Pitch Patterns of Intonational Phrases and Intonational Phrase Groups in Native and Non–Native Speech

Hiroko Hirano¹ and Goh Kawai²

¹ University of Tokyo, Department of Frontier Informatics, Tokyo, Japan
hiran@gavo.t.u-tokyo.ac.jp
² Hokkaido University, Institute of Language and Culture Studies, Sapporo, Japan
goh@kawai.com

Abstract
We examined pitch patterns within and across intonational phrases of Japanese read aloud by native and non-native (Mandarin Chinese) speakers. Japanese speakers change pitch ranges for each intonational phrase. The relative pitch ranges of neighboring intonational phrases indicate which intonational phrase belongs to which intonational phrase group. Chinese speakers are unable to acoustically convey intonational phrase groups because their intonational phrases have limited and inflexible pitch ranges. Our results might form the basis for developing automated pronunciation learning systems that assist learners in acquiring intonation contours spanning intonational phrases.

1. Introduction
When learning to speak a non-native language, acquiring appropriate prosody is as important as saying segmentals. However, scant attention has been paid to comparing native and non-native pitch patterns. Particularly in the case of learning Japanese as a foreign or second language, instructor-led classroom instruction often omits prosodic training, partly due to a lack of instructor knowledge (“can’t teach what you can’t see in print”), a reliance on language universals (“learners will figure it out somehow because all languages have similar prosody”), and paucity of pronunciation coursework (“prosody belongs to the performing arts”). Instructors are beginning to realize that, for Japanese at least, prosody affects nativeness and naturalness more than segmentals. Non-native prosodic disfluencies yield utterances that are sometimes subjectively referred to as “choppy”, “disjoint”, “harsh” or “rude”. Exactly how native and non-native pitch patterns differ remains unknown.

We wish to suggest a systematic method for analyzing and comparing native and non-native pitch patterns within and across intonational phrases. Eventually, we hope this method can be used to develop prosody coursework, and to automatically determine pitch errors so that learners can receive corrective feedback from a computerized tutor.

2. Speech data
Chinese learners comprise over two-thirds of learners of Japanese as a second language. We collected speech from 10 Mandarin-dialect speakers and 10 Tokyo-dialect speakers. Subjects read a script containing common greetings, short phrases, sentences and a fairly long passage. Recordings were made in a soundproof video recording studio, using a Sony TCD-D10PRO digital audio tape recorder (sampling rate 48 kHz, 16 bits) and a Sony C-38B microphone placed at a 45-degree angle 60 cm away from the subject (desktop, monaural, cardioid directivity, FET condenser, frequency response 80–18000 Hz). Total speech time was approximately 20 minutes per speaker. To avoid pitch artifacts caused by lexical pronunciation errors such as incorrect pitch accent, we chose 15 sentence types for analysis (300 utterances, 324 mora each). Total speech durations were 420 seconds for natives and 614 seconds for non-natives.

A native Japanese language instructor trained in phonetic labeling labeled utterances at the phone level using [1]. Intonational phrase boundaries were marked wherever intonational phrase boundaries were perceived. There were 569 intonational phrases for natives, and 760 for non-natives.

3. Speech analyses
3.1. Utterance durations
Utterance durations were normalized separately for natives and non-natives by dividing the duration of each utterance by the median duration for that sentence (for instance, if sentence 14’s median duration was 3.8 [s] for natives, then each native utterance’s duration was divided by 3.8).

Table 1 summarizes speech durations of native and non-native utterances. The table shows that, on average, non-natives speak slowly. Natives and non-natives have roughly the same normalized minimum durations, but the natives have larger normalized maximum durations with greater standard deviation. These differences are statistically significant (p<0.01).

This means that utterances spoken the fastest sounded as having the same speed for natives and non-natives, but utterances spoken the slowest sounded slower for natives. Natives produce various durations; non-natives do not.

Table 1: Durations and normalized durations for 10 natives and 10 non-natives reading 15 sentences.

<table>
<thead>
<tr>
<th></th>
<th>native</th>
<th>non-native</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean duration</td>
<td>2.80</td>
<td>4.10</td>
</tr>
<tr>
<td>min duration / median</td>
<td>0.62</td>
<td>0.64</td>
</tr>
<tr>
<td>max duration / median</td>
<td>1.57</td>
<td>1.33</td>
</tr>
<tr>
<td>S.D. of (duration / median)</td>
<td>0.24</td>
<td>0.15</td>
</tr>
</tbody>
</table>

3.2. Pitch contours within intonational phrases
In [2], we saw a non-native use narrow pitch ranges and the same pitch contour regardless of context. Our present study confirms this finding. Non-natives have a remarkably limited pitch repertoire, at least partly due to their inability to control pitch over multiple syllables—most pitch changes occurred
abruptly within one or two syllables, suggesting L1 interference. Pitch ranges are limited—both pitch analyses and impressionistic listening indicate what might be generalized as a high–low binary pitch distinction, either due to L1 interference and/or high–low lexical pitch accent training for L2. The limited pitch ranges and abrupt changes give non-native pitch contour a pulse-train shape.

By comparison, natives' pitch contours are hump-like curves that gradually rise and fall over many syllables. Pitch ranges vary considerably according to the degree of prominence. Figure 1 compares pitch contours of a noun phrase “Hokkaido [+particle]”.

![Figure 1](image1.png)

Figure 1: A non-native and a native saying the noun phrase “Hokkaido [+ particle]” in 7 different contexts. Horizontal axes are intonational phrase lengths normalized for time.

Table 2 shows normalized pitch for the declarative sentence “meganeno niau naoyano anini indono ryoorio naraini iku”. Figure 2 shows a typical native and a non-native saying this sentence. Most natives would say this sentence in 4 intonational phrases. We compared pitch maxima and minima using pitch normalized by dividing non-zero pitch values by the median pitch value of the utterance.

Both natives and non-natives adjust their pitch range by raising their pitch above their baseline pitch. Although the baselines are comparable, natives have higher and more varied maxima. In non-read speech, pitch maxima are expected to rise and vary further. This means that non-natives should be trained to raise pitch maxima. No instruction seems necessary for pitch baselines.

Pitch minima tend to occur towards the ends of utterances. Vowels following unvoiced sibilants tend to become devoiced or deleted. Of the 150 utterance-final vowels in the 15 sentences studied, natives devoiced or deleted 55, while non-natives devoiced or deleted 32. While voicing per se does not degrade intelligibility, non-native voicing is often accompanied by loudness carried through to the end of the utterance, terminated by an abrupt pitch drop that natives perceive as harsh or overly strong. Decreasing loudness gradually towards the end of the sentence allows content-bearing portions of the utterance to bear prominence. Instructing non-natives to delete vowels may indirectly assist them to acquire effective prominence skills.

### 3.3. Pitch contours at intonational phrase boundaries

In [2], we saw a non-native often unwittingly use acoustic cues that natives may perceive as intonational phrase boundaries. Our present study confirms this finding. More intonational phrases were found in non-native speech because pitch control was ambiguous. Natives combine multiple acoustic cues (such as pitch changes, segmental lengthening and pause insertions) to unambiguously mark intonational phrase boundaries. We focused on pitch as the first order of approximation.

![Figure 2](image2.png)

Figure 2: A typical native and non-native saying the declarative sentence “meganeno niau naoyano anini indono ryoorio naraini iku”. Both time and pitch are raw values.

![Figure 3](image3.png)

Figure 3: Native and non-native means of maximum, median and minimum non-zero pitch values within each intonational phrase Above: Native intonational phrase groups become increasingly ambiguous because maximum pitch values start small, resulting in a loss of pitch headroom. n=10 each.

| Table 2: Durations and normalized pitch for 10 natives and 10 non-natives reading a sentence. Pitch was normalized by dividing min and max pitch values by the median non-zero pitch value of each utterance. n=10 each. |
|---------------------------------|-----------------|-----------------|
| mean duration [s]              | native          | non-native      |
| S.D. of duration               | 0.40            | 1.21            |
| min pitch / median             | 0.67            | 0.66            |
| S.D. of (min pitch / median)   | 0.08            | 0.05            |
| max pitch / median             | 1.77            | 1.32            |
| S.D. of (max pitch / median)   | 0.18            | 0.10            |
Natives and non-natives have much in common with minimum pitch value trends, but the maximum values are different. Figure 3 shows means of maximum, median and minimum non-zero pitch values within each intonational phrase (i.e., for instance, for each intonational phrase, the maximum non-zero pitch values for natives and non-natives were averaged).

The shaded lines in the Figure 3 show that native pitch maxima drop steeply for the first 3 intonational phrases, followed by an arrow indicating a pitch rise (but not rising as high as the 1st phrase), then dropping again, followed by another rise (again, not as high as the 1st phrase of the preceding group). Impressionistic listening confirms that these blocks of phrases sharing a sequential pitch drop correspond to intonational phrase groups. An intonational phrase starts a new intonational phrase group when its maximum pitch value is greater than its preceding phrase. When the maximum pitch value of an intonational phrase group is larger than its preceding group, a new intonational sentence begins.

Non-native pitch changes are similar, perhaps due to L2 learning and/or language universals. However, the slopes of pitch maxima are not as steep as natives (compare the shaded lines in the top and bottom figures). This means that pitch rises are of necessity smaller than natives—indeed, because pitch rises become successively smaller towards the end of the intonational sentence, non-natives eventually run out of headroom (see the flat arrow to the right of the bottom figure).

Obviously natives can exhaust their pitch maneuvering range too—however this happens for longer intonational sentences, consistent with sophisticated pitch–planning associated with natives. This suggests that one measure of fluency may be the number of intonational phrases a person can say before they run out of pitch–rise headroom.

By combining the two figures in Figure 4, we obtain an acceptance range of pitch. (This is analogous to voltage tolerances when testing electronic circuits with an oscilloscope—measurements from the circuit being tested must fall within a specified range.) We define the target as maximum pitch values that are within natives and above non-natives. Exceeding the natives’ maximum pitch will be tagged as an error (“overdoing it”, so to speak). An example is the last intonational phrase in Figure 4, where non-natives did not drop pitch fast enough.

3.4. Summary of speech analyses

Table 3 summarizes our findings. Figure 5 shows stylized schematics of pitch contours.

The key to nativeness is variety. The degree of freedom in the time and pitch domains is a consistent characteristic of native speech. Although there is ample evidence that natives speak faster than non-natives in almost all languages, speech rate per se is probably not the unique factor. Rather, natives are capable of speaking at various speeds, while non-natives are usually slow. Likewise, natives have a wide pitch range, whereas non-natives are restricted to a narrow, low frequency region. The limited room to maneuver prevents non-natives from employing a wide array of acoustic cues that characterize nativeness.

Another key feature that goes hand in hand with limited room to maneuver is combining multiple acoustic cues to clearly convey the speaker’s intent. Natives’ intonational phrases are perceived consistently because natives choose cues redundantly. Non-natives tend to emit cues in isolation—as a result they fail to unambiguously distinguish between intended and unintended events. Saying each cue separately at first, and then progressing to using them in unison might train cue coordination.

Figure 4: Native and non-native pitch ranges overlaid. The overlapping shaded areas correspond to native and non-native pitch ranges. Learners can be instructed to use pitch that falls within the native–only area. The last intonational phrase shows that non-natives did not drop pitch fast enough. n=10 each.
Table 3: Summary of differences between native and non-native speech. n=10 each.

<table>
<thead>
<tr>
<th></th>
<th>native</th>
<th>non-native</th>
</tr>
</thead>
<tbody>
<tr>
<td>durations</td>
<td>short, uniform</td>
<td>long, varied</td>
</tr>
<tr>
<td>pitch range</td>
<td>large, varied</td>
<td>small, varied</td>
</tr>
<tr>
<td>maximum pitch</td>
<td>large, varied</td>
<td>medium, uniform</td>
</tr>
<tr>
<td>intonational phrase boundaries</td>
<td>clear, combines multiple acoustic cues</td>
<td>ambiguous, uses few and/or weak cues</td>
</tr>
<tr>
<td>intonational phrase groups</td>
<td>downstep + pitch range adjustments</td>
<td>absent or weaker than natives</td>
</tr>
<tr>
<td>sentence final</td>
<td>low pitch, small loudness</td>
<td>high pitch, abrupt drop, loud loudness</td>
</tr>
</tbody>
</table>

Figure 5: Stylized schematics of typical native and non-native intonational phrase groups and their corresponding intonational phrases. Natives form gentle humps. Non-natives have limited pitch ranges and abrupt rise–falls that form a pulse train.

4. Discussion

4.1. Automated method for intonational phrase grading

We intend to develop an automated tutor for intonational phrase learning using the following strategy:

- Select reading material. Longer, more complex sentences are more challenging. Include voiced segments for reliable pitch estimation.
- Record natives and non-natives reading the material.
- Force-align recordings at the phone level using a speech recognizer, and match phones with pitch.
- Determine acceptable pitch range, as shown in section 3.3.
- Record learners reading material.
- Force-align and extract pitch.
- Determine if pitch is within acceptable range.
- Provide learner with feedback.

4.2. Implications for instructor-led learning

Japanese language instructors may find the following relevant:

- Remedy errors at the beginning level (but do not discourage students).
- Pitch errors persist through intermediate and advanced levels. Continue pronunciation practice.
- Help learners make a good oral impression. Advanced learners may be particularly motivated.
- Segmental errors affect intonational phrase perception. Chinese lexical tones interfere with intonational phrases. Practice segmentals and suprasegmentals together.

5. Conclusion

Compared with Japanese speakers, Chinese speakers generate (a) utterance with durations of larger means and variance, (b) a larger number of intonational phrases, and (c) intonational phrase boundaries that are perceived less consistently by native speakers.

Japanese speakers produce a series of gentle hump-shaped pitch contours that vary widely according to context. Chinese speakers produce linear step-shaped with reduced pitch ranges, abrupt pitch changes, and limited variation.

Japanese speakers change pitch ranges for each intonational phrase. The relative pitch ranges of neighboring intonational phrases indicate which intonational phrase belongs to which intonational phrase group. Chinese speakers are unable to acoustically convey intonational phrase groups because their intonational phrases have limited and inflexible pitch ranges.

Japanese speakers produce prosody that is rich, expressive, consistent, and facilitate understanding. Chinese speakers, however, generate monotonous prosody that hinders communication. Given that these mistakes persist into the intermediate and advanced learning levels, prosodic training should probably be intensified at the beginning level.

This study’s results might form the basis for developing automated pronunciation learning systems that (a) explicitly explain lexical pitch accent rules and prosodic expressions, and (b) assist learners in acquiring gentle hump-shaped curves spanning multiple syllables.

References