L2 Perception and Production of Japanese Lexical Pitch: A Suprasegmental Similarity Account

Tim Joris Laméris[1] and Calbert Graham
University of Cambridge

Adults are known to have difficulties acquiring suprasegmental speech that involves pitch (f_0) in a second language (L2) (Graham & Post, 2018; Hirata, 2015; Wang, Spence, Jongman, & Sereno, 1999; Wong & Perrachione, 2007). Previous research has suggested that the perceived similarity between L1 and L2 phonology may influence how easily segmental speech is acquired, but this notion of 'similarity' may also apply to suprasegmental speech (So & Best, 2010; Wu, Munro, & Wang, 2014). In this paper, the L2 acquisition of Japanese lexical pitch was assessed under a 'Suprasegmental Similarity Account', which is a theoretical framework inspired by previous models of segmental and suprasegmental speech (Best & Tyler, 2007; Flege, 1995; Mennen, 2015) to account for the L2 acquisition of word prosody. Eight adult native speakers of Japanese and eight adult English-native advanced learners of Japanese participated in a perception and production study of Japanese lexical pitch patterns. Both groups performed similarly in perception, but non-native speakers performed significantly worse in production, particularly for 'unaccented' Low-High-High patterns. These findings are discussed in light of the 'Suprasegmental Similarity Account'.

Keywords: JAPANESE, LEXICAL PITCH, PERCEPTION, PRODUCTION

[1] Email (corresponding author): tjl44@cam.ac.uk

BACKGROUND

L2 Acquisition of Word Prosody and Japanese Lexical Pitch

When acquiring a second language (L2), adults are faced with the challenge of not only acquiring the segmental aspects of speech, i.e. vowels and consonants, but also the suprasegmental aspects at the word level, such as lexical stress, and, in many languages, lexical tone. Suprasegmental speech at the word level, also known as word prosody, is known to be a difficult aspect of L2 speech acquisition, as has been reported in studies investigating L2 acquisition of lexical stress (Dupoux, Sebastián-Gallés, Navarrete, & Peperkamp, 2008; Ortega-Llebaria, Gu, & Fan, 2013, among others) and L2 acquisition of lexical tone (a good overview can be found in Antoniou & Chin (2018)).

The similarity between L1 and L2 phonology is often thought to affect how easily non-native sounds are perceived, as has been shown in various studies on segmental acquisition (MacKain, Best, & Strange, 1981; Pallier, Bosch, & Sebastián-Gallés, 1997). An influential theory that supports this notion is the Perceptual Assimilation Model (PAM) (Best, 1995; Best & Tyler, 2007). In a nutshell, the PAM proposes that learners are good at perceptually discriminating L2 phonological categories when those categories map, or ‘assimilate’, readily onto separate L1 categories in a one-to-one fashion (Two-Category Assimilation). Learners have more difficulty telling L2 sounds apart when there is a phonological mismatch and L2 categories map onto L1 categories in a many-to-one (Single-Category Assimilation) or a one-to-many fashion (Category Goodness Assimilation). Although the PAM’s predictions seem to fit well with many empirical observations in segmental speech acquisition, which is indeed its focus, whether categorical assimilation also occurs in suprasegmental speech is a topic that has been less studied. However, in recent years, the PAM is increasingly being applied to suprasegmental speech acquisition, in particular lexical tone, in which, according to some studies, similar categorical assimilation effects can be observed (see: Braun, Galts, & Kabak (2014); Hao (2012); So & Best (2010); Wu, Munro, & Wang (2014), among others).

Despite its widespread applications, the PAM only provides a limited account of acquisition because it focuses solely on perception, i.e. listening abilities. An alternative model that examines the learnability of both the perception and the production of L2 speech is the Speech Learning Model (SLM) which is concerned with 'the ultimate attainment of L2 pronunciation' (Flege, 1995, p. 238). The SLM is similar to the PAM in that it assumes that acquisition is primarily affected by categorical assimilation in perception, although it posits that incompatible L2 sounds may be stored as new 'phonetic categories'. Being a model of L2 production, the SLM further postulates that production may be subject to 'motoric output constraints' (p. 238). Unlike PAM however, the SLM appears to have been applied only sparsely to the acquisition of suprasegmental speech, with one exception being a study on acquisition of lexical tone by Y. Hao (2014).

Perhaps one of the most suitable recent models of suprasegmental speech acquisition is the 'L2 Intonation Learning theory' (LILt) (Mennen, 2015). An important property of this model is that it recognises four dimensions (systemic, realisational, semantic, and frequency) that are relevant in acquisition of intonation. Any L1-L2 discrepancy along these four dimensions is thought to affect the relative difficulty of acquiring intonation in an L2. For instance, acquisition may be hindered when there are: (1) phonologically incongruent L1 and L2 intonational categories (the 'systemic dimension'), (2) different alignments of pitch targets (the 'realizational dimension'), (3) different functions and meanings associated to pitch curves (the 'semantic dimension'), or (4) different frequential distributions of intonational categories (the 'frequency dimension'). Although designed as a model for intonation, which is phrasal suprasegmental speech, the LILt framework may also be applied to lexical suprasegmental speech. For instance, the notion of the 'semantic dimension' resonates with earlier reported problems in 'levels of representation' in the acquisition of lexical tone (Francis, Ciocca, Ma, & Fenn, 2008, p. 269). That is, speakers of non-tone languages may have more difficulty acquiring lexical tones because they do not strongly associate pitch patterns with lexical meaning. Similarly, in studies on L2 lexical stress acquisition, it has been argued that speakers of non-stress languages struggle with acquiring lexical stress not because they are psychoacoustically 'deaf' to stress, but because they do not associate stress to lexical contrastivity (Dupoux et al., 2008, p. 701).

As described above, there appears to be no single theoretical model that directly applies to the acquisition of lexical suprasegmental speech. In this paper, which focuses on the acquisition of Japanese lexical pitch, we will therefore draw upon the commonalities of these three models and summarise these in what we will call a 'Suprasegmental Similarity Account' of lexical suprasegmental speech. The Suprasegmental Similarity Account postulates a theoretical framework that describes the acquisition of suprasegmental speech at the word level, since similar theoretical models like the LILt have only focused at suprasegmental speech at the phrasal level. Our account primarily assumes the assimilation of L2 (suprasegmental) categories in terms of L1 (suprasegmental) categories in perception, as inspired by the PAM model, but we also acknowledge that L2 production may be subject to motoric output constraints, as highlighted by the SLM and by the 'realisational dimension' of the LILt. Finally, our account will also consider effects of different sound-meaning associations that learners may have with certain suprasegmental features, in line with the LILt's 'semantic' dimension, as well as effects of 'frequency' of those suprasegmental features. As we aim to describe the acquisition of Japanese lexical pitch in a linguistic framework, our account will, for simplicity purposes, not consider any speaker-specific extralinguistic factors that may have an influence on acquisition of suprasegmental speech, such as musical experience (Antoniou & Chin, 2018).

The reason why Japanese provides an interesting case for a Suprasegmental Similarity Account of speech acquisition is because its word prosody has features which are similar to both lexical tone and lexical stress (Gussenhoven, 2004, pp. 26–47; Yip, 2002, p. 2). On the one hand, Japanese appears to have features of a tone language: words carry lexically predefined pitch patterns of High (H) and Low (L) pitches which are individually assigned to each mora. These pitch patterns are an inherent property of the Japanese word and can be crucial in differentiating meaning between segmental homonyms. For instance, the word *bùdou*, when pronounced with a High-Low-Low H*LL pattern[2], means ‘martial art’, whereas its segmental twin *budou* pronounced with a Low-High-High LHH pattern means ‘grape’, as shown in Figure 1.

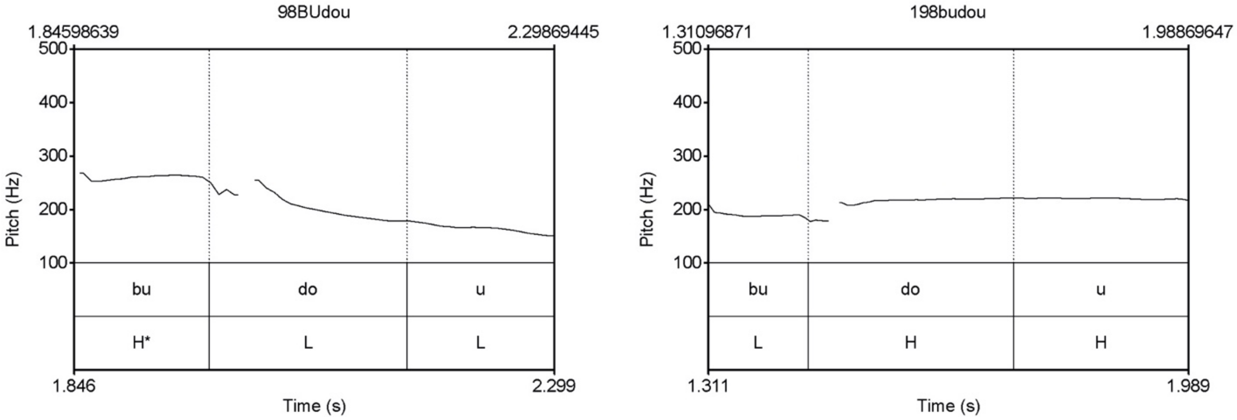

Figure 1. f_0 contours for H*LL bùdou 'martial art' and LHH budou 'grape'

[2] Lexical pitch patterns will be described in an adapted form of the style by Kawahara (2015), where each mora is described by a relatively Low (L) or High (H) pitch value. An asterisk on a High pitch (H*) indicates a lexically predefined pitch accent and a period (.) a word boundary. In a written word, like *bùdou,* a grave accent indicates a pitch accent.

At the same time, Japanese lexical pitch patterns have properties similar to lexical stress. This is because pitch patterns can have a pitch accent, which is a lexically predefined fall from H* in one mora (the accented mora) to L in the subsequent mora. This lexically predefined fall in f_0 is the primary acoustic cue for prominence in a Japanese word (Beckman, 1986; Kawahara, 2015) and can be compared to lexical stress in English as it marks prominence and can make lexical distinctions (e.g. English obJECT (verb) and OBject (noun)). An important difference between Japanese pitch accents and English lexical stress however, is that f_0 is the primary cue for Japanese lexical pitch accent, whereas English lexical stress is formed of not only f_0, but also of other acoustic elements such as duration, intensity, and vowel quality (Pierrehumbert & Beckman, 1988, p. 270; Shport, 2011, p. 11). Another important difference is that the Japanese pitch accent may be optional, whereas English lexical stress is required in each word (Kawahara, 2015, p. 2; Sugiyama, 2012, pp. 1–9). Japanese therefore can have ‘unaccented’ words, which are words that start with a Low pitch followed by a continuous High pitch until the end of the word. These words lack a pitch accent and, as a result, a clear sign of prominence. The earlier mentioned word *budou* ‘grape’, with a LHH pitch configuration, is an example of such an unaccented word. In fact, most of native Japanese words and Sino-Japanese words are unaccented (Kubozono, 2012). The current study investigated the acquisition of four trimoraic Japanese pitch pattern categories, with a pitch accent located on the first, the second, the third, or on none of the morae. In terms of relative pitch values, these patterns are described as: [1] H*LL, [2] LH*L, [3] LHH*, and [0] LHH. It is important to note that realisation of pattern [3] depends on whether it is uttered in isolation or in a sentence context. In isolation, the pitch fall is prescribed but not realised, because there is no space within the mora for the fall in f_0 to take place (this will be indicated by ‘°’). As a result, LHH° pattern [3] in *isolation* is audibly the same as LHH pattern [0] (Ota, 2015, p. 695; Sakuma, 1929; Sugiyama, 2012, pp. 23–28). The prescribed fall in f_0 in pattern [3] only occurs and only becomes audible in a sentence context, in which a mora from a subsequent word can accommodate the pitch fall. Therefore, only in sentence contexts will we refer to the final-mora accented pattern [3] as LHH*.

It has been suggested in previous studies that second-language learners of Japanese struggle with lexical pitch (Ayusawa et al., 1995; Hirano-Cook, 2011; Hirata, 2015; Iimori, 2014; Lanz, 2003; Sakamoto, 2011; Shport, 2016). Despite this observation, why exactly L2 learners of Japanese struggle so much with learning lexical pitch is relatively unclear. Moreover, whether L2 learners of Japanese rely on L1 suprasegmental categories when attempting to acquire lexical pitch, in line with a Suprasegmental Similarity Account of speech learning, is a topic that is rarely addressed, with a few exceptions: T. Ayusawa et al. (1995) suggest for instance that French learners' performance in identifying HLLLL patterns was relatively bad because these patterns have no clear correlate in French intonation, indirectly hinting that French listeners may have assimilated Japanese lexical pitch to French intonation. Sakamoto (2011) mentions that English learners of Japanese may benefit from 'cross-language phonetic similarity between L1 and L2' (p. 265), highlighting that listeners may perceive L2 sounds in a more phonetic rather than phonological way. One of the most in-depth interpretations in terms of a Suprasegmental Similarity Account is a study by Shport (2016), who argues that English learners of Japanese often confused unaccented LH with final-mora accented LH* patterns because of their similarity to English second-syllable stress patterns, which constitutes a 'Category Goodness Assimilation' in PAM terms (Best & Tyler, 2007).

MOTIVATION FOR CURRENT STUDY

The primary aim of this study is to provide new insights into the interactions between L1 and L2 suprasegmental phonology in the acquisition of Japanese lexical pitch, and how an overall degree of similarity between English and Japanese prosody affects the acquisition of particular lexical pitch patterns. Our hope is that these insights will help to better define the currently available theoretical accounts on L2 acquisition of lexical pitch, in Japanese as well as in other languages.

Furthermore, production data are relatively scarce within the literature on the L2 acquisition of Japanese lexical pitch. Although the canonical study by Hirano-Cook (2011) does include a production study, the analysis is limited to accuracy scores reporting whether speakers produced target patterns correctly, based on native speaker judgments. The only other important study including L2 production data is the one by Sakamoto (2011), who conducted quantitative analyses of three parameters (f_0 peak location, degree of f_0 fall and f_0 range) to investigate foreign accent in L2 productions of Japanese pitch patterns. Our study aims to supply the literature with more production data, which is important because data on perception only show a limited glimpse of actual speech acquisition.

RESEARCH QUESTIONS AND HYPOTHESES

Our study assessed identification and production of four trimoraic Japanese lexical pitch patterns, as described in the previous section, by English-native advanced learners and native Japanese speakers. This was assessed in order to provide an answer to the following research question:

1. To what extent can a Suprasegmental Similarity Account explain the L2 acquisition of Japanese lexical pitch by English-native learners?

We will forward two hypotheses, based on two scenarios of categorical assimilation from L2 Japanese lexical pitch to L1 English suprasegmental categories.

Scenario 1: Japanese Lexical Pitch perceived as English Intonation

Under this scenario, we would predict that Japanese lexical pitch categories are perceived by English listeners in terms of their L1 intonational pitch categories. Assimilation from lexical pitch to utterance intonation has been considered in earlier studies (Ayusawa et al., 1995; Braun et al., 2014), and may be intuitively plausible because both types of prosody share pitch (f_0) as their primary acoustic component. For hypothetical purposes, we acknowledge the 22 categories defined by Ladd (2008, p. 82) to describe the L1 English intonational category inventory, even though the categorical nature of English intonation may be disputed (Post, Stamatakis, Bohr, Nolan, & Cummins, 2015). Under Scenario 1, each of our four lexical pitch categories would assimilate in a one-to-one fashion onto separate English intonational categories. For instance, H*LL pattern [1] would assimilate to the 'Fall' category H*L(L%). Similarly, LH*L pattern [2] would assimilate to the 'Stylised Low Rise' category L*+H(L%) and LHH* [3] and LHH patterns [0] to the 'Stylised Low Rise' L+H*H(L%) category. Because of this clear one-to-one 'Single-Category Assimilation', this hypothesis would predict that none of these particular patterns should be difficult to perceive for English listeners. However, there may be an overall, added difficulty in perceiving Japanese lexical pitch for English listeners because under this scenario, Japanese lexical pitch is processed as phrasal prosody rather than as word prosody, therefore occurring in a different 'semantic dimension'. We may therefore see an overall lower accuracy in identifying Japanese lexical pitch, but this should occur across the board rather than in one particular pattern. As for production, we do not expect any particular 'motoric output constraints', given the fact that in terms of pitch shape, our lexical pitch patterns can be deemed similar to English intonational curves.

Scenario 2: Japanese Lexical Pitch perceived as English Lexical Stress

Another scenario, as pointed out by Shport (2016, p. 762), is that English listeners compare Japanese lexical pitch categories to English lexical stress categories. This scenario may also be probable because both these types of word prosody are used for lexical distinctions, and because they can indicate word prominence. In this scenario, we could assimilate Japanese patterns [1], [2], [3] onto English first-syllable, second-syllable, or third-syllable patterns, respectively. The unaccented pattern [0] however, would not map readily onto a stress category because it lacks prominence. In this case, we apply the same logic as that of (Shport, 2016, p. 763) and expect pattern [0] to assimilate to the English third-syllable stress category as a 'Bad Exemplar'. This would imply that LHH pattern [0] would be perceptually difficult to tell apart from LHH* pattern [3], which is a 'Good Exemplar' of English third-syllable stress. As a result, we would predict that under this scenario, English speakers will perform worse than native Japanese speakers in the perception of pattern [0] because they confuse it often with pattern [3]. Unlike our prediction under Scenario 1, we do not foresee an increased difficulty in the 'semantic dimension' because in Scenario 2, English listeners process Japanese lexical pitch at the lexical level and not at the phrasal level. In terms of production however, we may expect to see difficulties with the production of the unaccented pattern [0], which has no correlate in English lexical stress because it lacks prominence.

Our hypotheses are summarised in Table 1. The study conducted to assess our research questions and predictions is outlined in the following sections.

Table 1. Hypotheses

	Scenario [1] **Assimilation to English intonation**	**Scenario [2]** **Assimilation to English lexical stress**
Predicted perception by L2 learners	Good for all patterns because of one-to-one mapping with intonational categories, but possibly worse perception overall because of processing at phrasal level instead of at lexical level.	Good for all patterns except for LHH pattern [0] because of confusion with LHH* pattern [3].
Predicted production by L2 learners	Good for all patterns because of similar pitch curves in English intonation.	Good for all patterns except for pattern LHH [0] because of lack of prominence.

METHODOLOGY

Participants

Two groups of speakers were recruited for this study. Group EN consisted of eight native speakers of English (six female, two male) who grew up in the UK in English-speaking households. At the time of the experiment (March 2016) their mean age was 21.5 (SD = 0.75). All were fourth year university students at the Japanese department of SOAS, University of London and had learned Japanese for 3.5 years in a formal setting for approximately 10 hours per week. They had all spent one year in Japan in their third year of university. Group JA consisted of eight native Japanese speakers (four female, four male) of standard Tokyo. At the time of the experiment (March-May 2016), their mean age was 29.5 (SD = 7.75). They had lived in Japan for the largest part of their lives and were in the UK as visitors, university students or company workers. During their stay abroad, they continued to speak Japanese at home. None of the participants were simultaneous bilinguals.

All participants voluntarily participated in a perception and production task, as outlined below, for which they signed a consent form and for which they were rewarded a token fee.

**Perception Task Stimuli**

Audio stimuli for the perception task consisted of 48 trimoraic nonce words and 6 real Japanese filler words[3] in isolation and in the carrier sentence _x ga kàite arimasu_ ‘x is written down’. This resulted in a total of (48*2) + (6*2) = 108 audio stimuli. All nonce words were phonotactically viable words in Japanese and were formed of sonorant syllables, for instance /mamano/. We decided to have a majority of nonce words to discount any effects of lexical knowledge of real words, which may have affected identification. The stimuli were produced by a female native speaker of standard Tokyo Japanese. The speaker was instructed to produce the stimuli with the following pitch patterns, which differ in the presence and location of the pitch accent (the lexically predefined fall in f_0, indicated by ‘*’) in the trimoraic word:

-H*LL pattern [1]: first-mora accented

-LH*L pattern [2]: second-mora accented

-LHH*$^{/\circ}$_pattern [3]: third-mora accented_

-LHH pattern [0]: unaccented

The f_0 contours of all the perception stimuli per pattern are shown in Figure 2.

[3] _Bùdou_ ‘martial art’; _budou_ ‘grape’; _goyòu_ ‘order, need’; _goyou_ ‘misuse’; _namarì_ ‘accent, regionalism’; _namari_ ‘lead (metal)’.

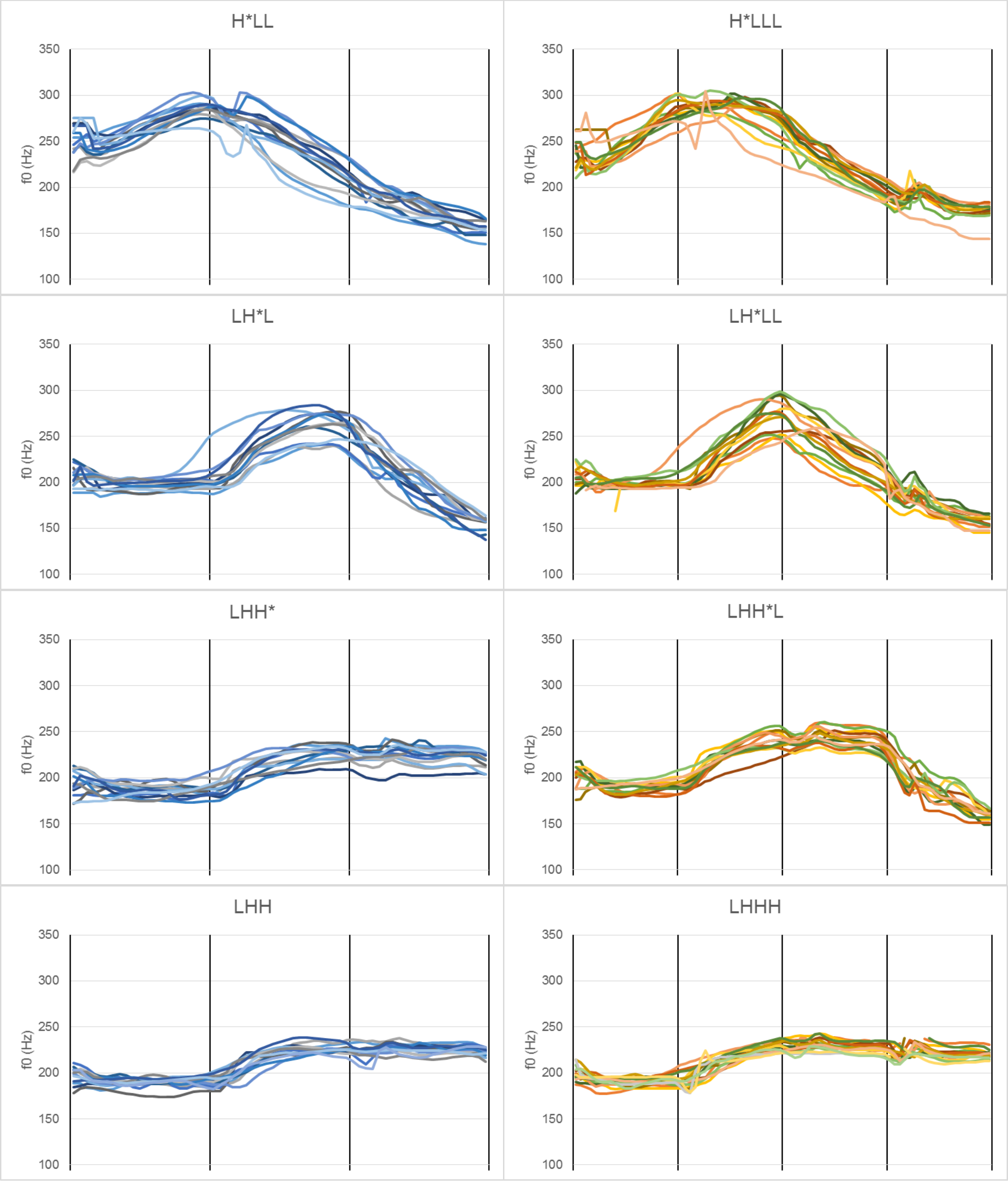

Figure 2. f_0 contours of the 108 perception stimuli. The left-side windows show stimuli in isolation. The right-side windows show stimuli in a sentence context including the monomoraic particle /ga/. The vertical black bars indicate mora boundaries

The contours in Figure 2 clearly show high f_0 values in H morae and low f_0 values in L morae. The drop in f_0 in the accented H* morae is also visible, with the f_0 values starting to fall towards the end of the H* mora. Sometimes, the fall only starts in the mora after the H*: this phenomenon is known as *ososàgari* 'delayed pitch fall' and is a typical feature of native pitch patterns, particularly in first-mora accented words (Ishihara, 2006). This can also be observed in our stimuli of H*LL pattern [1]. The earlier-mentioned similarity between final-mora accented LHH° patterns [3] and unaccented LHH patterns [0] in isolation can also be observed: both appear to show the same f_0 contours (bottom two left windows). The prescribed f_0 fall in final-mora accented patterns [3] can only be observed in the sentence context, when a fourth mora from a subsequent word can accommodate the transition to a lower pitch, resulting in a LHH*.L pitch pattern.

Despite the apparent similarity between isolated final-mora accented LHH° stimuli and isolated unaccented LHH stimuli, some studies have suggested that the two patterns are in fact different because accented final morae exhibit a higher f_0 than unaccented final morae, which can be auditorily perceived by some native listeners (Pierrehumbert & Beckman, 1988; Warner, 1997, p. 58). Although the native speaker who provided the stimuli indicated that she did not distinguish between final-mora accented and unaccented patterns, neither in listening nor in speaking, it may have been that she inadvertently produced these patterns differently. To make sure that this was not the case, the f_0 peaks in the final morae of the isolated final-mora accented LHH° stimuli and isolated unaccented LHH stimuli were assessed.

For isolated LHH° stimuli, the mean f_0 peak in the final mora was 231.53 Hz (SD = 6.31). For isolated LHH stimuli, the mean f_0 peak in the final mora was 230.35 Hz; (SD = 5.84). A one-way ANOVA revealed no significant difference between these values *F*(1,26)

= .263, $p > 0.05$. (Statistical significance was tested using *SPSS* (IBM Corp., 2017)). We therefore concluded that there was no difference between final-mora accented and unaccented pitch patterns in isolation. As such, we counted LHH° pattern [3] stimuli as LHH pattern [0] stimuli in the analysis. This led to an overall distribution of patterns across the stimuli as described in Table 2. As will be explained later, we will primarily focus on the sentence-embedded nonce stimuli because these allow for a four-way comparison between four distinctive pitch patterns.

Table 2. Pitch pattern distribution of perception stimuli

Pattern	**Notation**	**N° isolated stimuli *(Real words)***	**N° sentence-embedded stimuli *(Real words)***
[1]	H*LL	12 *(1)*	**12 *(1)***
[2]	LH*L	12 *(1)*	**12 *(1)***
[3]	LHH*	-	**12 *(1)***
[0]	LHH LHH°	24 *(4)*	**12 *(3)***
Total		48 *(6)*	**48 *(6)***

Perception Task Stimuli: Acoustic Details

To gain a little more insight in the sentence-embedded nonce stimuli beyond the f_0 contours, we report on two acoustic parameters: f_0 peak and f_0 decrease. These two parameters are of particular interest because they are considered to be important perceptual cues to pitch pattern recognition (Hasegawa & Hata, 1992; Shport, 2011, pp. 48–49).

Values for f_0 peak were taken from observations of the maximum f_0 value in the accented morae H* (and in the last mora of [0] LHH) using the 'Get Maximum Pitch' command in *Praat* (Boersma & Weenink, 2019). Any f_0 maxima elicited by noise were disregarded.

Values for f_0 decrease were obtained by calculating the percentage change from the maximum f_0 value in the accented mora H* to the minimum f_0 value in the subsequent mora. Note that for final-mora accented LHH* pattern [3] and unaccented LHH pattern [0], the maximum f_0 was taken from the final mora in the word and the minimum f_0 from the word *ga* from the carrier sentence. The values for both parameters are shown in Table 3.

Table 3. f_0 peak and f_0 decrease of sentence-embedded perception stimuli

Parameter		**H*LL [1]**	**LH*L [2]**	**LHH* [3]**	**LHH [0]**
f_0 peak	Value (Hz)	286.61	274.04	245.10	235.05
	SD	*11.26*	*18.62*	*7.05*	*5.21*
f_0 decrease	Value (%)	9.60	26.27	35.18	7.74
	SD	*6.03*	*8.42*	*2.17*	*1.82*

Table 3 shows that the accented mora in H*LL pattern [1] had the highest average f_0 peak (286.61 Hz) and the unaccented final mora in LHH pattern [0] the lowest peak (235.05 Hz). All f_0 peak values were subjected to a Welch's ANOVA with *Pattern* ([1], [2], [3], [0]) as repeated measure. This showed that there was a significant effect of *Pattern* $F(3, 22.920) = 74.437$ $p < 0.001$. Games-Howell corrected pairwise comparisons further showed that all the f_0 peak values significantly differed between patterns (all $p < 0.001$), except between H*LL [1] and LH*L [2], and between LHH* [3] and LHH [0].

The values for f_0 decrease show that the f_0 decrease was largest in LHH* pattern [3] (35.18%) and smallest in LHH pattern [0] (7.74%). A Welch's ANOVA with *Pattern* ([1], [2], [3], [0]) as repeated measure showed that the effect of *Pattern* was significant $F(3, 23.197) = 453.031$, $p < 0.001$. Games-Howell corrected pairwise comparisons further showed that all the f_0 decrease values significantly differed between patterns (all $p < 0.001$) except between

H*LL [1] and LHH [0].

Production Task Stimuli

The stimuli used for the production experiment were 18 trimoraic words that were also used in the perception experiment. The target words were two members of a minimal pair for all contrastive combinations (i.e. a minimal pair with pitch patterns [1]&[2], [1]&[3], etc.) and all the six real words used in the perception experiment, with an equal distribution of nonce words across each pattern (i.e. three nonce words per pattern). The nonce stimuli were presented to the participants written in the Japanese hiragana syllabic script and the real word stimuli in a combination of Sino-Japanese kanji and hiragana scripts. Additionally, target pitch patterns were visually displayed with dots and lines, which is a common method used in textbooks and Japanese dictionaries to indicate the pitch pattern (Hasegawa, 1995).

Procedures: Perception Task

The perception task was carried out in a silent classroom at SOAS, University of London for group EN and a sound-attenuatead room at the University of Cambridge for group JA. Participants listened to the audio stimuli through headphones at a comfortable hearing level. After signing a consent form, participants were seated in front of a PC on which the experiment software *OpenSesame* (Mathôt, Schreij, & Theeuwes, 2012) was running. Oral and written instructions were given in the participant's native language by the first author. Participants were told that they would listen to a mix of real and nonce trimoraic words, presented in isolation and in a carrier sentence, with an accent on the first, second, third, or on none of the morae, and that their task was to select one of these four accent patterns by pressing keys 1, 2, 3, or 4 on a keyboard. Both in the English and the Japanese instructions, only the generic word 'accent' or アクセント /akusento/ was used, rather than using terminology such as 'prominence', 'stress', 'pitch fall', 'intonation', etc.

Participants first completed a practice round with eight stimuli which had the same distribution of pitch patterns as in the main task. Feedback was given in the practice round. After being given the opportunity to ask any questions and to adjust the volume to a comfortable hearing level, participants took the main task without feedback. Timeout per trial was 10 seconds and in case of non-response the trial was marked as 'incorrect' (the average percentage of non-responses was 1.61% across all participants). The total task took approximately 8 minutes. Reaction times were not measured.

Procedures: Production Task

The production task was conducted directly after the perception task. Participants read out loud the production stimuli from a piece of paper (from a PC monitor for group JA). We specifically opted for a reading task rather than a mimicry task because we were interested how L2 learners would produce the pitch patterns without aid of an auditory example. Participants' productions were recorded using a portable 24bit/96KHz *H4 Next Hand Recorder* (Zoom Corporation) at a sampling frequency of 44.1 KHz. Participants were asked to read out loud the 18 words twice in isolation and twice in the carrier sentence *x ga kàite arimasu* 'x is written down'. The participants first read out loud all the words for pattern [1], and then for patterns [2], [3], and [0]. The first author did not pronounce the words. Participants were given the opportunity to pronounce a word again if they felt that they had produced it incorrectly. The production task took approximately 5 minutes.

Production Task: Labelling Procedure of Recorded Data

In the production task, participants were asked to pronounce words with a target pitch pattern [1], [2], [3], or [0]. The recorded audio files were saved as .wav files and analysed in the software *Praat* (Boersma & Weenink, 2019). They were then labelled by the authors with a categorical (phonological) pitch pattern in the style of Kawahara (2015), such as H*LL for pattern [1] in order to determine whether a production was phonologically correct (with regard to the target pattern) or incorrect. Although participants produced each stimulus twice, only the first production was labelled. However, if the first production was phonologically incorrect, or if the f_0 contour was unclear, the second production was labelled, provided that this was a better alternative.

The labelling was done *Praat* using a combination of auditory analysis and visual inspection of f_0 contours. We deemed this labelling method appropriate for the purposes of our research question, which is aimed at overt phonological accuracy. It needs to be mentioned however, that it was not always directly clear what categorical pitch pattern to assign to f_0 contours. In particular, differentiating between LHH* and LHH patterns in sentence context sometimes had to be based on subtle differences. As is shown in Figure 3, it was sometimes difficult to tell from observing the f_0 contour alone whether a decrease in f_0 in the mora following the trimoraic word was due to a pitch accent or due to a general 'downtrend' effects (Pierrehumbert & Beckman, 1988, p. 57).

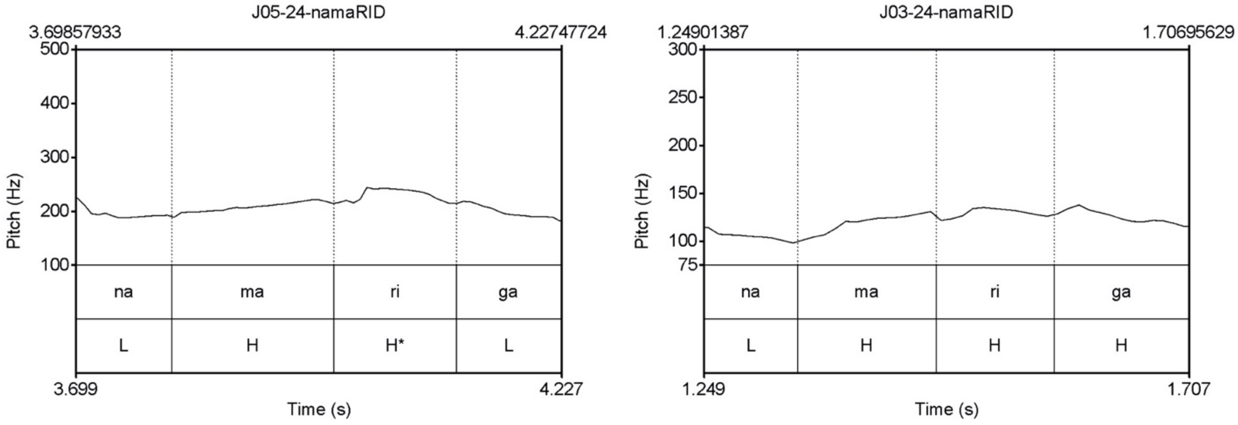

Figure 3. Labelling of [3] LHH* and [0] LHH stimuli

Therefore, in these cases, we would apply a threshold of at least 25% in f_0 decrease from the f_0 maximum in the final H of the word to the f_0 minimum in the word ga from the carrier sentence in order to label a mora as accented (H*). This threshold was based on the values of f_0 described in Table 3.

Participants sometimes produced atypical patterns, such as monotonous tones, which were labelled accordingly as LLL or HHH.

Production Task: Phonetic Analysis

Previous studies have indicated that even though when audibly similar, production of pitch patterns by native Japanese speakers can still be discerned from non-native production when looking at certain phonetic properties. An example is the location of the f_0 peak in accented H* morae. Sakamoto (2011) found for instance that the f_0 peak was more consistently located in native speakers as opposed to non-native speakers in audibly similar patterns. To investigate if similar phonetic differences could be observed in our study, we analysed f_0 peak location from a subset of phonologically correct stimuli produced by six participants from EN and six participants from JA. The subset of these stimuli and the participants who produced them are described in Table 4. They are four stimuli from H*LL pattern [1]. There were not enough stimuli correctly produced by the same set of participants from patterns [2], [3] or [0] to analyse. From the subset of stimuli, values were taken from both productions per participant, resulting in eight tokens per participant.

The parameter for f_0 peak location was calculated by dividing the duration (in ms) from stimulus onset until the f_0 peak by the duration (in ms) of the accented mora /a/. This resulted in a percental value indicating the relative temporal location of the f_0 peak. The f_0 peak location was determined by observing f_0 contours in *Praat* and by identifying the highest f_0 value in the vowel of the accented mora (or of the subsequent mora, if the f_0 peak was only realized then) before the drop in f_0 using the 'Get Maximum Pitch' command. Any f_0 peaks elicited by other events or noise were disregarded.

Table 4. Subset of stimuli for phonetic analysis

Stimulus	Pattern	Participants EN	Participants JA
àmaro	H*LL [1]	EN-1 (f) EN-2 (f) EN-3 (m) EN-4 (f) EN-5 (f) EN-8 (f)	JA-1 (m) JA-2 (f) JA-3 (m) JA-5 (f) JA-7 (f) JA-8 (m)
àmaro ga	H*LL.L [1]		
àrino	H*LL [1]		
àrino ga	H*LL.L [1]		

RESULTS

Perception Task: Overall Perception Accuracy

Overall perception accuracy was calculated as percentage of correctly identified stimuli of the total of 96 nonce stimuli. The accuracy means were 51.30% (SD = 18.14) for EN and 65.76% (SD = 10.61) for JA. A one-way ANOVA with *Group* (EN, JA) as between-subject factors revealed that the difference between groups was statistically insignificant *F*(1,14) = 3.782, *p* > 0.05.

Perception Task: Accuracy per Pattern

Previous studies have indicated that some pitch patterns may be inherently more difficult to perceive than others (see Hirata (2015, p. 735)). To assess if this was the case in our study, the perception accuracies per pattern ([1], [2], [3], and [0]) were compared. In order to make comparisons with four audibly different pitch patterns, we only report on accuracy of sentence-embedded nonce stimuli (Recall that patterns [3] and [0] are only acoustically different from one another in a sentence context). The per-pattern accuracies are displayed in Table 5 and Figure 4.

Table 5. Mean perception accuracies per pattern

	H*LL **Pattern [1]**	**LH*L** **Pattern [2]**	**LHH*** **Pattern [3]**	**LHH** **Pattern [0]**
EN	**63.54**	**45.83**	**21.87**	**45.84**
SD	*29.19*	*30.86*	*20.86*	*21.36*
JA	**83.33**	**63.54**	**20.83**	**60.42**
SD	*16.06*	*21.33*	*14.08*	*29.12*

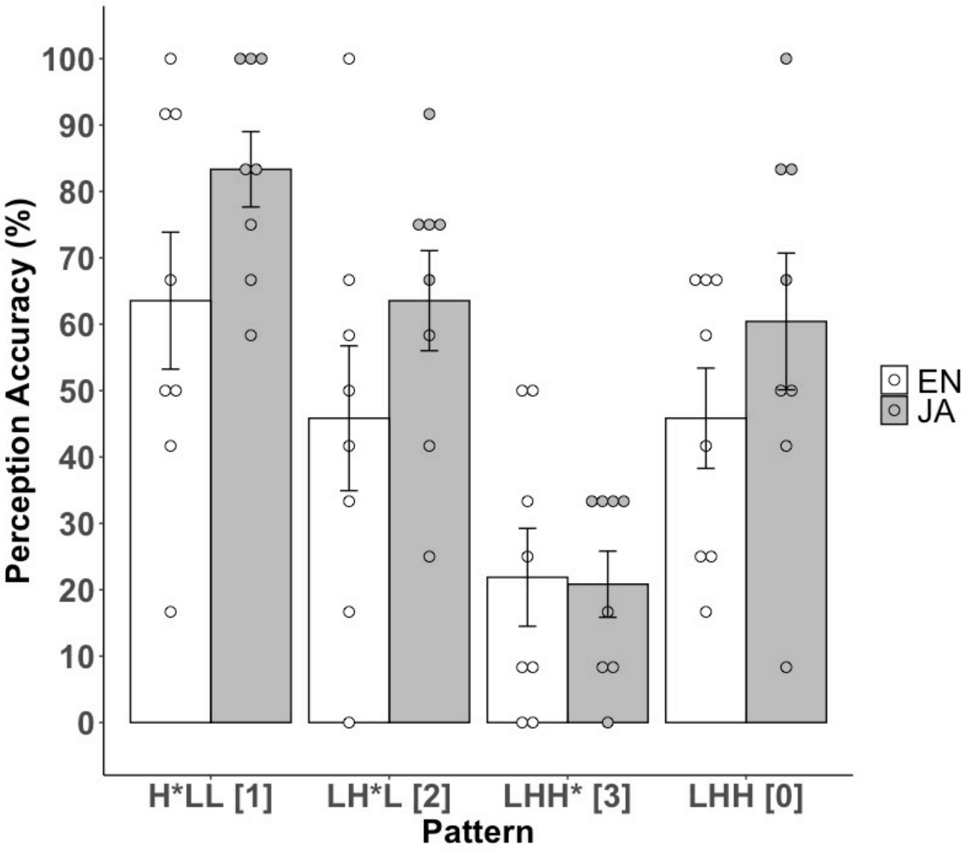

Figure 4. Bar charts showing mean perception accuracies per pattern (sentence-embedded nonce stimuli only). Errors bars = +/- 1 SE. Dots represent individual data points.

To investigate the effects of Pattern type, a two-way mixed ANOVA with *Pattern* ([1], [2], [3], [0]) as within-subject factor and *Group* (EN, JA) was carried out. This revealed a significant main effect of *Pattern* $F(3,42) = 23.987$, $p<0.001$ but not of *Group* $F(1,14) = 2.013$, $p > 0.05$. There was no significant interaction between *Pattern* and *Group* $F(3,42) = 1.147$, $p > 0.05$. For both participant groups, Bonferroni-adjusted pairwise comparisons showed that perception accuracy for the H*LL pattern [1] was significantly higher than that of patterns [2] and [3] (respectively $p < 0.05$ and $p < 0.001$). LHH* pattern [3] yielded significantly lower perception accuracies than all other patterns [1], [2], and [0] (respectively $p < 0.001$; $p < 0.001$ and $p < 0.01$). It thus appears that pattern [1] was notably easy to perceive, and pattern [3] notably difficult to perceive, and that this was the case in both native and non-native listeners.

Perception Task: Confusion Matrices

This section presents confusion matrices (cf. Lee, Tao, & Bond (2010)) in order to investigate the type of mistakes participants made. The perception confusion matrices are shown in Table 6. The vertical axis of the table displays the target answer, and the horizontal axis displays the given response. The values in each cell represent the percentage of given response options per target answer. For instance, cell [1];[1] in the top table in Table 6 shows that on average, 64% of the time English participants *correctly* responded [1] when the target answer was [1]. By contrast, cell [1];[2] shows that on average, 8% of the time participants *incorrectly* responded [2] when the target answer was [1].

Table 6. Perception task: Confusion matrices

	Group:	**Response**			
	EN	H*LL [1]	LH*L [2]	LHH* [3]	LHH [0]
Target	H*LL [1]	**64**	8	3	22
	LH*L [2]	15	**46**	15	22
	LHH* [3]	13	40	**22**	21
	LHH [0]	8	18	23	**46**

	Group:	**Response**			
	JA	H*LL [1]	LH*L [2]	LHH* [3]	LHH [0]
Target	H*LL [1]	**83**	5	0	10
	LH*L [2]	13	**64**	10	13
	LHH* [3]	1	29	**21**	46
	LHH [0]	5	10	22	**60**

The confusion matrices reveal that, in group EN, the least correctly identified pattern [3] was most often confused with pattern [2]. In group JA, pattern [3] was most often incorrectly identified as pattern [0], followed by pattern [2].

Production Task: Overall Production Accuracy

The overall production accuracy was calculated as the percentage of correct productions, based on the labelling as described in the methodology section, of the total amount of nonce stimuli. Production accuracy of the 24 nonce stimuli was 59.37% (SD = 21.21) for EN and 83.85% (SD = 8.46) for JA. We report only the accuracy of the nonce stimuli to discount any influence that lexical knowledge of the real words may have had on the production of the real words. The production accuracy means were subjected to a one-way ANOVA, which revealed a significant effect for *Group* $F(1,14) = 9.188$, $p < 0.01$, demonstrating that the mean production accuracy of the native Japanese group was significantly higher than that of the English-native L2 learner group.

Production Task: Accuracy per Pattern

Just as in the perception task, we analysed per-pattern accuracy to see whether any patterns yielded particularly higher or lower accuracy scores. The production accuracies per pattern are shown in Table 7 and Figure 5.

Table 7. Mean production accuracies per pattern

	H*LL	**LH*L**	**LHH$^{\circ/*}$**	**LHH**
	Pattern [1]	**Pattern [2]**	**Pattern [3]**	**Pattern [0]**
EN	**85.42**	**75.00**	**43.75**	**33.33**
SD	*35.00*	*38.83*	*38.77*	*36.73*
JA	**95.83**	**89.58**	**72.92**	**77.08**
SD	*11.78*	*23.47*	*26.63*	*25.10*

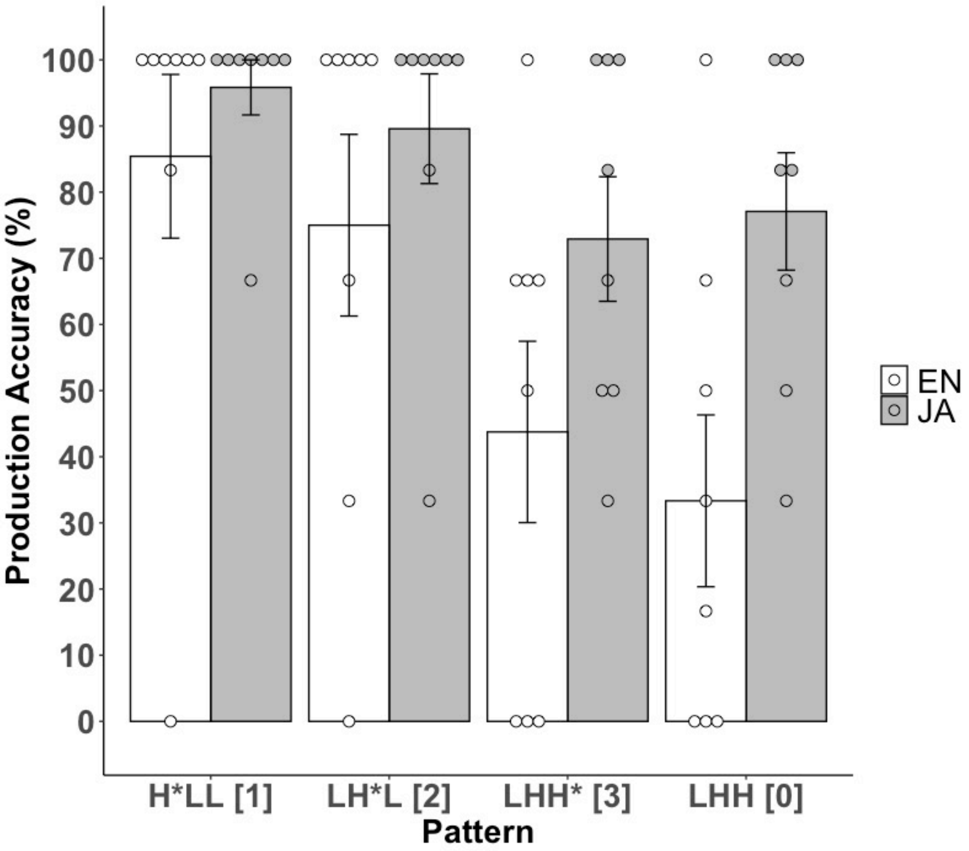

Figure 5. Bar charts showing mean production accuracies per pattern (nonce stimuli only). Errors bars = +/- 1 SE. Dots represent individual data points.

To investigate the effects of Pattern type, a two-way mixed ANOVA with *Pattern* (1], [2], [3], [0]) as within-subject factor and *Group* (EN, JA) was carried out. This revealed a significant main effect of *Pattern* $F(3,42) = 5.332$, $p < 0.01$ and a significant main effect of *Group* $F(1,14) = 9.187$, $p < 0.01$. There was no significant interaction between *Pattern* and *Group* $F(3,42) = 0.997$, $p > 0.05$. Bonferroni-adjusted pairwise comparisons revealed that overall, H*LL pattern [1] was produced significantly better than LHH pattern [0] ($p < 0.01$), and that overall, group JA had a significantly higher mean production accuracy than group EN ($p < 0.01$).

Although no combined effect of *Pattern*Group* was observed, it is of interest to compare the different accuracies between the two participant groups per pitch pattern because *Group* yielded a significant main effect on accuracy scores in the production task.

This suggests that at least in some patterns, English participants may have significantly performed differently than Japanese participants. To investigate this, four independent sample t-tests were conducted for the accuracy scores per pattern between groups. This showed that only within the LHH pattern [0], phonological production accuracy of group EN was significantly lower than that of group JA with a mean difference of 43.78 percentage points ($t(14) = -2.781$, $p < 0.05$).

Production Task: Confusion Matrices

Confusion matrices for production accuracies are shown in Table 8. The confusion matrices show that in group EN, the pattern with the lowest mean accuracy, LHH pattern [0], was often mispronounced as LH*L pattern [2] or as an atypical, often monotonous pattern, which is listed under ‘other’. LHH* pattern [3] was on average equally confused with LH*L pattern [2] and H*LL pattern [1]. For group JA, the patterns yielding relatively low accuracies, namely LHH* pattern [3] and pattern LHH [0], appeared to be confused with one another to the same extent.

Table 8. Production task: Confusion matrices

	Group:			*Response*		
	EN	[1]	[2]	[3]	[0]	*Other*
Target	[1]	**85**	13	0	0	2
	[2]	13	**75**	8	4	0
	[3]	17	21	**44**	10	8
	[0]	15	27	6	**33**	19

	Group:			*Response*		
	JA	[1]	[2]	[3]	[0]	*Other*
Target	[1]	**96**	4	0	0	0
	[2]	0	**90**	2	8	0
	[3]	4	4	**73**	17	2
	[0]	6	0	17	**77**	0

Production Task: Phonetic Accuracy

Analysis of 96 data points (8 tokens x 12 participants) from the subset of tokens described in Table 4 showed that on average, the f_0 peak location in the accented initial mora was located at 63.9% (SD = 22.8) in EN and at 86.6% in JA (SD = 22.4). This is shown in Figure 6.

These data were subjected to two-way mixed ANOVA with *Token* (8 levels: Token 1, Token 2 (..), Token 8) as within-subject factor and *Group* (EN, JA) as between-subject factor. This revealed no significant effect of *Token F*(7,70) = 1.820, *p* > 0.05 but did yield a significant main effect of *Group F*(1,10) = 12.975, *p* < 0.01. There was also significant interaction between *Token***Group F*(7,70) = 2.921, *p* < 0.05. Bonferroni-adjusted pairwise comparisons revealed that f_0 peak location was significantly later in JA compared to EN by 27.7 percentage points (*p* < 0.01). Pairwise comparisons for Token are not reported.

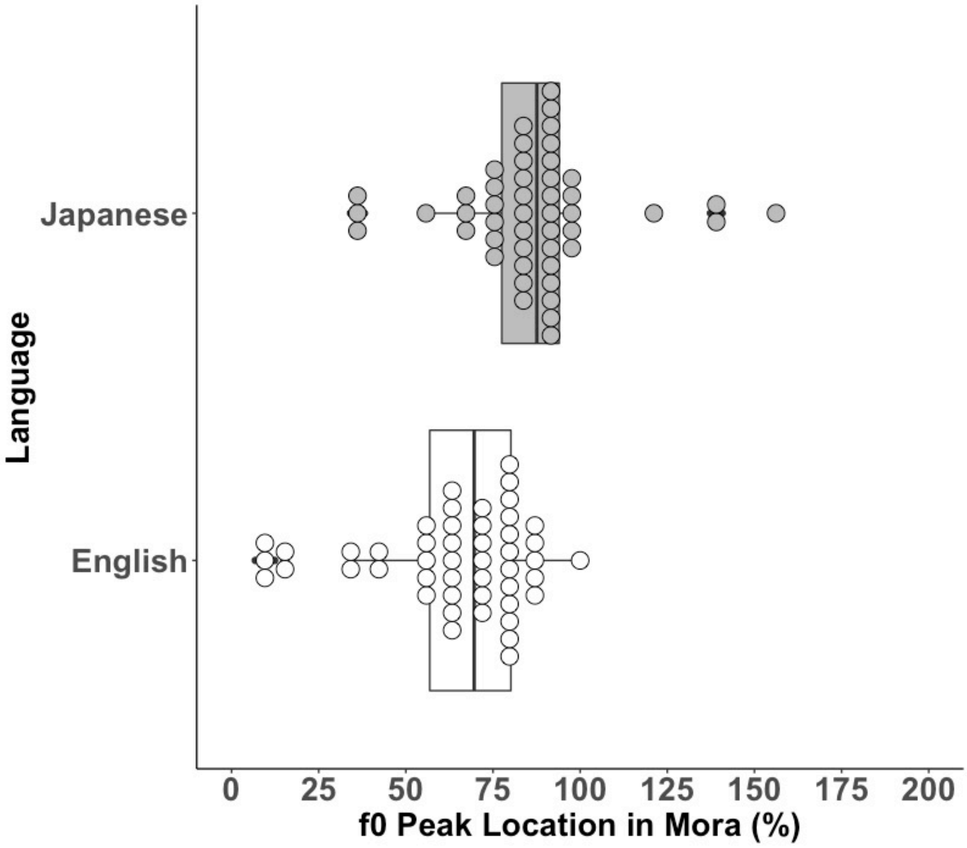

Figure 6. Box-and-whisker plot showing f_0 peak location in initial-accented morae /a/. Dots are individual data points.

Correlation between Perception and Production

The majority of work on speech acquisition focuses on perceptual studies, as perception is often seen as an indicator of overall speech acquisition (Munro & Bohn, 2007, p. 9). To assess to what degree individual perception performance predicted production performance in our study, we assessed the correlation between total accuracy in the perception task and in the production task.

The scatterplot displaying total production accuracy against total perception accuracy is shown in Figure 7. It can be clearly observed that the Japanese group perform relatively well in both perception and production, and that there were some English participants who performed relatively poor in perception, but also in production. A simple linear regression showed that for all participants taken together, perception performance significantly predicted production performance $F(1, 15) = 22.844$, $p < 0.001$, $R^2 = 0.620$.

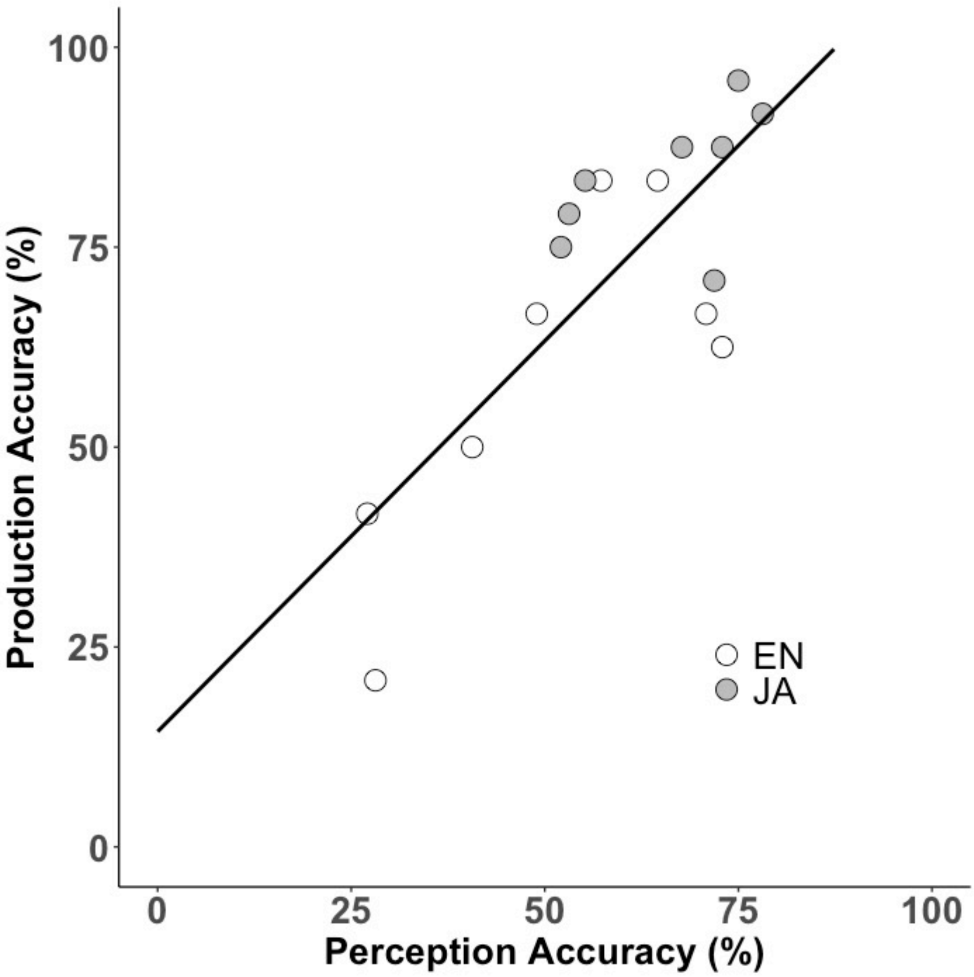

Figure 7: Scatterplot and linear regression curve showing production accuracy against perception accuracy ($p < 0.001$, $R^2 = 0.620$)

DISCUSSION

Overall Performance

We found that advanced L2 learners of Japanese did not perform significantly different to native speakers in perception of pitch patterns. This finding was not completely surprising, and it falls in line with findings of advanced L2 listeners displaying perception accuracy levels nearing that of native listeners (Hirano-Cook, 2011; Sakamoto, 2011). This indicates that L2 experience can lead to native-like perceptual performance in identification of pitch patterns.

However, we did find that advanced L2 learners significantly underperformed in comparison to native speakers when it came to producing phonologically correct target pitch patterns on nonce words, and this difference was particularly stark in unaccented LHH pattern [0].

Performance per Pattern

We also looked at the perception and production of specific pitch patterns, and observed that some patterns were easier to perceive and produce than others. In the perception task, accuracy per pattern ranked from high to low in the order [1]>[2]>[0]>[3]. That is, H*LL pattern [1] was most easily identified, followed by LH*L pattern [2], LHH pattern [0] and finally LHH* pattern [3]. These findings, at least for perception, are largely in line with earlier studies on L2 perception, such as Toda, (2001) and Shport (2016), although Ayusawa (2003) and Hirano-Cook (2011) report different hierarchies, with unaccented patterns (similar to our pattern [0]) being perceived most accurately.

We observed a largely similar per-pattern difficulty hierarchy in production, with accuracy per pattern ranked from high to low in the order [1]>[2]>[3]≈[0]. The only difference with regard to perception was performance in the unaccented LHH pattern [0].

This pattern was least accurately produced among English speakers, and the production performance was significantly lower than that of the native Japanese speakers. This suggests that in production, unaccented LHH patterns [0] pose a particular difficulty, even for the advanced English-native L2 learner of Japanese. This finding partially corroborates earlier findings from previous literature (Hirano-Cook, 2011, p. 93). Sakamoto (2011) found however, and in contrast to our findings, that unaccented LH words (similar to our pattern [0]) were the easiest to produce for L2 learners.

A variety of methodology in different studies (different moraic lengths of stimuli; nonce and real words; different participant groups, different speech accuracy rating methods, etc.) may explain why there seems to be no conclusive answer on what pattern is the most difficult for L2 learners (Hirata, 2015). Yet, it is worth discussing why we observed our particular per-pattern hierarchy. As shown in the results of the perception test, the final-mora accented LHH* pattern [3] yielded significantly lower perception accuracies than all other patterns. By contrast, first-mora accented H*LL patterns [1] yielded significantly higher accuracy scores than other patterns. It may be that inherent acoustic properties of the patterns were of influence here. As described earlier in Table 3, first-mora accented H*LL patterns [1] had a high f_0 peak in the accented H*, but a relatively small f_0 decrease onto the next mora. By contrast, final-mora accented LHH* patterns [3] had a relatively low f_0 peak in H*, but a drastic f_0 change onto the next mora. These acoustic properties may have contributed to the fact that pattern [1] was identified so well, and pattern [3] so poorly: it has been widely established that in speakers of stress languages, f_0 height is the primary cue for pitch perception (Antoniou & Chin, 2018, p. 2; Francis et al., 2008; Gandour & Harshman, 1978; Shport, 2011, pp. 48–49; Wayland & Guion, 2004). This could explain why English listeners identified first-mora accented H*LL patterns [1] (high f_0) so well and final-mora accented LHH* patterns [3] (low f_0) so poorly. For speakers of Japanese, which is typologically in between a stress language and a tone language, there are indications that rather than f_0 height, both f_0 peak location and f_0 decrease affect pitch identification (Hasegawa & Hata, 1992). It is therefore slightly puzzling why LHH* pattern [3] also yielded such low perceptual identification scores in Japanese listeners, as it would have been expected that this pattern, which had the strongest f_0 decrease, would be the most acoustically salient. However, if for Japanese listeners, f_0 height was also a strong acoustic cue, stronger than f_0 decrease, then the low identification rates for LHH* pattern [3] and high identification rates for H*LL pattern [1] can be explained in the same way as for English listeners.

There may be other, non-phonetic factors that could explain the significantly high accuracies for pattern [1] and significantly low accuracies for pattern [3] in perception. For instance, most words in English have stressed initial syllables (Cutler & Carter, 1987; Ernestus & Neijt, 2008), and it has been reported that adults tend to produce stress on the initial syllable when confronted with trisyllabic pseudowords (Baker & Smith, 1976). It may therefore be that English listeners were better attuned to hearing first-mora accented words. Similarly, among the accented trimoraic pitch patterns in Japanese, first-mora accented patterns [1] are the most frequent, and final-mora accented patterns [3] the most infrequent (Kubozono, 2012). Therefore, word frequency effects, from both English and Japanese, may have had a general influence on the relative ease of identifying and producing pattern [1] and the relative difficulty of identifying and producing pattern [3]. This would fall in line with the 'frequency dimension' in suprasegmental speech learning as proposed by Mennen (2015).

Phonetic Accuracy

A phonetic analysis of a subset of stimuli revealed that native Japanese speakers' productions differed significantly from productions by L2 learners in terms of f_0 peak location, even though the productions were phonologically and audibly identical. Native speakers produced significantly later f_0 peaks in H*LL stimuli than did L2 learners. These findings are slightly puzzling as they do not fall in line with similar findings by Sakamoto, who found no such difference between native speakers and L2 learners (Sakamoto, 2011, p. 280), but they may be explained by the fact that a relatively late f_0 peak realised at the end of the mora, i.e. *ososàgari* 'delayed fall', is a typical aspect of native production of first-mora accented words (Ishihara, 2006). Indeed, the f_0 peak in some of the native speaker productions was located in the mora after the accented H* mora. Moreover, findings from a recent study by Graham and Post (2018) suggest that L1 background can influence the systematic location of the f_0 peak in pitch accents. In their study, English native speakers realised a significantly earlier f_0 peak in initial-syllable accented English words than did Japanese L2 learners. It could therefore be argued that in our data, English speakers were influenced by their L1, resulting in a significantly earlier f_0 peak on first-mora accented Japanese words in comparison to native Japanese speakers.

Compatibility with Hypotheses

We originally asked to what extent a Suprasegmental Similarity Account of L2 speech acquisition, rooted in the PAM, SLM, and LILt models, could explain the acquisition of Japanese lexical pitch by advanced English-native learners.

We had two hypothetical scenarios of categorical assimilation: In Scenario 1, we hypothesised that English listeners would perceive Japanese lexical pitch in terms of similar-sounding intonational categories. This implied that because of a clear one-to-one Single-

Category Assimilation, no particular pattern should be more difficult than others to perceive for English listeners. However, because of a discrepancy in the functional ‘semantic dimension’, i.e. processing at the phrasal rather than at the lexical level, overall perception by English listeners was expected to be worse. As for production, we hypothesised that because of the similarity in terms of pitch contours of Japanese lexical pitch patterns to English intonational patterns, there should be no particular disadvantage for L2 learners in producing the pitch patterns. Our findings were incongruent with the hypothesis under Scenario 1, because we observed native-like perception accuracy, but significantly lower production accuracy in L2 learners.

In Scenario 2, we considered that Japanese lexical pitch categories would assimilate to English lexical stress categories. Following Shport (2016), we assumed that the unaccented LHH patterns [0] would assimilate onto English third-syllable stress categories as ‘Bad Exemplars’ (Category Goodness Assimilation, in PAM terms (Best & Tyler, 2007)), thus making patterns [0] relatively difficult to perceive and easily confused with third-mora accented LHH* pattern [3]. For production, we also predicted relatively lower accuracies for the unaccented pattern [0], because there is no ‘unaccented’ English lexical stress.

Our findings appear to fit better under this scenario, although not perfectly. For perception, we did find, as predicted, that English listeners relatively often confused pattern [0] with pattern [3], as shown in the confusion matrix in Table 6, but this did not result in a significantly worse perception for pattern [0] compared to native listeners. This however, and as mentioned before, may have been a result of the fact that our L2 learners were advanced learners (cf. Hirano-Cook, 2011; Sakamoto, 2011). In addition, it may be result of the nature of the identification task, in which selecting an ‘unaccented’ pattern may be a safe option in case of doubt, yielding to higher accuracy scores for unaccented tokens

(Shport, 2016, p. 741).

For production, our predictions under Scenario 2 seem to be congruent, as we indeed found that for English speakers, the unaccented LHH pattern [0] was notably harder to produce than other patterns in comparison to native Japanese speakers. Our production confusion matrix in Table 8 showed that English speakers tended to mispronounce the unaccented LHH pattern [0] in quite an inconsistent way: either as second-mora accented LH*L pattern [2] or first-mora accented H*LL pattern [1], or as other patterns such as a hypercorrected monotonous HHH. This suggests that rather than just due to systematic confusion with one other pitch pattern category (a problem in the 'systemic dimension'), the production of the unaccented, prominence-lacking LHH pattern [0] is problematic due to L1-L2 dissimilarities in several other 'dimensions' (cf. Mennen, 2015). First of all, it may be due to a dissimilarity in the 'frequency dimension': In Japanese, prominence-lacking lexical pitch patterns are the most frequent (Kubozono, 2012), but in English prominence-lacking lexical stress patterns do generally not occur at all. Therefore, it may be that our English speakers were inclined to indicate prominence *somewhere* in the word through a fall in f_0, even though unaccented Japanese words do not require this. Indeed, the rather sporadic distribution of incorrectly accented productions of pattern [0] in our confusion matrix in Table 8 would suggest that this was the case. In addition, unaccented LHH patterns may be inherently problematic because they require 'maintaining a higher pitch for more than two morae', which may be difficult for English speakers as pointed out by Hirano-Cook (2011, p. 94). This may be the kind of 'motoric output constraints' as well as issues in the 'realisational dimension' of pitch height and timing that earlier theoretical models of speech acquisition refer to (Flege, 1995, p. 238; Mennen, 2015, p. 176). Our phonetic analysis, which showed a subtle, but significantly later alignment of f_0 peaks in Japanese productions compared to English productions, would support this notion of a general articulatory difficulty in accurately producing Japanese lexical pitch targets for English speakers.

All in all, we therefore suggest that in acquiring Japanese lexical pitch, English advanced L2 learners indeed rely on similarities with English suprasegmental speech by assimilating Japanese word prosody (lexical pitch) to English word prosody (lexical stress). As a result, English learners are good at perceiving Japanese lexical pitch, but they struggle at accurately producing it, particularly for lexical pitch patterns that are multidimensionally dissimilar to English word prosody. Although we showed that, overall performance in perception largely predicted performance in production, as illustrated in Figure 7, we also observed that for specific suprasegmental categories, difficulties in L2 production may persist even though there are not many difficulties in L2 perception. A Suprasegmental Similarity Account attempts to encompass these discrepancies between performance in perception and in production.

CONCLUSION

This paper has aimed to provide new insights in the applicability of a 'Suprasegmental Similarity Account' to the L2 acquisition of lexical pitch in Japanese. Combining elements of existing theories of segmental and suprasegmental speech acquisition (the PAM, the SLM and the LILt), we showed how perception and production in L2 learners was indeed guided by the notion of (dis)similarity in different 'dimensions' (Mennen, 2015). We argue that any L2 suprasegmental category that is multidimensionally different from any L1 category is particularly difficult to acquire. In our study, this particular category was the unaccented LHH pattern [0] for English-native L2 learners. Given that this pitch pattern is the most common in Japanese, an important pedagogical implication is thus that these patterns deserve more attention in Japanese language instruction.

An important question that remains is to what degree L1-L2 differences in each of the dimensions (systemic, realisational, semantic, and frequency) weigh in on the overall difficulty of acquiring L2 prosody (Mennen, 2015, p. 184) and how extralinguistic factors, such as general pitch sensitivity and musical experience (Antoniou & Chin, 2018) affect this. We hope that this paper will have laid a foundation for such future work to better understand the acquisition of word prosody in a second language.

ACKNOWLEDGEMENTS

We thank the volunteer participants and the academic staff at the School of Oriental and African Studies and at the University of Cambridge for helping in participant recruitment. We are grateful for the for the valuable feedback provided by anonymous reviewers on earlier versions of this paper.

BIBLIOGRAPHY

Antoniou, M., & Chin, J. L. L. (2018). What Can Lexical Tone Training Studies in Adults Tell Us about Tone Processing in Children? *Frontiers in Psychology*, *9*, 1–11. https://doi.org/10.3389/fpsyg.2018.00001

Ayusawa, T. (2003). Gaikokujingakusyūsya no nihongo akusento-intonēshon gakusyū [Acquisition of Japanese Accent and Intonation by Foreign Learners]. *Journal of the Phonetic Society of Japan*, *7*(2), 47–58.

Ayusawa, T., Nishinuma, Y., Lee, M. H., Arai, M., Odaka, K., & Hoki, N. (1995). Analysis of Perceptual Data on the Tokyo Accent: Results from 10 Language Groups. *Japic Research Report*, *2*, 25–32.

Baker, R. G., & Smith, P. T. (1976). A psycholinguistic study of english stress assignment rules. *Language and Speech*, *19*(1), 9–27. https://doi.org/10.1177/002383097601900102

Beckman, M. (1986). *Stress and non-stress accent*. Dordrecht: Foris.

Best, C. T. (1995). A direct realist view of cross-language speech perception. *Speech Perception and Linguistic Experience. Issues in Cross-Language Research*, 167–200. https://doi.org/10.1016/0378-4266(91)90103-S

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception. In M.J Munro & O.-S. Bohn (Eds.), *Second Language Speech Learning: the role of language experience in speech and production* (pp. 13–34). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.17.07bes

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer. Retrieved from http://www.praat.org/

Braun, B., Galts, T., & Kabak, B. (2014). Lexical encoding of L2 tones: The role of L1 stress, pitch accent and intonation. *Second Language Research*, *30*(3), 323–350. https://doi.org/10.1177/0267658313510926

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*(3–4), 133–142. https://doi.org/10.1016/0885-2308(87)90004-0

Dupoux, E., Sebastián-Gallés, N., Navarrete, E., & Peperkamp, S. (2008). Persistent stress ‘deafness’: The case of French learners of Spanish. *Cognition*, *106*(2), 682–706. https://doi.org/10.1016/j.cognition.2007.04.001

Ernestus, M., & Neijt, A. (2008). Word length and the location of primary word stress in Dutch, German, and English. *Linguistics*, *46*(3), 507–540. https://doi.org/10.1515/LING.2008.017

Flege, J. E. (1995). Second Language Speech Learning: Theory, Findings, and Problems. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233–277). Timonium, MD: York Press. https://doi.org/10.1111/j.1600-

0404.1995.tb01710.x

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics*, *36*(2), 268–294. https://doi.org/10.1016/j.wocn.2007.06.005

Gandour, J. T., & Harshman, R. A. (1978). Crosslanguage Differences in Tone Perception: a Multidimensional Scaling Investigation. *Language and Speech*, *21*(1), 1–33. https://doi.org/10.1177/002383097802100101

Graham, C., & Post, B. (2018). Second language acquisition of intonation: Peak alignment in American English. *Journal of Phonetics*, *66*, 1–14. https://doi.org/10.1016/j.wocn.2017.08.002

Gussenhoven, C. (2004). *The phonology of tone and intonation*. *The Phonology of Tone and Intonation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511616983

Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279. https://doi.org/10.1016/j.wocn.2011.11.001

Hao, Y. (2014). The Application of the Speech Learning Model to the L2 Acquisition of Mandarin Tones. *Proceedings of the 4th International Symposium on Tonal Aspects of Languages (TAL 2014)*, (1992), 67–70.

Hasegawa, Y. (1995). Against Marking Accent Location in Japanese Textbooks. *Japanese-Language Education Around the Globe*, *5*, 95–103.

Hasegawa, Y., & Hata, K. (1992). Fundamental Frequency as an Acoustic Cue to Accent Perception. *Language and Speech*, *35*(1–2), 87–98. https://doi.org/10.1177/002383099203500208

Hirano-Cook, E. (2011). *Japanese Pitch Accent Acquisition by Learners of Japanese: Effects of Training on Japanese Accent Instruction, Perception, and Production.* University of Kansas.

Hirata, Y. (2015). 18 L2 phonetics and phonology. In H Kubozono (Ed.), *Handbook of Japanese Phonetics and Phonology* (pp. 720–762). Berlin; Boston: De Gruyter Mouton.

IBM Corp. (2017). IBM SPSS Statistics for Macintosh, Version 25.0. Armonk. NY. https://doi.org/10.1080/02331889108802322

Iimori, Y. (2014). *Japanese Learners' Awareness of Pitch Accent And its Relationship to Their Oral Skills and Study Habits*. The Ohio State University.

Ishihara, T. (2006). *Tonal alignment in Tokyo Japanese*. University of Edinburgh.

Kawahara, S. (2015). The phonology of Japanese Accent. In H. Kubozono (Ed.), *Handbook of Japanese Phonetics and Phonology* (pp. 445–492). Berlin; Boston: De Gruyter Mouton.

Kubozono, Haruo. (2012). Japanese Accent. In S. Miyagawa (Ed.), *The Oxford Handbook of Japanese Linguistics*. https://doi.org/10.1093/oxfordhb/9780195307344.013.0007

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

Lanz, L. A. (2003). *Japanese Pitch-Accent: Cross-Linguistic Perceptions by Speakers of Stress- and Pitch-Accent Languages*. University of Hawaii.

Lee, C. Y., Tao, L., & Bond, Z. S. (2010). Identification of acoustically modified mandarin tones by non-native listeners. *Language and Speech*, *53*(2), 217–243. https://doi.org/10.1177/0023830909357160

MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, *2*(4), 369–390. https://doi.org/10.1017/S0142716400009796

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*. https://doi.org/10.3758/s13428-011-0168-7

Mennen, I. (2015). Beyond Segments: Towards a L2 Intonation Learning Theory. In E. Delais-Roussarie, M. Avanzi, & S. Herment (Eds.), *Prosody and Language in Contact : L2 Acquisition, Attrition and Languages in Multilingual Situations* (pp. 171–188). Springer. https://doi.org/10.1007/978-3-662-45168-7_9

Munro, Murray J., & Bohn, O.-S. (2007). The study of second language speech learning (pp. 3–11). https://doi.org/10.1075/lllt.17.06mun

Ortega-Llebaria, M., Gu, H., & Fan, J. (2013). English speakers' perception of Spanish lexical stress: Context-driven L2 stress perception. *Journal of Phonetics*, *41*(3–4), 186–197. https://doi.org/10.1016/j.wocn.2013.01.006

Ota, M. (2015). L1 phonology : phonological development. In H. Kubozono (Ed.), *Handbook of Japanese Phonetics and Phonology* (pp. 682–717). Berlin; Boston: De Gruyter Mouton.

Pallier, C., Bosch, L., & Sebastián-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition*, *64*(3), B9–B17. https://doi.org/10.1016/S0010-0277(97)00030-9

Pierrehumbert, J., & Beckman, M. (1988). *Japanese Tone Structure*. Cambridge (MA): The MIT Press.

Post, B., Stamatakis, E. A., Bohr, I., Nolan, F., & Cummins, C. (2015). Categories and gradience in intonation A functional Magnetic Resonance Imaging study. In J. Romero & M. Riera (Eds.), *The Phonetics/Phonology Interface: Sounds, representations, methodologies* (pp. 259–284). Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.335.13pos

Sakamoto, E. (2011). *Investigation of factors behind foreign accent in the l2 acquisition of japanese lexical pitch accent by adult english speakers*. *PQDT - UK & Ireland*.

Sakuma, K. (1929). *Nihon onseigaku [Japanese phonetics]*. Tokyo: Kyoubunsha.

Shport, I. A. (2011). *Cross-linguistic perception and learning of Japanese lexical prosody by*

English listeners. University of Oregon.

Shport, I. A. (2016). TRAINING ENGLISH LISTENERS TO IDENTIFY PITCH-ACCENT PATTERNS IN TOKYO JAPANESE. *Studies in Second Language Acquisition*, *38*(4), 739–769. https://doi.org/10.1017/S027226311500039X

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech*, *53*(2), 273–293. https://doi.org/10.1177/0023830909357156

Sugiyama, Y. (2012). *The Production and Perception of Japanese Pitch Accent*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Toda, T. (2001). Hatsuon shidō ga akusento no chikaku ni ataeru eikyō [The effect of pronunciation practice upon the perception of Japanese accents]. *Bulletin of Center for Japanese Language, Waseda University*, *14*, 67–88.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*(6), 3649–3658. https://doi.org/10.1121/1.428217

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

Wayland, R. P., & Guion, S. G. (2004). Training English and Chinese listeners to perceive Thai tones: A preliminary report. *Language Learning*, *54*(4), 681–712. https://doi.org/10.1111/j.1467-9922.2004.00283.x

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, *28*(4), 565–585. https://doi.org/10.1017/S0142716407070312

Wu, X., Munro, M. J., & Wang, Y. (2014). Tone assimilation by Mandarin and Thai listeners with and without L2 experience. *Journal of Phonetics*, *46*(1), 86–100. https://doi.org/10.1016/j.wocn.2014.06.005

Yip, M. (2002). *Tone*. Cambridge University Press.