, the combinations WORD A + CARRIER U and WORD U + CARRIER A might be possible when context is manipulated to induce focus on a specific word. The $f_o$ contour represented by the former combination indicates that an accented accentual phrase is followed by an accentual phrase that is not downstepped. According to Pierrehumbert and Beckman (1988), focus blocks downstep, and the $f_o$ contour of the focused word will be fully realized even when it follows an accented word. Given that suki was part of the carrier sentence that was constant for all the stimuli, in the current experiment, the interpretation that this word is focused is unlikely. The combination WORD U + CARRIER A might be possible when special focus is on the first word. This combination represents an $f_o$ contour in which an unaccented accentual phrase is followed by an accentual phrase that is downstepped. This type of strong focus is typically examined by intentionally contrasting words to be examined. Since no contextual manipulation was implemented in the current study, it is not likely that listeners would interpret this combination to be a plausible one. The four combinations, WORD A + CARRIER A, WORD A + CARRIER U, WORD U + CARRIER A, and WORD U + CARRIER U, allow the assessment of which had a stronger influence on listeners’ judgments: accent information from an accentual phrase that includes a test word or the information from a downstep. In total, 140 stimuli were created. As summarized in Table III, for each pair of words, seven slopes that have varying degrees of steepness were created, which were then combined with the carriers with and without downstep: 5 pairs × (WORD A + WORD U) × 7 slopes × (CARRIER A + CARRIER U).

C. Procedure

Each listener was tested individually in a sound-treated room. All stimuli were presented with a MacBook Pro laptop computer, through an RME (Haimhausen, Germany) Babyface Pro audio interface. The participants listened to the stimuli on Audio-Technica (Tokyo, Japan) ATH-M50x headphones connected to the audio interface. The stimuli were presented at 70 dB sound pressure level (SPL), as measured using a RION (Tokyo, Japan) sound level meter NL-42 connected to an artificial ear [Brüel & Kjær (Nærum, Denmark) type 4153]. Since response times were not analyzed in this study, listeners gave their responses using a wireless mouse.

Before a practice session and actual trials, listeners were briefed on the utterances they would hear. As stated in Sec. II B, all test words were presented embedded in the carrier sentence watasi wa__ ga suki (I like/prefer__). This carrier sentence was selected because it was more or less semantically compatible with many words. Even so, the resulting sentences sounded more natural with some words than others. In perception studies in which test words appear embedded in a sentence, participants often report that their responses were likely to be influenced by whether the

---

**TABLE III. Factors tested in the experiment.**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word pairs</td>
<td>hana, hasi, hati, osu, sita</td>
</tr>
<tr>
<td>Word accent</td>
<td>Accented (WORD A), unaccented (WORD U)</td>
</tr>
<tr>
<td>$f_o$ contour slopes</td>
<td>Slopes #1 to #7</td>
</tr>
<tr>
<td>Carrier</td>
<td>With downstep (CARRIER A)</td>
</tr>
<tr>
<td></td>
<td>Without downstep (CARRIER U)</td>
</tr>
</tbody>
</table>
sentences they heard were plausible or not. In an effort to minimize the effect of plausibility on the listeners’ responses, the experimenter gave a scenario in which they might hear the carrier sentence with each of the test words. In the practice session, a pair of words, namı “wave” and namı “standard quality,” which did not appear in actual trials, produced in the same carrier sentence by the same speaker was presented to familiarize participants with the task they would be asked to perform in the actual trials. These two stimuli were presented five times in a random order. At each trial, two alternatives, the final-accented word and its unaccented counterpart, appeared on the computer screen. The final-accented word was written with kanji (Chinese characters used in Japanese), and the unaccented word was written in a combination of kanji and hiragana, a Japanese syllabary, based on how they are commonly written in Japanese. After each stimulus was presented, listeners were asked to click on the word they thought they heard. The actual trials consisted of ten blocks in which one block consisted of 140 stimuli (2 words × 7 slopes × 2 carrier phrases × 5 pairs). Ten blocks (ten repetitions per item) were presented to each listener considering that eight repetitions were presented to each listener in Sugiyama (2017). Although the nature of the stimuli was quite different, the differences among the stimuli seemed even more subtle in the current study. We deemed that collecting ten responses per item would allow us to obtain robust data. Within each block, the order in which the five pairs of words appeared was randomized, but once a certain pair of words started, all of the 28 stimuli for the pair were presented in a random order. When all the stimuli were presented for the pair, the trials proceeded to the next pair of words. Listeners were allowed to take a break before proceeding to the next pair if they wished. The words used in the actual trials were written solely in Chinese characters, except osu “vinegar,” which was written in a combination of hiragana and a Chinese character because it looked more natural.

After data were collected, it turned out that four participants did not qualify as native speakers of Tokyo Japanese because they had lived in areas where varieties of Japanese other than Tokyo Japanese were spoken. Therefore, 28,000 responses (140 stimuli × 10 repetitions × 20 listeners) were analyzed.

D. Analysis

Data were analyzed to assess how the listeners’ judgments of final-accented and unaccented words were affected by (1) source PA (WORD A and WORD U: whether an accentual phrase was created from a phrase containing a final-accented word or unaccented word), (2) carrier (CARRIER A and CARRIER U: whether the carrier phrase watași wa and suki was taken from the utterance in which a final-accented word was produced or its unaccented counterpart), (3) slope number (the \( f_o \) slope continuum from an accented accentual phrase to an unaccented accentual phrase; see Figs. 3 and 5), and (4) word (five pairs of words). Slope examined how the \( f_o \) contour within the accentual phrase that included a test word affected the perception of the test word, while carrier examined how the presence or absence of downstep affected the perception of accent of the test word. Source PA assessed how cues other than slope and carrier affected the perception of the test word. While it may seem that word should be treated as a random factor, it was treated as a fixed factor (independent variable) because the number of items was limited. This approach also seemed appropriate in light of the fact that each pair showed different results from one another, as will be discussed in Sec. III.

All the data were analyzed with generalized linear mixed-effects regression with a binomial family link function in R version 3.6.1 (R Core Team, 2019) with the lme4 package version 1.1–21 (Bates et al., 2015). Plots were created using ggplot2 version 3.2.1 (Wickham, 2016).

III. RESULTS

Data collected were analyzed using a generalized linear mixed-effects regression with the following four fixed effects: (1) source PA, (2) carrier, (3) slope number, and (4) word. Listeners were coded as a random effect. The maximal model proposed by Barr et al. (2013) would include four-way, three-way, and two-way interactions as well as random slopes for all the fixed effects, as all the fixed effects were within-listener factors in the current experiment. However, the maximal model would be too complex to interpret with too many possible interactions (cf. Bates et al., 2015). Indeed, the model failed to converge presumably due to overparameterization. The mixed-effects model that reached convergence included one two-way interaction between source PA and word with no random slopes. Thus, the maximal model justified by the design reads as follows in the syntax of glmer (Bates et al., 2015): ratios of responses as accented words \( \sim \) source PA \* word \+ carrier \+ slope + (1 | listener). A type II analysis of variance (ANOVA) conducted on the model using the car package (Fox and Weisberg, 2019) version 3.0.6. revealed significant main effects of source PA \( (f^2 = 6.03, \ p = 0.01) \), carrier \( (f^2 = 1164.28, \ p < 0.01) \), slope \( (f^2 = 5674.28, \ p < 0.01) \), and word \( (f^2 = 405.36, \ p < 0.01) \) and a significant interaction between source PA and word \( (f^2 = 307.37, \ p < 0.01) \). Figures 6–10 illustrate the probabilities of responses as an accented word for each pair. The shaded bands along the lines indicate the 95% confidence intervals. The significant interaction of source PA and word found in the analysis above can be visually confirmed in the figures in which different pairs showed different response patterns from listeners.

To further investigate how listeners’ responses were affected by the \( f_o \) contour in the accentual phrase, downstep, and other acoustic cues, each pair was analyzed separately using type II ANOVA. In analyzing each pair of words, the fixed effects were slope, carrier, and PA source, and listeners were coded as a random factor: the ratios of responses as the final-accented word \( \sim \) source PA \* carrier \* slope + (1|...
listener). For all of the pairs, this was the maximal model justified by the design.

The results of the type II ANOVA for *hana* found significant main effects of carrier ($\chi^2 = 47.44, p < 0.01$) and slope ($\chi^2 = 1562.82, p < 0.01$), while the main effect of source PA was marginal ($\chi^2 = 3.59, p = 0.06$). The interaction between carrier and slope was also marginal ($\chi^2 = 11.5, p = 0.07$). The other two-way interactions and three-way interaction failed to reach significance, all $p > 0.1$. Figure 6 shows that, overall, probabilities of response as the final-accented word are similar regardless of the type of carrier (CARRIER A and CARRIER U) and the type of source PA (WORD A and WORD U) with all of the four lines close to one another. The effect of carrier is more clearly observed in the slopes that are in transition from the final-accented word to the unaccented word, in which the stimuli with CARRIER A elicited more accented responses than the stimuli with CARRIER U, suggesting that listeners used the acoustic information from carrier when information from slope was ambiguous between accented and unaccented. As expected, slope had a clear effect on the listeners’ responses. In all four conditions of WORD A + CARRIER A, WORD A + CARRIER U, WORD U + CARRIER A, and WORD U + CARRIER U, slope #1 elicited the most responses as the final-accented word, and the responses as the final-accented word decreased toward slope #7.

For the *hasi* pair, the results from the ANOVA revealed significant main effects for all the fixed effects: $\chi^2 = 39.49, p < 0.01$ for source PA; $\chi^2 = 880.23, p < 0.01$ for carrier; and $\chi^2 = 684.22, p < 0.01$ for slope. Two-way interactions were significant for source PA and slope ($\chi^2 = 30.72, p < 0.01$) and carrier and slope ($\chi^2 = 23.94, p < 0.01$). The two-way interaction between source PA and carrier and the three-way interaction...
interaction did not reach significance ($\chi^2 = 0.0005, p > 0.1$ and $\chi^2 = 5.84, p > 0.1$, respectively). Figure 7 shows that the effect of carrier (downstep) was relatively strong with no overlap between CARRIER A and CARRIER U. Probabilities of response as the final-accented word clearly higher for CARRIER A than for CARRIER U, showing the relevance of downstep in perceiving accent in the preceding phrase, consistent with findings in previous work, such as Beckman and Pierrehumbert (1986) and Poser (1984). Within the stimuli created with the same carrier, listeners’ responses were somewhat influenced by the accent information within the accentual phrases (WORD A or WORD U) when the slopes were ambiguous between the final-accented word and the unaccented word (i.e., slopes #4 and #5). However, the pattern is opposite from what was expected. For some reason, WORD U elicited more final-accented responses than WORD A.

The hati pair showed the same pattern of statistical results as the hasi pair. ANOVA found significant results for all the main factors: $\chi^2 = 37.31, p < 0.01$ for source PA; $\chi^2 = 192.38, p < 0.01$ for carrier; and $\chi^2 = 1549.07, p < 0.01$ for slope. Two-way interactions were reliable for source PA and slope ($\chi^2 = 12.74, p < 0.05$) and carrier and slope ($\chi^2 = 14.00, p < 0.05$). The two-way interaction between source PA and carrier was not reliable ($\chi^2 = 1.28, p > 0.1$); nor was the three-way interaction of source PA, carrier, and slope ($\chi^2 = 3.39, p > 0.1$). While the statistical results were similar between the hasi pair and the hati pair, a comparison of Figs. 7 and 8 reveals that each factor had a different influence on the listeners’ responses. Slope had a greater impact on the listeners’ responses for the hati pair than for the hasi pair, which can be confirmed by listeners’ responses spanning wider ranges for the hati pair. On the other hand, the effect of carrier was clearer for the hasi pair, showing a greater effect of downstep for this pair. The overall pattern of probabilities of responses as the final-accented word is more similar to that of the hana pair than to that of the hasi pair.

For the osu pair, the ANOVA revealed significant main effects for source PA ($\chi^2 = 204.04, p < 0.01$) and carrier ($\chi^2 = 284.94, p < 0.01$), but the main effect of slope failed to reach significance ($\chi^2 = 6.50, p > 0.1$). A two-way interaction between source PA and slope was the only interaction that proved reliable ($\chi^2 = 17.34, p < 0.01$). The two-way interactions between source PA and carrier and carrier and slope were not significant ($\chi^2 = 2.18, p > 0.1$ and $\chi^2 = 2.88, p > 0.1$, respectively); nor was the three-way interaction ($\chi^2 = 4.08, p > 0.1$). This is the only pair for which slope did not have a significant main effect, shown in Fig. 9 as relatively flat predicted probabilities. The probability of a final-accented word response does not change as a function of slope. While we found significant effects of carrier and source PA and a two-way interaction between source PA and slope, the figure suggests that listeners’ responses were not as consistent, as can be confirmed by the wide confidence intervals. Listeners’ responses for word A + carrier U are hardly separable from the ones for word U + carrier A. It seems as if listeners’ judgments were neutralized by the conflicting information from carrier and slope. Probabilities of listeners’ responses for the final-accented word are clearly higher for word A + carrier A than for word U + carrier U, but slope does not seem to exercise a strong influence. A possible factor for this finding will be explored further in Sec. IV.

ANOVA for the sita pair found significant main effects of carrier ($\chi^2 = 75.40, p < 0.01$) and slope ($\chi^2 = 1416.96, p < 0.01$), while the main effect of source PA was not significant ($\chi^2 = 0.35, p > 0.1$). A significant interaction was found for carrier and slope ($\chi^2 = 15.89, p < 0.05$), but the other interactions were not significant ($\chi^2 = 0.14, p > 0.1$ for source PA and carrier; $\chi^2 = 4.88, p > 0.1$ for source PA and slope; and $\chi^2 = 3.50, p > 0.1$ for the three-way interaction of source PA, carrier, and slope). For this pair of words, source PA had little influence on the listeners’ judgments. For all of the four conditions, the listeners’ judgments were affected by slope with slope #1 eliciting the most responses as the final-accented words, as expected. The effect of carrier was slightly stronger for slopes #4 and #5, which seems to be responsible for the two-way interaction between carrier and slope. It suggests that listeners relied more on the information from carrier when other acoustic information was not clear.

IV. DISCUSSION AND CONCLUSION

The statistical results for all pairs are summarized in Table IV. The table shows that the main effect of carrier was significant for all pairs, while the effects of slope and source PA were not uniform across different pairs of words. The two-way interaction between source PA and carrier was not reliable for any of the pairs. The three-way interaction of source PA, carrier, and slope was not reliable for any of the pairs. The two-way interaction between carrier and slope was reliable for the hasi, hati, and sita pairs; it was marginal for the hana pair; and it was not reliable for the osu pair. A
two-way interaction between source PA and slope was significant for the hasi, hati, and osu pairs, but it was not reliable for the hana pair or the sita pair.

While fewer pairs of words and only the carrier taken from unaccented words were used in Hui et al. (2019), the results found in the current study basically replicate their results. Their results, found for elderly listeners who had relatively good difference limens for frequency, were similar to the results found in the current study. In both studies, listeners’ responses show a clearer effect of the f₀ contour in the accentual phrase for the hana pair than for the sita pair. Some minor difference can also be observed between the two. While the results for elderly listeners and young listeners are similar for the hana pair, the effect of the f₀ contour appears to be stronger for young listeners than for elderly listeners for the sita pair, suggesting that young listeners were more sensitive to change in f₀ than elderly listeners.

### A. The effect of slope on perceiving PA

Probabilities of responses for the final-accented words illustrated in Figs. 6–10 provide further insights in assessing how source PA, carrier, and slope affected listeners’ responses. The listeners’ judgments were clearly affected by slope for the pairs hana, hati, and sita. The listeners’ responses indicating final-accented words clearly decreased as the slope number increased (as the slopes became flatter). Comparing the three pairs, the listeners’ responses were somewhat more weakly related to slope for the hana pair; they did not change as clearly as they did for the three pairs. For the osu pair, the listeners’ responses were little affected by slope, as evidenced by the virtually flat lines in Fig. 9. These varying results among the pairs made us suspect that the size of f₀ fall might have been different depending on the pair. If this was the case, it would not come as a surprise if the slope number had little effect on the listeners’ responses for some of the pairs. To investigate this possibility, the f₀ difference between the word-final syllable and the following particle was measured by subtracting the minimum f₀ of the particle from the maximum f₀ of the word-final syllable. As the f₀ tracks in Figs. 2 and 4 show, f₀ typically falls for final-accented words (slope #1), whereas it stays relatively flat for unaccented words (slope #7). Therefore, f₀ fall should be larger for final-accented words than for their unaccented counterparts. In other words, the value would be positive when the size of f₀ fall for an unaccented word is subtracted from its final-accented counterpart. Table V summarizes the result of the measurements. The fact that the values are all positive indicates that the f₀ fall was greater for the final-accented word than for the unaccented word for all the pairs, which is consistent with the phonological characterization of Japanese PA (Kubozono, 1988; Pierrehumbert and Beckman, 1988). At the same time, it is also the case that the size of the difference is not uniform across the pairs. The difference is roughly 40 Hz for the pairs hana, hati, and sita, while it is less than 30 Hz for the osu pair and only a little over 10 Hz for the hasi pair. Referring to Figs. 6–10, it is noticeable that the pairs that had a relatively large difference in the f₀ fall (i.e., the hana, hati, and sita pairs) had a clearer effect of slope, where the listeners’ judgments as final-accented words decrease as the slope number increases. However, when the results for the hasi pair and the osu pair are compared, we see that the size of the f₀ fall does not solely determine listeners’ judgments; although the f₀ fall was greater for the osu pair than the hasi pair, the listeners’ responses were less influenced by slope for the osu pair than the hasi pair. It suggests that some acoustic information other than the f₀ fall at the accentual phrase affected the listeners’ judgments, which we turn to in Secs. IV B–IV D.

#### B. The effect of carrier on perceiving PA

The effect of carrier was significant for all the pairs, although Figs. 6–10 show that the magnitude of its effect differed from one pair to another. Comparing the probabilities for carrier A (solid lines) and carrier U (dotted lines), the former elicited more responses as an accented word than the latter, indicating that downstep can be a cue to PA in the preceding phrase. The result indicates that listeners used the f₀ information after the accentual phrase (i.e., the presence of downstep or its absence) to judge if the test word in the accentual phrase was a final-accented word or an unaccented word. Because different tokens of suki were used for the five pairs, we explored whether the magnitude of downstep was related to the listeners’ perception. Since the word-initial syllables were completely voiceless because of vowel

### TABLE IV. Summary of statistical analyses. In the table, an asterisk indicates a significant effect (p < 0.05), “m” indicates a marginal effect (0.05 ≤ p < 0.10), and “n.s.” indicates non-significant effect (p ≥ 0.10).

<table>
<thead>
<tr>
<th>Carrier</th>
<th>hana</th>
<th>hasi</th>
<th>hati</th>
<th>osu</th>
<th>sita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Source PA</td>
<td>m</td>
<td>*</td>
<td>*</td>
<td>n.s.</td>
<td>*</td>
</tr>
<tr>
<td>Carrier * slope</td>
<td>m</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Source PA * slope</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Source PA * carrier</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
</tbody>
</table>

### TABLE V. Differences in the f₀ fall from the word-final syllable into the following particle between final-accented words and unaccented words (Hz).

<table>
<thead>
<tr>
<th>Word</th>
<th>hana</th>
<th>hasi</th>
<th>hati</th>
<th>osu</th>
<th>Sita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
<td>39</td>
<td>11</td>
<td>42</td>
<td>27</td>
<td>43</td>
</tr>
</tbody>
</table>
devoicing, \( f_o \) measurements were possible only for the word-final syllable. Given that *suki* was subject to downstep after an accented word, \( f_o \) movements would be compressed with a lowered peak when the word followed a final-accented word. By contrast, \( f_o \) would be expected to show an ascending pattern toward the word-final accented syllable when it followed an unaccented word. We subtracted the maximum value of \( f_o \) in the word-final syllable of *suki* following a final-accented word from the maximum \( f_o \) of *suki* following the corresponding unaccented word. Table VI summarizes these figures. The results were consistent with the prediction except for the *osu* pair, which showed a higher \( f_o \) maximum for the word following the final-accented word than for the one following the unaccented word. A closer inspection of the stimuli found that the final vowel was creaky with very low \( f_o \) for the one following the unaccented word, which was rather unusual. Since the final word was not subject to downstep, the \( f_o \) should be typically high for the word-final accented vowel. The low \( f_o \) may have given rise to the perception that the preceding phrase contained a final-accented word. The result for the *osu* pair may be interpreted as an example where the \( f_o \) contour in the following phrase had a strong effect on the perception of accent in the preceding phrase. While the values in Table VI are the same for the *hasi* and *sita* pairs, the listeners’ responses are more clearly separated by carrier for the *hasi* pair than the *sita* pair. Although it is beyond the scope of the current study, this suggests that something other than the simple difference in \( f_o \) in carrier affected the listeners’ judgments.

C. The effect of source PA on perceiving PA

The acoustic information from source PA had an unclear and inconsistent effect on the listeners’ judgments. The effect of non-\( f_o \) information was most clearly observed in the *osu* pair in the way expected. Probabilities of response as the final-accented word were higher for word A (gray lines) than for word U (blue lines). The effect of source PA was not clear for the other pairs. For the *hasi* and *hati* pairs (Figs. 7 and 8), it is puzzling that, if anything, *word U* shows higher probabilities of responses as final-accented words than word A. As discussed in the Introduction, the motivation for creating stimuli using both final-accented words and unaccented words was that if lexical accent manifested itself in acoustic cues other than in \( f_o \), stimuli created with final-accented words were likely to elicit more responses as final-accented words than the ones created with unaccented words. The inconsistent and unexpected effect of accent needs further investigation, but at least the results of the present study show that non-\( f_o \) acoustic information does not act as strong secondary cues. Alternatively, the \( f_o \) is so dominant as a cue that it overrides other acoustic information as long as it is present. Research on tone languages suggests that when the \( f_o \) is present, its perceptual saliency conceals other weaker cues to tone identity (Brunelle, 2012; Liu and Samuel, 2004). It is possible that \( f_o \) has a similar effect in PA in Japanese. To identify secondary cues to PA, a perception experiment with stimuli that contain no \( f_o \) or its harmonics, such as the ones used in Sugiyama (2017), but with more controlled stimuli would be needed. As discussed in Sec. 1A, past research typically found no correlation between duration and accent. The stimuli used in the current study were consistent with this finding. Table VII shows duration differences between syllable-final accented syllables and their unaccented counterparts. Since the duration of the unaccented syllable was subtracted from the duration of its corresponding final-accented syllable, the value would be positive if the duration of the final-accented syllable is longer than the duration of the unaccented syllable. The results indicate that duration seems to have no consistent relation to accent. Duration difference seems to have no clear relation to the listeners’ word identification either (Figs. 6–10). The same holds true for word duration, as shown in Table VIII. Here, the duration of an unaccented word was subtracted from the duration of its corresponding accented word. The positive values indicate that word duration was longer for a final-accented word than its unaccented counterpart. Duration difference shows no clear relation to accent or the listeners’ judgments. It is not likely that syllable duration or word duration can be a secondary cue to PA in Japanese. Studies that measure intensity typically find that accented syllables tend to have greater intensity than unaccented syllables (Beckman, 1986; Sugiyama, 2017; Weitzman, 1970), but they also find that the intensity difference is relatively small, making it less clear if intensity would be a cue to PA in Japanese. The stimuli used in the current study are consistent in that, although intensity difference was consistently greater for accented words than unaccented words, the difference seems too small to be a reliable cue in natural speech. Table IX shows the difference in intensity fall between accented words and unaccented words. In this measurement, first, intensity fall was measured by subtracting the minimum intensity of the particle from the maximum intensity of the word-final syllable for each word. Then

<table>
<thead>
<tr>
<th>hana</th>
<th>hasi</th>
<th>hati</th>
<th>osu</th>
<th>sita</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>37</td>
<td>32</td>
<td>−78</td>
<td>37</td>
</tr>
</tbody>
</table>

TABLE VI. Differences in the maximum \( f_o \) values in the final syllable in *suki* between final-accented words and unaccented words (Hz).

<table>
<thead>
<tr>
<th>hana</th>
<th>hasi</th>
<th>hati</th>
<th>osu</th>
<th>sita</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>46</td>
<td>66</td>
<td>−16</td>
<td>−7</td>
</tr>
</tbody>
</table>

TABLE VIII. Differences between accented words and unaccented words (ms).
intensity fall for an unaccented word was subtracted from its accented counterpart. The values were all positive for this measurement, indicating that intensity fall was greater for accented words than for unaccented words. According to Fletcher (1953), difference limen for intensity for pure tone is less than 1 dB when it is presented at 40 dB. Considering that the stimuli used in the present study were natural speech with dynamic properties where frequencies of complex tones and intensity change over time, it seems difficult for intensity alone to be a reliable cue to PA.

### D. Other factors that may influence listeners’ word identification

As explained in Sec. II C, we tried to control for how natural the carrier sentences would sound with the test words. Some readers may wonder if some sentences still sounded semantically more natural than others. Listeners were not asked any specific questions about the semantic naturalness of the sentences in the present study. However, in perception or production studies, participants often report some test words are more natural than others when they feel so even though an experimenter does not specifically ask the question. The fact that listeners did not say anything specific about the naturalness of the sentences probably indicates that they were more or less comfortable with all the sentences.

Further research is needed to determine the contribution to the perception of lexical PA from downstep or its absence (carrier), the \( f_o \) contour within the accentual phrase in question (slope), and secondary cues to lexical PA (source PA). When findings from Sugiyama (2012), Sugiyama (2017), and the current study are viewed as a whole, they suggest that different parts of an utterance contribute to word recognition. When an accentual phrase consisting of a test word and a particle was taken out of a carrier sentence and presented to listeners, listeners’ word identification was not as accurate as expected, reaching only up to 70% or so (Sugiyama, 2012). When a whole utterance was presented to listeners, their accuracy was approximately 96% (Sugiyama, 2017). In both studies, the carrier sentences had the same structure of subject–object–verb, where test words were placed in the object position. It is not possible to compare results from two different studies directly, but it suggests that downstep (or its absence) in the verb phrase carries important information in judging lexical accent in the preceding phrase. The results from the present study are consistent with those found in the previous studies. They uphold the idea that listeners’ word identification was influenced by both the \( f_o \) contour in the accentual phrase containing the test word and the \( f_o \) contour in the following phrase, which indicates information about downstep. What happens if a sentence consisting of only a subject and a predicate, such as \( kore wa ___ da \) (this TOPIC ___ be-PRESENT), is presented to listeners? If listeners can correctly identify minimal pairs of lexical accent at a rate close to 100%, it would mean that speakers compute how much accent information to include in a certain phrase based on the whole structure of the utterance as they speak and produce the utterance accordingly.

To keep the stimuli in the present study as natural sounding as possible, for each pair of words, the initial and final parts of the carrier sentence (i.e., \( watsai wa \) and \( suki \)) were taken from their own utterances. In other words, these parts of the carrier sentence were not acoustically identical across different pairs of words. While it was an option to use exactly the same tokens, this approach was not chosen because the resulting stimuli did not always sound natural. This leaves room for the possibility that token differences may have made some test words sound more accented than others, especially when combined with \( suki \), which includes information about downstep. Another possible factor for varying results for different pairs was that the shape of \( f_o \) fall was not identical among the test words. It would be a challenge to control acoustic variations among different test words while keeping the stimuli sounding natural. However, since the results obtained from the present study suggest multiple acoustic cues play a role in an intricate way to affect the perception of Japanese lexical accent, a next step would be to control these variations. This should enable researchers to determine how much each part in the utterance contributes to word identification. Other limitations were the number of words tested and the types of accent contrast included. To gain greater general understanding of acoustic properties of Japanese PA, more words should be tested for each accent type. Furthermore, since contrast in accent is not limited to final-accented words and unaccented words, examining other accent patterns such as initial-accented words would be necessary for future research.

### ACKNOWLEDGMENTS

We would like to thank Doug Roland for his advice on conducting the statistical analyses of the data. We would also like to thank Associate Editor Benjamin V. Tucker for his constructive comments on an earlier version of the manuscript. This research was partly supported by the Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research (C) No. 19K00584.

---

1In this paper, “lexical PA” is often simply referred to as “PA” or “accent.” The term “pitch” is used interchangeably with “\( f_o \).”

2The particle following test words was also changed from to to go because there are some views about Japanese prosody that hold that the citation particle places a strong focus on the preceding word (i.e., test words) and the word is produced as if produced in isolation.

3In this particular token, affected by the initial voiceless consonant, the \( f_o \) is relatively high for the initial syllable of the test word. Readers will also notice that the \( f_o \) is actually slightly higher for the particle than for the word-final syllable of the test word, which is common for the particles following unaccented words.

4See Ishihara (2016) for an alternative view of downstep.


