

Problems with Audio Recordings for Elementary School English Textbooks in Japan¹

Nobuo YUZAWA

Faculty of International Studies

Utsunomiya University

Simon FRASER

Institute for Foreign Language Research and Education

Hiroshima University

In 2020, English became an official subject for 5th and 6th graders in Japan for the first time. To support this new initiative, seven publishing companies created textbooks specifically designed for these students to use when learning English: Gakko Toshō (*Junior Total English*), Kairyudo (*Junior Sunshine*), Keirinkan (*Blue Sky*), Kyoiku Shuppan (*One World Smiles*), Mitsumura Toshō (*Here We Go*), Sanseido (*Crown Jr.*), and Tokyo Shoseki (*New Horizon Elementary English Course*).² Each company produced two versions of their textbooks: one tailored for 5th graders and the other for 6th graders. All these textbooks, approved by MEXT (the Ministry of Education, Culture, Sports, Science and Technology), were introduced in April of the same year.

Given that English is not commonly used in daily life in Japan, the audio materials accompanying these textbooks play a crucial role in providing students with models of spoken English. In addition, MEXT (2017, p. 17) states that one of the three purposes of learning foreign languages in elementary school is to familiarise students with the *sounds* [emphasis added] and basic expressions of foreign languages. Therefore, it is worthwhile to conduct research into the quality and effectiveness of these audio materials for young learners. Yuzawa (2022a, 2022b, 2022c) examined these recordings, focusing on intonation, and found several notable problems, including:

1. Lack of uniformity in intonation patterns within the same context, where consistent patterns are preferred.
2. Misplacement of the tonic syllable, resulting in incorrect stress patterns without specific reasons.
3. Unnecessary use of a high pitch for sentence-initial pronouns. This may be connected to the common habit of many Japanese learners of English placing a high pitch on “I” in expressions like “I think.”

The present study is a continuation of Yuzawa’s research. In the study, the impressionistic judgment of Fraser, an L1 (first-language) English speaker, is very important. This approach aims to shed new light on the issues Yuzawa has addressed over the past two and a half years, based on the expectation that an educated L1 English speaker can make accurate judgements about the naturalness of spoken English. The authors of this paper believe that this naturalness is an important factor that should be employed in making a model for learners, especially those who learn English as a foreign language, so that they can develop “English ears,” the ability to understand spoken English like their mother tongue.

The purpose of this paper is, then, to examine the intonation patterns in the audio recordings of two English textbooks for Japanese elementary school students, as perceived by an L1 English speaker.

METHODS

The present research utilises a mixed-methods approach, collaboratively undertaken by two researchers: Yuzawa, a phonetician, and Fraser, an applied linguist. Both have over 30 years of teaching and research experience in Japanese universities.

Firstly, Fraser employed a qualitative method to evaluate the audio materials accompanying the textbooks, assessing any unnatural aspects in the recordings based on his perceptual judgment as an L1 English speaker. Such an approach is valuable when assessing the naturalness of these materials, which serve as significant models of spoken English. This is especially pertinent in countries like Japan, where English is not commonly used in everyday communication. Upon identifying unnatural or problematic recordings, Fraser compiled and sent a list to Yuzawa, along with his own recordings using intonation patterns he deemed natural.

Secondly, adopting a quantitative method, Yuzawa conducted an acoustic analysis using PRAAT (version 6.2.14), to examine why the recordings on the list sound unnatural and to suggest improvements for the next edition of the textbooks. The results of this analysis, which include the waveform, the spectrogram, and the fundamental frequency (F0), are presented through figures when necessary. Each figure displays the waveform at the top, with the spectrogram and the F0 underneath. The F0 is depicted as a line within the spectrogram. Due to the use of greyscale in this paper, which may render the F0 contour difficult to discern, white arrows are used to highlight key points in the F0 contour.

The Textbooks

Among the seven sets of elementary school English textbooks, a set published by the Tokyo Shoseki Publishing Company was selected for this study. These textbooks are the most widely used in Japan³, including Hiroshima City, making them highly influential in the teaching and learning of English in elementary schools in Japan. They are titled *New Horizon Elementary English Course 5* and *New Horizon Elementary English Course 6*. In this paper, they are referred to as the G5 textbook and the G6 textbook, respectively, and when used with an accompanying CD, they are named G5 and G6, respectively (e.g., G5 CD1).

Data Collection

There are four accompanying CDs for the G5 textbook and three for the G6 textbook. The audio data recorded on these seven CDs are used as data in this paper. The number of tracks on each CD and the total duration are shown in Table 1.

TABLE 1. CDs Accompanying the Textbooks

CD	No. of tracks	Total duration
G5 CD1	69	48 minutes 55 seconds
G5 CD2	72	51 minutes 30 seconds
G5 CD3	65	49 minutes 16 seconds
G5 CD4	68	61 minutes 53 seconds
G6 CD1	83	52 minutes 44 seconds
G6 CD2	68	49 minutes 55 seconds
G6 CD3	68	64 minutes 13 seconds

Yuzawa extracted the audio recordings from the seven CDs, converted them into MP3 format at a bit rate of 160 kbps, and sent the files to Fraser via the internet. Then, Fraser listened to all the files, listed problematic recordings when he noticed them, and sent the list to Yuzawa. Yuzawa undertook acoustic analysis of the parts of the recordings that Fraser selected. For efficient internet transfer of large audio data, the MP3 format was adopted due to its smaller size compared to WAV format, as well as its compatibility with PRAAT. Although 160 kbps is a lower bit rate, it did not impede Fraser's impressionistic judgment, nor did it affect the F0 analysis. Furthermore, Fraser recorded his readings of the problematic segments in MP3 format at a bit rate of 705 kbps and sent these to Yuzawa, facilitating a comparative acoustic analysis with the original recordings.

RESULTS

In this section, the comments relevant to naturalness have been tabulated. These provide the data from which the key issues can be identified. A skeleton summary of the problems found on a CD precedes each table. The comments are then carefully analysed in relation to the three major issues that are identified in the audio materials.

The findings from Fraser's impressionistic analysis are presented in Tables 2–7. Tables 2, 3, and 4 show the comments made by Fraser after listening to the CDs accompanying the G5 textbook, and Tables 5, 6, and 7 list his comments on the recordings accompanying the G6 textbook.

Table 2 presents Fraser's comments on G5 CD1. His observations on this recording include perceptions of exaggerated utterances, unnatural pauses, and unnecessarily slow reading.

TABLE 2. Fraser's Comments on G5 CD1

	Section	Comment
1	Classroom English	At 0:39, "What's this?": To me, the reading of this sounds unnecessarily exaggerated, as if the speaker has seen something unexpected or unpleasant.
2	Unit 1, Let's Sing	Around 5:00, in the song, stress is placed on 'you' in "How are you?". That's probably OK, given that stress often changes to fit the rhythm of a song, but this pronunciation is not usual unless the speaker is replying with something like "I'm fine, how are <i>you</i> ?", or perhaps to express delight when addressing a person one has not seen for a long time.
3	Unit 1, Let's Watch and Think, No. 1	At 8:55, the way "What's your name?" is read sounds odd to me, and unnecessarily exaggerated (because of the rise-fall at the end?)
4	Unit 1, Let's Watch and Think, No. 2	At 9:30, a different pattern of intonation is used with "What's your name?" here. I'm not really sure which would be better, but it would be good to be consistent.
5	Unit 1, Let's Watch and Think, No. 2 and Let's Listen 1	At 9:39 and 10:40, "How do you spell your name?" is pronounced differently: in the first instance 'name' seems to be drawn out unnecessarily (rise-fall?); in the latter it seems to be more straightforward. The example at 11:04 seems reasonable to me, but that at 11:30 is different again, ending high, with the emphasis on "name" (why?). This variety seems unnecessary, and may be confusing for the students.
6	Unit 3, Let's Listen 2	At 44:40, there seems to be an unnatural pause between 'study' and 'home economics', and also in the sentence "I want to be a baker" (c.f. 45:39 "I want to be a scientist", where the flows seem natural, without a pause. There are other similar examples like this throughout the recordings. I found it difficult, however, to notice such instances, as allowance has to be made for the unnaturally slow rate of speech.

Table 3 presents Fraser's comments on G5 CD2. His observations here relate to the inappropriate use of stress and unnecessarily exaggerated reading.

TABLE 3. Fraser's Comments on G5 CD2

	Section	Comment
1	Unit 4, Enjoy Communication	The questions "Who is this?" and "Who is Mark Smith?" (15:50) came across as rather inquisitorial to me, rather than normal conversational style (also at 16:23).
2	Unit 5, Enjoy Communication	At 32:19, "You're welcome" – with a rise and fall, primary stress on first syllable of "welcome", seems appropriate to me; however, earlier there is an instance of the sentence with the primary stress on "You're" – we do sometimes hear this, but it isn't usual, and may lead to confusion.
3	Unit 6, Let's Watch and Think	At 44:00, . . . sounds a bit unnatural, exaggerated; also at 44:22 "potatoes and some spices are inside" – is such a high pitch necessary? Presumably to capture the students' attention and interest?

Table 4 lists Fraser's comments on G5 CD3. Here, he notes an unnecessary variety of intonation patterns presented to the listener, as well as what he perceives to be unnaturally slow reading by the narrators.

TABLE 4. Fraser's Comments on G5 CD3

	Section	Comment
1	Unit 7, Starting Out	At 0:21, "Happy New Year!", with the emphasis on 'New', sounds very strange to me. (I know that AmE and BrE have different preferences with regard to 'New Year's Day' and 'New Year's Day', and 'New Year's' and 'New Year', but I think in both varieties "Happy New Year!" is usual.)
2	Unit 7, Let's Try 1	At 1:18, I was expecting to hear "it's round in <i>western</i> Japan" to express contrast.
3	Unit 7, Enjoy Communication	At 2:53, We hear 'New Year's Day', the usual AmE pronunciation, whereas at 14:50 it is "New Year's Day", closer to what is usual in BrE. I think it probably is just a result of personal preference, as either is possible, but it may be confusing for the students to hear both versions.
4	Unit 8, Let's Listen 1	At 30:08, "I sometimes clean my house . . . but I never go shopping" – there seems to be an unnatural pause, but presumably it is here to emphasise the contrast, so it's probably OK.

The findings in Table 5 indicate that there were a number of problems in G6 CD1 concerning incorrect intonation patterns and placement of prominence.

TABLE 5. Fraser's Comments on G6 CD1

	Section	Comment
1	Unit 1, Starting Out	At 4:38, "Do you like chocolate? I <i>like</i> chocolate" seems unnatural; "I <i>love</i> chocolate" would work (or "I like chocolate?").
2	Unit 1, Let's Listen 2	At 14:40, "My birthday is May 5th", there is a strong emphasis on '5th', whereas at 14:50 and 15:11 a more regular pattern is used. This difference was very obvious to me, so it may well confuse the students.
3	Unit 2, Let's Try 2	I noticed different intonation patterns are used at 30:00 and 30:16 with "I usually watch soccer games on Sundays", with the former seemingly replying to a 'what' question, and the latter a reply to a 'when' question.
4	Unit 8, Let's Listen 1	At 48:37, "You can see the Eiffel Tower": I'm wondering why 'Eiffel' is stressed rather than 'Tower'; as far as I am aware <i>Eiffel Tower</i> is the norm in both AmE and BrE, and there is no comparison being made with any other tower.

Table 6 lists Fraser’s comments on G6 CD2. The main problems he noted in this recording concern the incorrect use of tonicity.

TABLE 6. Fraser’s Comments on G6 CD2

	Section	Comment
1	Unit 5, Starting Out, Let’s Read and Write	At 19:12, it seems odd to put the emphasis on ‘live’ in “Where do sea turtles <i>live</i> ?”, unless it is a follow-up question.
2	Unit 5, Your Turn, Let’s Listen 2, No. 1	At 26:00, here the stress patterns seem to be as would be expected, with the emphasis put first on ‘sea turtles’ and then on ‘eat’.
3	Unit 2, Let’s Try 2	At 27:39, “And the second question: What do <i>frogs</i> eat?”: I’m wondering why there is emphasis on ‘frogs’, as this is a follow-up question where contrast would be expected between ‘live’ and ‘eat’. I suppose the inferred contrast could be between frogs and previously mentioned sea turtles and eagles in earlier rounds of the quiz, but it must be rather confusing for the students, especially as at 28:03, where the questions are about lions, and the expected patterns are used (the emphasis first on ‘lions’, then in the second question on ‘eat’).
4	Unit 3, Enjoy Communication	At 29:20, again, “Where do sea turtles <i>live</i> ?”

Table 7 presents Fraser’s comments on G6 CD3. Here, his observations concern the use of an unnatural intonation pattern, different intonation patterns being used for the same utterance, and unnatural pausing.

TABLE 7. Fraser’s Comments on G6 CD3

	Section	Comment
1	Unit 7, Your Turn, Let’s Try 2	At 10:00, “What did you see?”: It might be kinder to the students to use a more basic intonation pattern here (I wonder whether this pattern is perhaps used because the question is a simpler form of the more usual “What did you see there?”, where the rise-fall would be natural?).
2	Unit 7, Your Turn, Let’s Read and Write 4	At 10:25 and 10:36, The students hear two different readings of “I enjoyed talking with my friends”, with ‘friends’ highlighted in the former but not the latter. At 11:03 (Enjoy Communication, Step 1), we hear the same sentence again with a fall at the end.
3	Unit 8, Starting Out	At 17:40, It sounds as though there is a slightly unnatural pause between ‘here’ and ‘enjoy’, but I may be imagining it!

ANALYSIS AND DISCUSSION

The problematic points identified by Fraser can be categorised into three main groups:

- (1) slow speech rate
- (2) incorrect tonicity
- (3) unnecessary exaggeration

In this section, acoustic analysis is employed to thoroughly investigate and discuss each of these three categories.

Slow Speech Rate

This issue is highlighted in Fraser’s sixth comment on G5 CD1. He states that (1) the recordings are

made with an “unnaturally slow rate of speech,” (2) there is an “unnatural pause” in the middle of a sentence throughout the recordings, and (3) this unnatural pause is not easily detected because of the slow rate of speech.

To confirm his intuitive judgement that the speech rate is slow, monologues⁴ were chosen to calculate the average rate. No suitable monologues were found in the G5 textbook, but two appropriate ones were identified in Units 1 and 8 of the G6 textbook. The analysis revealed a speech rate of 85 words per minute (wpm) in the former (51 words in 36 seconds) and 95 wpm in the latter (40 words in 25 seconds)⁵. Although an increase in speech rate from earlier to later units⁶ was observed, these rates are relatively low when compared with the average speech rate in conversations, which is, according to Virtual Speech (2022), between 120 and 150 wpm. In addition, as will be shown later, when the speech rate of the textbook narrators is compared with that of Fraser by using sentences that he identified as being problematic, it was found that their speech rate is 1.5 times as slow as his speech rate, even though Fraser tried his best to read these sentences as slowly as possible without losing naturalness. As this comparison indicates, there seems to be a close relationship between the slow speech rate adopted in the textbooks and the loss of naturalness. If this is the case, a faster speech rate should be considered in order to maintain naturalness.

An argument might be made in favour of a slow speech rate, considering that people adjust their speech rate depending on their listener. In the audio materials, everything is pre-recorded, and no such flexibility can be possible in class. If various speech rates are used in the recordings as learning progresses, the use of a slow speech rate in the initial stage may be valid. However, if an unnaturally slow rate of speech is used throughout the entirety of the recordings for the two-year course, there is no such validity; as Field (2008, p. 271) states, “Exposing a listener only to a graded material is like feeding a child exclusively on baby food and then wondering why the child cannot cope with an adult diet.”

One realistic solution to improve this situation may lie in the leadership of MEXT. They should develop and publicly disseminate guidelines specifying an appropriate speech rate for each course level in Japanese schools. Such guidelines should consider a gradual increase in speech rate throughout the course levels and draw from successful practices observed in other educational systems around the world. In addition, textbook publishers must ensure that their production of audio materials adheres to these guidelines.

It is also conceivable that this slow speech rate is directly connected to unnatural, long pauses between words. For instance, there is a 365 millisecond (ms) pause between “study” and “home economics” in the phrase “I want to study home economics” (G5 CD1 Track 6). Between “here” and “enjoy” in “Please make good friends here and enjoy your school life” (G6 CD3 No. 3), there is a 515 ms pause. This use of pauses, perceived as unnatural by Fraser, will naturally be resolved using a faster speech rate.

The danger is that constant exposure to such a slow speech rate may give students an inaccurate impression of how English is naturally spoken. This could impede their ability to comprehend authentic spoken English.

Incorrect Tonicity

The correct placement of the tonic syllable is important in accurately conveying a message. Levis (2018, p. 162) emphasises this, asserting, “Unexpected placements of accented words can compromise intelligibility, and thus it is important to place the tonic syllable appropriately to convey the speaker’s message correctly.”

However, as highlighted by Yuzawa (2022b, 2022c), many examples of incorrect tonic syllable placement are evident in the recordings. This section discusses such problematic examples, dividing them into two categories. The first subsection examines examples from non-contrastive contexts, while the second addresses those in contrastive contexts. Comparison with Fraser’s recordings of the same utterances are made when necessary.

Non-Contrastive Contexts

How are you? (G5 CD1 No. 2)

This expression is found in a song, and its lyrics are as follows:

Hi, hi, hi!
Hello, hello, hello!
Good morning!
How are you?
I’m fine!

As Wells (2006, p. 145) explains, in expressions like “How are you?” and “What is it?”, the tonic syllable goes on the *be* verb. This is not the case when the subject is contrasted with someone else. In this song, however, despite the lack of a contrastive context, the singer anomalously places emphasis on “you.”

She repeats this intonation pattern throughout the song. Such repetition, combined with the music, will naturally reinforce students’ memorisation of this pattern. However, this incorrect use of intonation patterns could hinder their learning process. It can be mitigated by modifying the musical arrangement. Fraser states that it might perhaps be used to express delight when addressing a person that one has not seen for a long time, for example, but it is not necessary for Japanese 5th graders to think of such a specific situation when they begin to learn English as a subject.

Indeed, songs have liberty to play with language in ways that may not align with the standard intonation patterns, but when they are used for language learning, especially for beginners like Japanese elementary school students, this can cause confusion and be educationally inappropriate. Examples of standard usage are crucial in the early stage of learning. Exposure to artistic variations is valuable, but it should wait until learners finish their foundational learning.

You’re welcome. (G5 CD2 No. 2)

This is another fundamental expression in our daily lives, and its basic intonation pattern, where the first syllable of “welcome” becomes the tonic syllable, must be carefully chosen for instructional purposes. Although this pattern is used in many instances in the recordings, there are also cases in which “You’re” is made prominent for no specific reason.

The same normal intonation pattern should be taught in situations where contrastive emphasis is not required. This principle is especially important for Japanese elementary school students so that they can establish a firm foundation of spoken English. Ignoring this may lead to confusion and frustration, contributing to the increasing number of those who dislike learning English.

Happy New Year! (G5 CD3 No. 1)

The G5 textbook presents two problematic renditions of “Happy New Year!” Both have the tonic syllable on “New,” with “Happy” read with a high head in one case and with a low head in the other. However, as Fraser noted, the default intonation pattern typically places the tonic syllable on “Year.” When examining various video clips on YouTube that include this expression, it becomes clear that people place the tonic syllable on “Year” more often. The only difference among these examples is found in the pitch level of “Happy”: Some use a high head, and others use a low head.

In addition, both “New Year’s Day” (G5 CD3 No. 3) and “the Eiffel Tower” (G6 CD1 No. 4) also have a fixed intonation pattern.⁷ In both phrases, the last word becomes the tonic syllable. In the case of “New Year’s Day,” however, two intonation patterns are used in the recordings without any apparent reasons. The tonic syllable is “New” in one case and “Day” in another. In the case of “the Eiffel Tower,” the first syllable of “Eiffel” is treated as the tonic syllable in a non-contrastive context, which is again incorrect. Textbook writers and narrators should take care to provide the correct intonation pattern for set phrases like these.

Since all the examples above are in non-contrastive contexts, it is challenging to logically explain why the standard intonation pattern is not observed in these instances. This likely results from simple errors by the narrators. Such errors should be avoided, as the default intonation pattern is a fundamental aspect of intonation.

Contrastive Contexts

If students are not correctly taught such basic knowledge of English intonation used in non-contrastive contexts, teaching the following contrastive intonation pattern becomes futile:

It was JOHN who gave it to me.

It was John who GAVE it to me.

This is cited from Cruttenden (1997, p. 87), who states, “Both sentences involve old information: in the first *gave it to me* is old information; in the second *John* is old information.” If “gave” in the first sentence is made prominent, it sounds strange to the listener. Likewise, if “John” in the second sentence becomes the tonic syllable, the listener perceives the same unnaturalness. Mastering this difference in tonic placement is important for effective real-life communication. It is essential for Japanese elementary school students to establish a solid foundation of the English language, including its intonation. The importance of presenting the correct use of tonicity is supported by the following examples from Dauer (1993, p. 230):

George is moving to Toronto next *month*⁸. (not some other time)

George is moving to Toronto *next* month. (not some other month)

George is moving to *Toronto* next month. (neutral; not some other city)

George is *moving* to Toronto next month. (not just going there)

George *is* moving to Toronto next month. (he really is)

George is moving to Toronto next month. (not someone else)

These examples clearly show the important use of tonicity in carrying the speaker's intended message correctly. However, as indicated in the results section, certain examples in the recordings demonstrate incorrect use of tonicity. These will be analysed in the following subsections.

I like chocolate. (G6 CD1 No. 1)

This utterance is used as a reply to the question “Do you like chocolate?” Unlike its default intonation pattern where the first syllable of “chocolate” becomes the tonic syllable, the subject “I” becomes the tonic syllable because both “like” and “chocolate” are old information. However, when the narrator read this sentence, he placed the tonic syllable on “like.” This placement is contextually incorrect. Figure 1 presents an acoustic analysis of Fraser’s utterance:

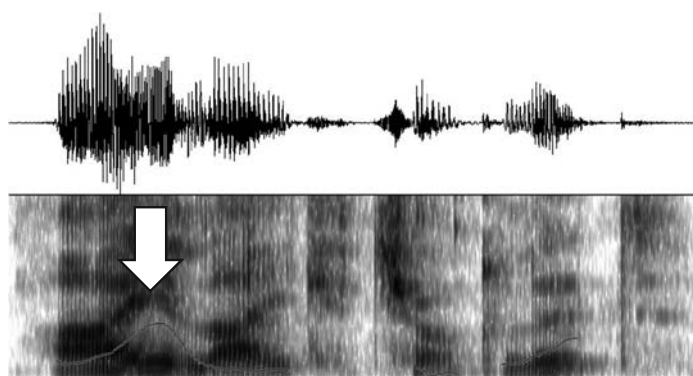


FIGURE 1. Acoustic Analysis of “I like chocolate.”⁹

The top layer displays the waveform, while the bottom layer presents both the spectrogram, depicted in shades of grey, and the F0, which is indicated by a line. The arrow corresponds to the F0 contour for “I.” He uses a fall-rise, which may be a personal choice to imply something afterwards. As Fraser mentioned above, when a reply is “I love chocolate,” “love” becomes the tonic syllable. Even though both verbs have the same meaning of fondness, their degree is much greater in “love.”

Textbook writers and narrators should choose proper intonation patterns carefully, paying attention to whether the default pattern should be used, and if not, which word should be made prominent. This correct placement of prominence is very important in carrying the speaker’s message correctly to his/her listener.

It’s round in western Japan. (G5 CD3 No. 2)

Another similar problematic case is found in this example. In a default, non-contrastive context, the tonic syllable is located on the second syllable of “Japan.” However, this sentence is used in the following conversation:

S: Do you like Japanese New Year’s food, Emily?

E: Yes, I like it very much.

G: Me, too. I especially like *zoni*. The *mochi* is soft and delicious.

S: Yes. You know? The *mochi* in eastern Japan is square. It’s round in western Japan.

G: Sounds interesting!

In this conversation, two contrasting elements – “square” vs. “round” and “eastern” vs. “western” – are present, but they are not appropriately reflected in the model reading, as illustrated in Figure 2 (a):

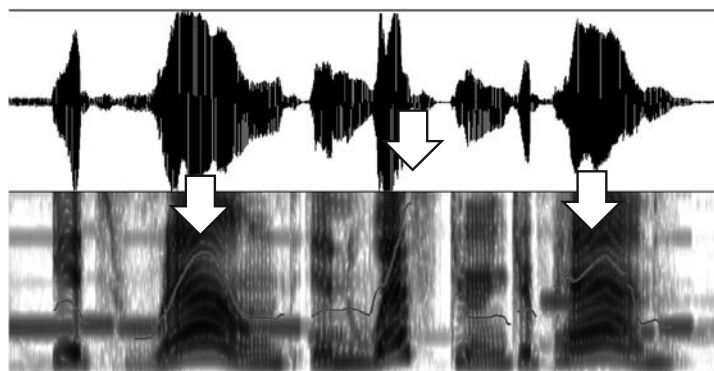


FIGURE 2 (a). Acoustic Analysis of “It’s round in western Japan.”¹⁰

As the arrows in the figure show, there are three noticeable F0 peaks. From the beginning, they correspond to “round,” the first syllable of “western,” and the second syllable of “Japan.” In fact, the peaks show that these first two tonic syllables are appropriate, but the third peak shows that the second syllable of “Japan” has incorrectly been made a tonic syllable. This sentence should be read as shown in Figure 2 (b):

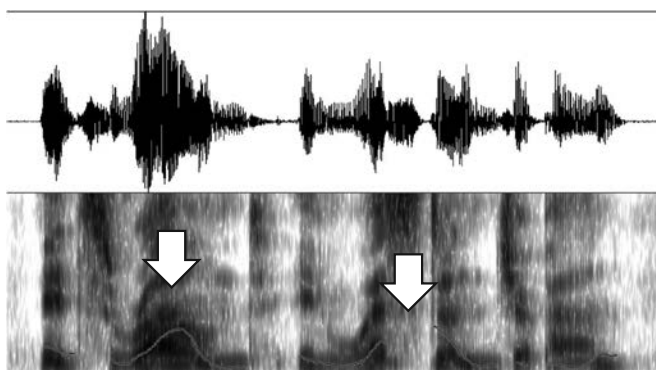


FIGURE 2 (b). Acoustic Analysis of “It’s round in western Japan.”¹¹

This is Fraser’s utterance. There are only two peaks, as shown by the arrows. From the beginning, they correspond to “round” and the first syllable of “western,” and show two contrasted items: “square” – “round” and “eastern” – “western.” “Round” is the onset of the head, and the first syllable of “western” is the tonic syllable. There is no noticeable F0 peak corresponding to “Japan,” except for the slight rise at the end. This may help to make the contrast more clearly between “eastern” and “western.”

Where do frogs live? What do frogs eat? (G6 CD2 Nos. 1-4)

These two utterances appear in the following dialogue:

H: OK. It's my turn again. Where do frogs live?
 E: Let's see ... Frogs live in the wetlands.
 H: That's right. And the second question, what do frogs eat?
 E: Well, frogs eat grasshoppers.
 H: Perfect! Good job!

Before this dialogue, the speakers talked about different animals: sea turtles, eagles, and lions. So, in “Where do frogs live?”, the fourth topic, “frogs,” should be highlighted. On the other hand, in “What do frogs eat?”, the topic “frogs” should be deaccented because it is already shared information between the two speakers. Instead, the contrastive word “eat” should be highlighted. In the reading, however, “frogs” is made prominent again. Such incorrect patterns are likely to confuse both students and teachers. Unknowingly, learners may gain incorrect knowledge about English intonation.

Where do sea turtles live?

In a non-contrastive environment, this utterance is spoken with the tonic syllable on the last lexical item “live.” This is the intonation pattern used outside the main text on page 42 in the G6 textbook. Fraser commented about this pattern, stating, “It seems odd to put the emphasis on ‘live’ in ‘Where do sea turtles live?’, unless it is a follow-up question.” As stated above, however, since this sentence is written outside the main text, it can be spoken in a default intonation pattern,¹² and this should be the main reason why this pattern is used here. Deciding which pattern to use in such textbook sections – whether to prioritise the default pattern or extract one from the main text – can be challenging.¹³

In the main text, this sentence is used in a successive series of questions in which two speakers talk about four animals (sea turtles, lions, eagles, and frogs) as a topic, focusing on their habitat (Where do ... live?) and food (What do ... eat?). When the basic framework of this talk is already understood as shared information with a conversational exchange such as “Let’s play a quiz game” – “OK,” then there is a good reason that each animal can be made prominent, including the first animal, where discursial information is prioritised over grammatical information. The location of the tonic syllable is influenced by “the speaker’s mental planning” (Wells, 2006, p. 114), as shown in Figure 3 (a):

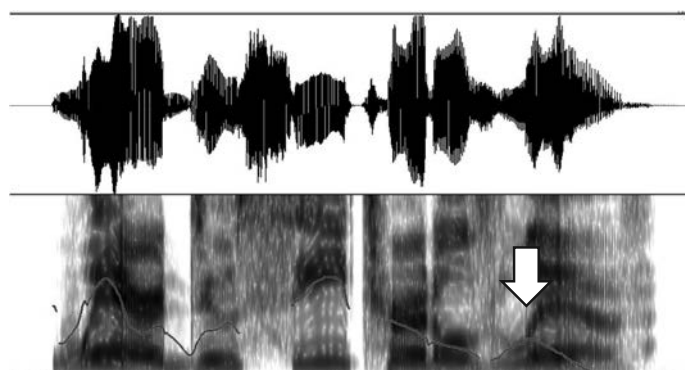


FIGURE 3 (a). Acoustic Analysis of “Where do sea turtles live?”¹⁴

This utterance is taken from a dialogue on page 42 of the G6 textbook. The arrow in the figure corresponds to the F0 contour for “live.” In contrast, Fraser placed much weaker prominence on this word, as shown in Figure 3 (b):

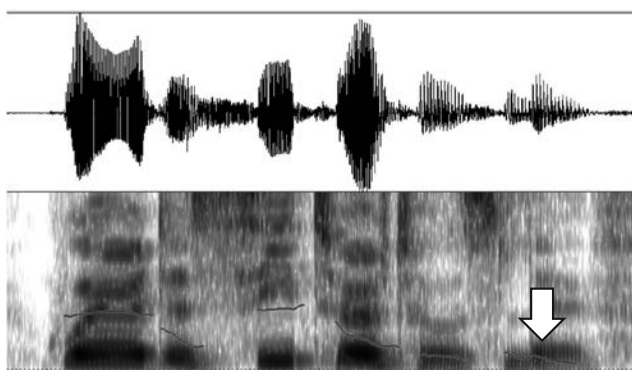


FIGURE 3 (b). Acoustic Analysis of “Where do sea turtles live?”¹⁵

This much lower F0 value on “live” clearly makes this word perceived as unaccented. The clear difference of this kind between an accented syllable and an unaccented syllable should be utilised in the narrators’ recordings.

The second subsection addressed the incorrect use of tonicity in contrastive contexts, as illustrated by examples presented in the results section. Although the recordings predominantly feature correct usage, instances of incorrect tonicity sporadically occur throughout. In a foreign language environment, the stimulus of spoken English is quite limited. Therefore, audio materials must be created with precision and care, adhering to the standard rules. Such deviations might be acceptable in students’ future learning, to expose them to the reality of human errors, but only after they have sufficiently and repeatedly listened to correct English intonation patterns and are able to discern between appropriate and inappropriate forms.

The following statement, cited from Jenkins (2000, p. 153), is noteworthy in demonstrating the significance of tonicity: “Nuclear stress is crucial for intelligibility in ILT¹⁶. Although the majority of the phonological errors causing intelligibility problems ... were segmental, a substantial minority involved intonational errors and, of these, almost all related to misplaced nuclear stress, particularly contrastive stress, either alone or in combination with a segmental error.” This statement emerges from her extensive research on factors that obstruct intelligibility in spoken interactions when English serves as a lingua franca. In addition, Jenkins (2007, pp. 23–24) summarised core features and non-core features in two tables. Naturally, the objectives for English as a lingua franca should be less challenging than those in traditional English education, which focuses on L1 English speakers. However, tonicity is identified as a critical core feature, whereas stress-timed rhythm is deemed unnecessary. Consequently, overlooking the incorrect use of tonicity in audio recordings for Japanese elementary school students is not an option.

Unnecessary Exaggeration

A recurring problem in the audio materials is unnecessary exaggeration. Six examples will be discussed

to illustrate this.

What's this? (G5 CD1 No. 1)

An acoustic analysis was conducted to evaluate Fraser's observation that this utterance sounds unnecessarily exaggerated, the result of which is displayed in Figure 4 (a):

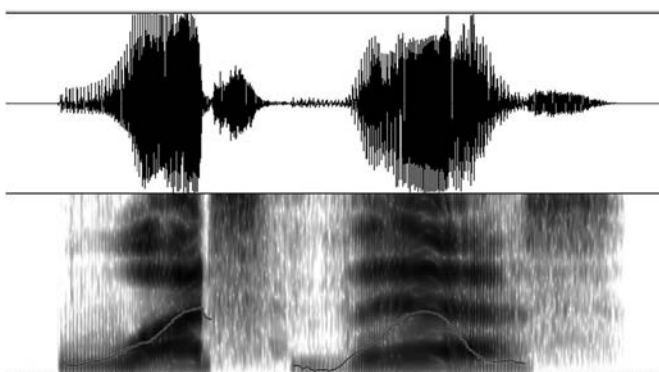


FIGURE 4 (a). Acoustic Analysis of “What’s this?”¹⁷

It is understood in this figure that a rising head is used for “What’s,” and that the tonic syllable “this” is spoken with a rise-fall, with its initial F0 value being low. The difference between this initial value and the peak value is 16.5 semitones (ST). This use of the F0 contour for both the head and the tonic syllable makes this utterance sound emphatic. On the other hand, Fraser’s utterance is different, as is shown in the acoustic analysis displayed in Figure 4 (b):

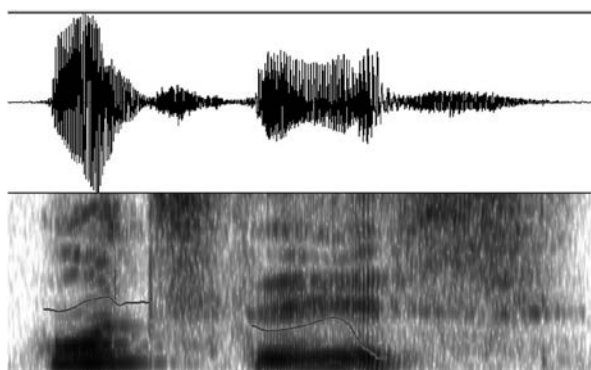


FIGURE 4 (b). Acoustic Analysis of “What’s this?”¹⁸

The major difference is found in the use of the F0 for both the head and the tonic syllable. The head is spoken with a high F0 value, with no major change. The tonic syllable is simply spoken with a fall. There is no noticeable initial low F0 value. The difference between the initial value and the peak value is only 2.7 ST.

This utterance is used simply to ask the name or feature of something that the speaker and the listener see. There is no reason to make this utterance sound exaggerated unless the speaker is irritated for some reason. The use of both the rising head and the rise-fall requires more time than the standard reading that Fraser shows. It is assumed from this that such an unnecessarily exaggerated intonation is a by-product of a slow speech rate.

What's your name? (G5 CD1 No. 3)

Regarding two utterances of this sentence, Fraser noted that one sounds odd due to unnecessary exaggeration, characterised by the use of a rise-fall, but that no such oddity is noticed in another. Figure 5 (a) and (b) show an acoustic analysis of these two utterances. The former is the utterance which sounds odd:

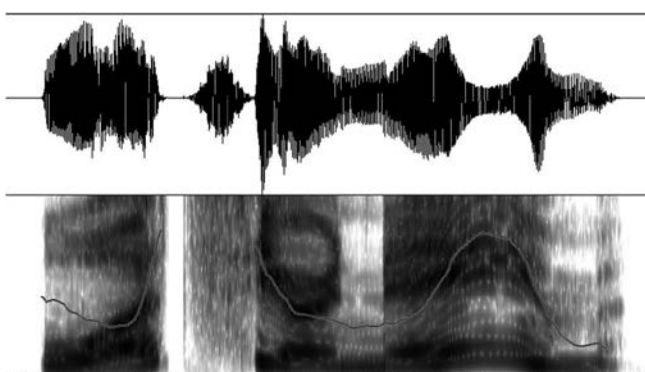


FIGURE 5 (a). Acoustic Analysis of “What’s your name?”¹⁹

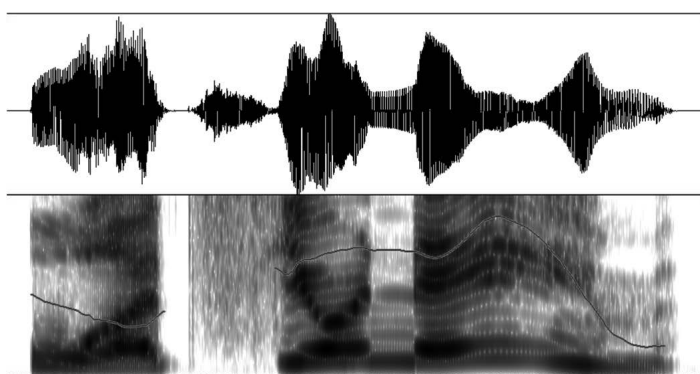


FIGURE 5 (b). Acoustic Analysis of “What’s your name?”²⁰

There are three noticeable differences in the F0 contour for “name” between (a) and (b). First, the initial value is low in (a) (190.3 Hz), but not in (b) (371.6 Hz). Second, a big jump is observed in (a) (13.3 ST), but not in (b) (3.3 ST). Third, the peak is located later in time in (a) than in (b): within the FACE vowel, it is in the 57.7% position in (a) and in the 45.6% position in (b). These three differences made the utterance analysed in Figure 5 (b) sound unnatural to Fraser. The initial low F0 value and the delayed peak are well-

known features of a rise-fall. As Roach (2009, p. 125) states, this tone is used for “strong feelings” including surprise. The use of a rise-fall in asking a person’s name is not appropriate in a normal conversational exchange, but this tone is applied to this utterance. The probability of the relation between the use of this tone and the slow speech rate may not be denied here, too. As stated above, a consistent intonation pattern should be used for the same utterance in the same social environment so as not to confuse learners.

Fraser’s reading of this sentence is shown in the acoustic analysis in Figure 5 (c):

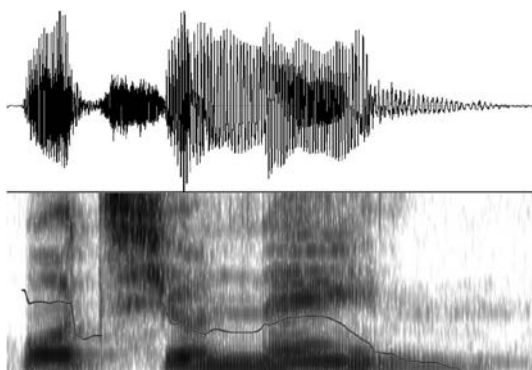


FIGURE 5 (c). Acoustic Analysis of “What’s your name?”²¹

As can be clearly observed, it is spoken with a fall. The initial low value for this word is 127.2 Hz. This is much lower than the value in (b), reflecting a physiological difference. The peak is located at the 46.4% position within the FACE vowel, which is almost the same position in (b). The vertical difference in the F0 value between the initial low and the peak is 2.6 ST in (c). This is much smaller than in (b). This difference is caused by an attitudinal difference applied in each recording. The high emotional level is applied to (b), as well as (a). Although (b) is more natural than (a), it may still not be good enough when it is compared with (c).

How do you spell your name? (G5 CD1 No. 5)

There are four utterances that Fraser commented on regarding this sentence. The initial low F0 value to the peak value in “name” is as follows: 5.3 ST, 10.2 ST, 11.3 ST, and 11.5 ST. The most serious comment he made was on the utterance whose F0 difference is the smallest. An acoustic analysis of this utterance is shown in Figure 6 (a).

Even though the difference in the F0 value between the initial low and the peak is only 5.3 ST, Fraser perceived this utterance as spoken with a rise-fall. On the other hand, there is an utterance that he judged as reasonable, and its acoustic analysis is displayed in Figure 6 (b).

In this utterance, the difference in the F0 value between the initial low and the peak is 11.3 ST. It is understood from these two figures that it is not the difference in the F0 contour between the low value and the peak value that makes these utterances sound different. The key issue is whether the F0 contour continues to ascend toward the peak for “name” or not. In Figure 6 (b), such a continuation is detected, but in Figure 6 (a), there is a sudden drop of the F0 contour at the beginning of “name” as indicated by an arrow. The same

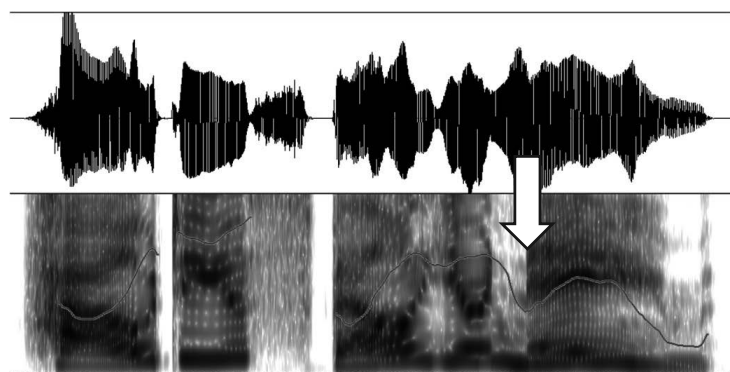


FIGURE 6 (a). Acoustic Analysis of “How do you spell your name?”²²

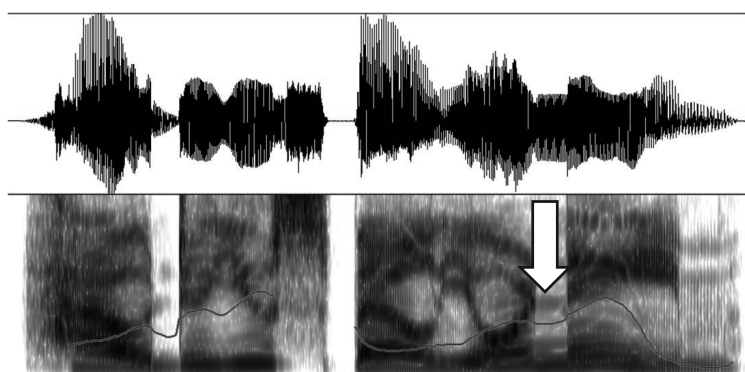


FIGURE 6 (b). Acoustic Analysis of “How do you spell your name?”²³

use of an arrow is also applied to Figure 6 (b), but no such drop is detected there. It is this sudden drop of the F0 contour that makes Fraser perceive this utterance as spoken with a rise-fall.

Since “spell your name” is a semantically cohesive unit as a verb phrase, a disrupted pitch movement is something not expected. Even such a slight change of pitch is perceived as linguistically significant, which may give an unintended impression to the listener. This also highlights the important role of the standard intonation, especially for those who have not finished a foundational stage of learning.

Who’s this? Who is Mark Smith? (G5 CD2 No. 1)

An acoustic analysis of this set of two utterances is shown in Figure 7 (a), and to make a comparison with this, an acoustic analysis of the same set spoken by Fraser is shown in Figure 7 (b).

Judging from the characteristics of the F0 contour in these two figures, it is conceivable that the main reason why these two utterances, acoustically analysed in Figure 7 (a), sound inquisitorial is the use of a rise-fall. When a rise-fall is used with a WH question, it can carry the meaning of “challenging and antagonistic” (O’Connor & Arnold 1973, p. 214). This use of a rise-fall may give an unnecessarily strong impression to the listener, and as a result, it does not sound like a normal intonation pattern in the context in which it is used.

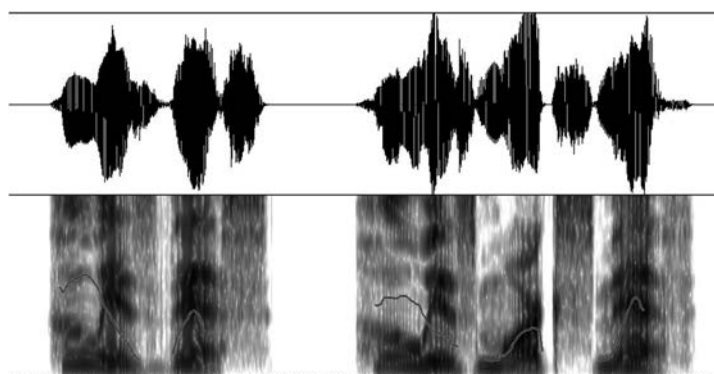


FIGURE 7 (a). Acoustic Analysis of “Who is this?” and “Who is Mark Smith?”²⁴

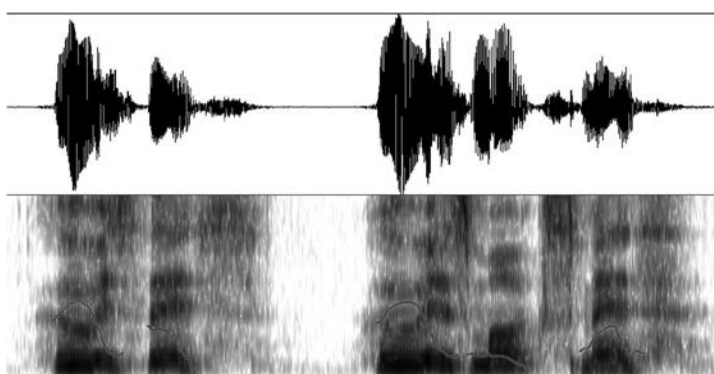


FIGURE 7 (b). Acoustic Analysis of “Who is this?” and “Who is Mark Smith?”²⁵

The duration of these four utterances was measured, yielding the following results:

TABLE 8. Comparison of Utterance Duration

	Textbook narrator	Fraser	Difference rate
Who is this?	1,301	867	1.5
Who is Mark Smith?	2,018	1,238	1.6

It is understood from Table 8 that there is a big difference in the speech rate between the two speakers. The narrator spoke 1.5 to 1.6 times slower than Fraser. Although Fraser tried to speak these utterances as slowly as possible while maintaining naturalness, he could not speak as slowly as the narrator. This example also shows the probability of a close relationship between unnaturalness and the slow speech rate. There may be a possibility that the narrator himself may not have been accustomed to this slow speech rate, but there should be a limit in the use of such a rate so that naturalness can be maintained.

Potatoes and some spices are inside. (G5 CD2 No. 3)

Figure 8 (a) shows an acoustic analysis of a textbook narrator’s utterance “Potatoes and some spices

are inside,” while Figure 8 (b) shows an analysis of Fraser’s utterance:

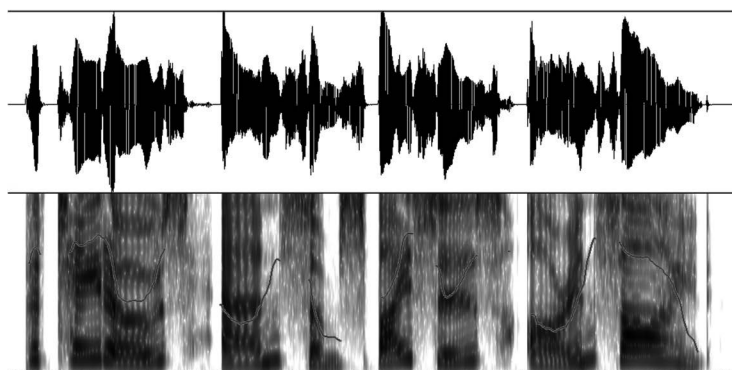


FIGURE 8 (a). Acoustic Analysis of “Potatoes and some spices are inside.”²⁶

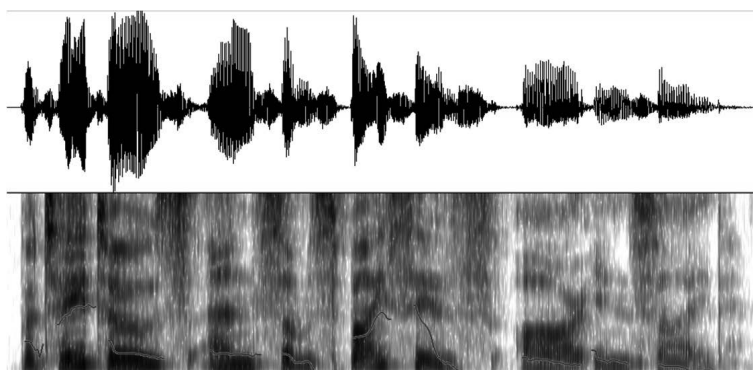


FIGURE 8 (b). Acoustic Analysis of “Potatoes and some spices are inside.”²⁷

There are three noticeable features of the F0 contour in Figure 8 (a): the use of a fall-rise for “potatoes” and “spices,” the use of a high fall for “inside,” and the overall use of a high pitch. It is understood from these tones that this utterance consists of three tone-units. On the other hand, in Figure 8 (b), these features of the F0 contour are not detected. There is no use of a fall-rise, and a fall is used for “potatoes” and “spices.” There is no accent on “inside” because Fraser thought it was predictable from the context. The interpretation of this word is a noticeable difference between the two utterances. There are two tone-units in this utterance. It has one fewer tone-unit than the one above. There is not much difference in duration: 3,973 ms in (a) and 3,089 ms in (b), but as with other examples (a) is longer than (b).

A similar pattern is observed in “My birthday is May 5th,” where the narrator uses an extremely high pitch for “5th.” In addition, there is a difference in duration between this narrator and Fraser. The narrator spoke 1.5 times as slowly as Fraser, taking 3,130 ms to utter this sentence, while Fraser took only 2,004 ms.

Considering the examples of unnecessary exaggeration presented here, it is reasonable to present models that create a lively atmosphere where appropriate. However, as the examples analysed in this section clearly show, the use of excessively high pitch in these audio materials, coupled with shorter tone-units and

an increased number of tones, raises questions. They end up sounding overly dramatic, which seemingly undervalues naturalness. This was Fraser's key concern. Typically, L1 English speakers do not adopt such an enthusiastic tone in the contexts portrayed in these textbooks. The rationale behind the textbook authors' choice of this intonation pattern remains unclear.

When the speaker's unnecessary exaggeration causes the listener to perceive an unintended oddity, it may not critically harm the interpersonal relationship between the interlocutors. However, sounding overly dramatic will certainly surprise or unsettle the listener. While these two unintended impressions may not pose a serious problem in communication, sounding inquisitorial when the speaker simply wants to know someone's identity in a normal social context is a different matter. It must surely be detrimental to the interlocutors, undermining the Cooperative Principle that Grice (1975, p. 45) advocates for effective conversational communication.

As can be understood from Pike's (1945) claim that "In English ... an INTONATION MEANING modifies the lexical meaning of a sentence by adding to it the SPEAKER'S ATTITUDE toward the contents of that sentence" (p. 21) and that "in actual speech, the hearer is frequently more interested in the speaker's attitude than in his words" (p. 22), the incorrect use of intonation should be avoided. In this sense, there is an issue with making use of unnecessary exaggeration in audio materials for Japanese learners when they learn English in elementary school. The most educationally desirable approach is to provide these students with as much natural spoken English as possible as a model.

Language teaching methodologies vary greatly and opinions on the effectiveness of exaggerated speech in teaching materials can differ. Some educators may see value in such speech for foundational learning, but there needs to be a critical limit. The authors believe that naturalness and authenticity should be the key concepts in English teaching. However, at the same time, a desirable and practical balance should be sought through constructive discussions among educators who hold different beliefs concerning language teaching methodologies. Such discussions are vital in contributing to a new phase of English teaching methods that will be implemented in Japanese elementary schools in the not-so-distant future.

CONCLUSION

This study examined the quality of audio recordings in English textbooks for Japanese elementary school students, utilising an impressionistic approach by an L1 English speaker with extensive experience teaching the language at a Japanese university. This approach is essential for evaluating the naturalness of these materials, which serve as crucial models of spoken English in a country where the language is not commonly used in daily communication. The introduction of English as an official subject in Japanese elementary schools in 2020 was intended to improve oral English skills among students. However, the findings of this paper reveal significant shortcomings in the audio materials examined here.

The study identified three primary issues with the audio materials: (1) an excessively slow speech rate resulting in unnatural pauses and intonation; (2) incorrect tonicity, which hinders the learning of correct intonation patterns and the accurate conveyance of the speaker's message; and (3) unnecessary exaggeration in overly animated and dramatised readings, which are not suitable for the target age group of 5th and 6th graders.

To solve these problems, textbook writers and narrators must ensure careful planning, recording, and

monitoring of the audio content, while MEXT inspectors should rigorously evaluate audio materials alongside written texts and establish clear recording guidelines. An interesting insight from a recent conference attended by the first author of this paper highlighted the rushed nature of recording sessions, indicating a systemic problem in the production process. To enhance the quality of oral English education in Japanese elementary schools, the process of materials creation, particularly in relation to audio, needs to be reviewed and tightened.

While research often focuses on learners' pronunciation, this study suggests that equal importance should be placed on the quality of model recordings. Thus, improving the quality of audio materials in English education is essential for fostering better language skills among Japanese students. In addition, future research could benefit from comparing methods used in other countries where English is learned as a foreign language. Such research will potentially offer valuable insights for improving Japan's approach to English education and also inform global best practices in foreign language education.

NOTES

- 1) This research is supported by the Japan Society for the Promotion of Scientific Grants-in-Aid for Scientific Research (C) 21K00672. We would like to thank the reviewers of this paper for their constructive and helpful comments, which have significantly improved the quality and clarity of the manuscript.
- 2) Words in the parentheses show the textbook name each company adopts.
- 3) According to Kaede Production (2023), the top three companies in terms of the adoption rate are Tokyo Shoseki (accounting for more than 50%), Mitsumura (accounting for 15%), and Kairyudo (no adoption rate given).
- 4) Unlike conversations, monologues will provide a more exact value of a speech rate because narrators can maintain their speech rhythm while reading without any interruptions from others, including turn-takings.
- 5) In the monologue of Unit 8, when the speaker mentions that she likes dogs, two dog barks are recorded. They were deleted when the duration of this monologue was measured.
- 6) The monologue in Unit 1 corresponds to the middle stage of the two-year English course in Japanese elementary school, and the one in Unit 8, which is the last unit of the G6 textbook, corresponds to the final stage. It is not certain, however, whether the increase of the speech rate between the two units was intended by the textbook writers.
- 7) In "X Tower," the first syllable of "Tower" becomes the tonic syllable, including "Tokyo Tower."
- 8) The original text employs bold font, but this paper adopts italic font in alignment with Cruttenden (1997).
- 9) The top scale of the F0 is 500 Hz.
- 10) The top scale of the F0 is 600 Hz.
- 11) The top scale of the F0 is 500 Hz.
- 12) Another example which uses this default pattern is found in the utterance "One of the big decisions I think every organisation makes is where do policies live." This is taken from a video clip on YouTube Microsoft Ignite (2017).
- 13) Similar examples are found in "I usually watch soccer games on Sundays" (G6 CD1 No. 3) and "I enjoyed talking with my friends." (G6 CD3 No. 2)
- 14) The top scale of the F0 is 500 Hz.

- 15) The top scale of the F0 is 500 Hz.
- 16) ILT means Interlanguage Talk.
- 17) The top scale of the F0 is 500 Hz.
- 18) The top scale of the F0 is 500 Hz.
- 19) The top scale of the F0 is 500 Hz.
- 20) The top scale of the F0 is 500 Hz.
- 21) The top scale of the F0 is 300 Hz.
- 22) The top scale of the F0 is 500 Hz.
- 23) The top scale of the F0 is 500 Hz.
- 24) The top scale of the F0 is 500 Hz.
- 25) The top scale of the F0 is 500 Hz.
- 26) The top scale of the F0 is 500 Hz.
- 27) The top scale of the F0 is 500 Hz.

REFERENCES

- Cruttenden, A. (1997). *Intonation* (2nd ed.). Cambridge University Press.
- Dauer, R. M. (1993). *Accurate English: A complete course in pronunciation*. Prentice Hall Regents.
- Field, J. (2008). *Listening in the language classroom*. Cambridge University Press.
- Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, 3: Speech acts* (pp. 41–58). Academic Press.
- Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.
- Jenkins, J. (2007). *English as a lingua franca: Attitude and identity*. Oxford University Press.
- Kaede Production. (2023). *Top three elementary school textbooks adopted in 2021 (Reiwa3nendo shogakko kyoukasho saitaku besuto 3)*. Retrieved November 15, 2023, from <https://kaede-pro.com/blog/2023/02/10/post-421/>
- Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.
- MEXT. (2017). *Commentary on the Course of Study for Elementary Schools: Foreign Language Activities and Foreign Language (Shogakko Gakushu Shido Yoryo Kaisetsu: Gaikoikugo Katsudo Gaikokugohen)*. Retrieved November 15, 2023, from https://www.mext.go.jp/component/a_menu/education/micro_detail/_icsFiles/afieldfile/2019/03/18/1387017_011.pdf
- Microsoft Ignite. (2017). *Build a modern intranet: Real-world planning, information architecture, governance and adoption* [Video]. YouTube. <https://www.youtube.com/watch?v=dosFfvee6xY&t=1402s>
- O'Connor, J. D., & Arnold, G. F. (1973). *Intonation of colloquial English: A practical handbook* (2nd ed.). Longman.
- Pike, K. L. (1945). *The intonation of American English*. University of Michigan Press.
- Roach, P. (2009). *English phonetics and phonology: A practice course* (4th ed.). Cambridge University Press.
- Virtual Speech. (2022). *Average speaking rate and words per minute*. Retrieved November 15, 2023, from <https://virtualspeech.com/blog/average-speaking-rate-words-per-minute>

- Wells, J. C. (2006). *English intonation: An introduction*. Cambridge University Press.
- Yuzawa, N. (2022a). An analysis of two English textbooks for elementary school in Japan: Focusing on teaching pronunciation. *Journal of the School of International Studies, Utsunomiya University*, 53, 103–116.
- Yuzawa, N. (2022b). An analysis of the intonation patterns in audio materials attached to English textbooks for 5th Graders in Japan. *Journal of the School of International Studies, Utsunomiya University*, 54, 105–114.
- Yuzawa, N. (2022c). A study of intonation unit in the audio materials for three English textbooks aimed at Japanese 6th Graders. *Journal of the School of International Studies, Utsunomiya University*, 54, 115–123.

ABSTRACT

Problems with Audio Recordings for Elementary School English Textbooks in Japan

Nobuo YUZAWA

Faculty of International Studies

Utsunomiya University

Simon FRASER

Institute for Foreign Language Research and Education

Hiroshima University

This study discusses issues with audio recordings attached to English textbooks for Japanese elementary school students. Out of seven sets of textbooks, approved by the MEXT (the Ministry of Education, Culture, Sports, Science and Technology), the one published by the Tokyo Shoseki Publishing Company, *New Horizon Elementary English Course 5* and *6*, was selected because these textbooks are the most extensively used in Japan and their impact on English teaching in Japan is considered high. Four CDs accompany the textbook for 5th graders, and three accompany the textbook for 6th graders. In the study, an L1 English speaker identified unnatural-sounding utterances, which were then acoustically analysed to understand why they sound unnatural.

Three problems with the recordings were found. Firstly, the speech rate is excessively slow, leading to unnatural pauses. Secondly, errors in tonicity are identified. Thirdly, the reading style is overly animated and dramatised, which is inappropriate for 5th and 6th graders.

The present study emphasises the critical role of audio materials for English learners, especially beginners. Recommendations include careful preparation by textbook writers and narrators, attentive monitoring during recording, and increased involvement of MEXT textbook inspectors to ensure recording quality. In conclusion, the authors suggest that the process of materials production and review be tightened and that comparative analyses with other countries be taken to improve the appropriateness of model recordings.

要 旨

日本の小学校英語教科書付属の音声教材に関する諸問題

湯 澤 伸 夫

宇都宮大学国際学部

サイモン・フレイザー

広島大学外国語教育研究センター

本研究では日本の小学校英語教科書に付属されている音声教材に関する諸問題を考える。7種類の文部科学省検定済教科書のうち、日本で最も多く使用され、日本の英語教育への影響が大きいと考えられる点から、東京書籍が出版する『New Horizon Elementary English Course 5・6』を使用した。付属のCDは、5年生用の教科書には4枚、6年生用には3枚用意されている。本研究では、まず、英語母語話者に不自然に聞こえる発話を特定してもらい、次に、その発話がなぜ不自然に聞こえるのかを音響的に分析した。

結果として3つの問題点が見つかった。第1に、発話速度が非常に遅く、不自然なポーズが生じている点である。第2に、音調核の位置に誤りがある点である。第3に、過度にドラマチックな読み方は、小学5・6年生には不適當である点である。

英語学習者、特に初級者に対し、音声教材は重要な役割を果たしている。現状の改善には、教科書執筆者とナレーターが入念に準備を行うこと、録音時には注意深くモニタリングすること、録音の品質管理のために文部科学省教科書調査官の関与を強化することが考えられる。制度全体の見直しが必要であり、モデルの妥当性の向上を目的とした諸外国との比較分析も重要である。