

Available online at www.sciencedirect.com**ScienceDirect**

Lingua 256 (2021) 103069

Lingua

www.elsevier.com/locate/lingua

Acoustic salience in the evaluations of intelligibility and foreign accentedness of nonnative vowel production

Na Zhi^{a,*}, Aijun Li^b^a College of Foreign Languages, Capital Normal University in Beijing, China^b Institute of Linguistics, Chinese Academy of Social Sciences in Beijing, China

Received 23 April 2020; received in revised form 17 February 2021; accepted 18 February 2021

Available online 19 March 2021

Abstract

The study investigates the relative weighting of spectral and temporal acoustic cues in native speakers' (NS) perception of vowels produced by nonnative speakers (NNS). Acoustic measurements were taken of the spectral and temporal properties of vowels produced by Chinese L2 speakers of English, and by native speakers of American English and British English, who were recruited to pronounce monosyllabic words containing 18 monophthongs and diphthongs in English. In the perception experiment, 25 native English speakers (16 American and 9 British) provided evaluations of intelligibility and ratings of foreign accentedness of the non-native vowels. A correlation analysis was conducted between the NS perception results and the acoustic computation of the NNS production data. The results show that spectral properties of vowels are significantly related to both American and British speakers' evaluation of English vowels produced by Chinese L2 speakers, whereas the temporal information has no significant correlation to NS perceptual evaluations. The correlation between acoustic measurements and perceptual evaluations confirms that the most significant acoustic cue of English vowels is spectral rather than temporal information in perception.

© 2021 Elsevier B.V. All rights reserved.

Keywords: Acoustic; Weighting; Perception; Intelligibility; Accentedness

1. Introduction*1.1. Nonnative production and perception relation*

Numerous studies have attributed production problems of L2 speakers to the lack of accuracy in their nonnative perception. Two influential theoretical frameworks, Flege's *Speech Learning Model* (SLM, 1995) and Best's *Perceptual Assimilation Model* (PAM, 1995), have focused on the relation between nonnative perception and production. Although these models are both based on the premise that the L1 sound inventory shapes the learners' perception of L2, they differ in several aspects. SLM postulates that L1 and L2 sounds, related at a "position-sensitive allophonic level", exist in a common phonological space, and due to the "equivalence classification", the L2 sounds similar to the L1 ones are more difficult to acquire than "new" or dissimilar ones. SLM focuses on the individual sounds of L2, and discusses the likelihood

* Corresponding author.

of forming new phonetic categories for L2 sounds, which could be influenced by various factors, such as the degree of perceived phonetic difference between an L2 sound and the closest L1 sound(s), the age of learners when first exposed to L2, the length of residence (LOR) by learners in an L2-speaking country, and the quantity and quality of L2 input. While PAM discusses the diverse types of nonnative contrasts discriminated and assimilated by naïve listeners according to their experience with the L1 phonological system. PAM-L2 (Best and Tyler, 2007), an extension of PAM, predicts the discrimination difficulty of L2 contrasts, and distinguishes different assimilation patterns. The model assumes that nonnative contrasts could be perceptually assimilated to the native phonological categories of learners in several different types: (a) *Two-Category (TC) assimilation*, in which the discrimination of two contrasting phones ranges from very good to excellent; (b) *Single-Category (SC) assimilation*, in which poor discrimination of contrasts is predicted if both are assimilated to the same native phonological category, and are equally deviant or acceptable from the native “ideal”; (c) *Category-Goodness (CG) assimilation*, in which pairwise discrimination is expected to be moderate, as one is assimilated as a good version and the other as a poor version of the same category; (d) *Uncategorized-Categorized (UC) assimilation*, in which one nonnative phone is uncategorized, and the other is categorized, with the sound pair discriminated very well; (e) *Uncategorized-Uncategorized (UU) assimilation*, for which nonnative phones are discriminated from poorly to moderately well; and (f) *Non-Assimilable (NA)*, in which phones are not perceived as speech sounds, and the discrimination might range from good to excellent.

Though both the above models could predict the various degrees of difficulty in L2 perception and production, a growing number of studies discover a dissociation between perception and production, postulating that the relation between the two could be complex. It is argued that L2 perception accuracy might not ensure the correct L2 production (Kartushina and Frauenfelder, 2014), due to the fact that L2 phonological perception (“input”) and production (“output”) have different underlying representations (Edwards, 1995). Rallo Fabra and Romero (2012) report the lack of an overall significant correlation between the data of perception and production in their study of different groups of Catalan L2 speakers with varied levels of English proficiency. They conclude that the non-alignment of perception and production relation could be accounted for by various factors, such as the target sounds elicited for production and perception experiments being embedded in different contexts; the tasks of perception and production not being equivalent in difficulty level for L2 speakers; participants in various studies might not be comparable owing to their differences in exposure to naturalistic L2 settings. Major (2014:55) holds that the relation between L2 perception and production is not clear-cut or predictable. He has proposed four possibilities in L2 phonology,

- (a) if perceived target = NS target, production = NS production,
- (b) if perceived target = NS target, production ≠ NS production,
- (c) if perceived target ≠ NS target, production ≠ NS production,
- (d) if perceived target ≠ NS target, production = NS production.

He also suggests that the interlanguage of L2 speakers is a combined product, which consists of certain parts of L1, parts of L2, and linguistic universals (which are neither parts of L1 or L2). We believe that the sensitivity of speakers to nonnative speech is related to their native sound patterns and linguistic experience. The difference in cue weighting for phonological contrasts between L1 and L2 deserve our further attention when exploring nonnative speech production and perception.

1.2. Acoustic cue weighting across languages

The phonological contrasts of a language are signaled by various acoustic information, such as spectral properties (static or dynamic), duration and intensity. The relative weighting of those acoustic cues can be different cross-linguistically, with some cues being primary and the others secondary. The diverse cue weighting could have a differential effect on the overall perception of the language (Escudero, 2000). In cross-language studies such as Bradlow (1995), Fox et al. (1995), and Gottfried and Beddor (1988), it is found that some acoustic cues are salient in the sound distinction in one language while being optional or not employed at all in another. For instance, native English speakers use both spectral cues and duration for vowel contrasts, whereas native French speakers do not often employ duration in categorizing vowels. It is claimed in Hillenbrand et al. (2000) that native speakers of American English (AE) could correctly perceive a great majority of vowels in an identification task with vowels altered in duration. The authors postulated that whether some acoustic cues, such as vowel length, should be considered as a contrastive or redundant phonological feature, depends on the degree of spectral difference between the contrasting phonemes. For example, they find differences in spectral properties in the tense and lax vowel pairs of /i/-ɪ/ and /u/-ʊ/. Duration, as a consequence, plays a minor role in native AE speakers' perception of the contrastive vowel identity in such a contrast. They are classified as “non-duration-sensitive” pairs in perception, despite the fact that in production a consistent durational difference is found between these

contrasting vowels. Other pairs such as /ɑ/-ɔ/-ʌ/ and /æ/-e/ have a greater overlap of spectral features, and as a result, duration is given more weight by native AE speakers in the recognition of contrasts.

Previous studies on acoustic cue weighting in L1 have shed light on the L2 learning research. The relative preference of cues for phonemic distinction is proved to be different between native and nonnative speakers. L2 learners are often easily identified by native speakers as making errors or speaking with a foreign accent in their production of either utterances or single words (Flege, 1984). Various factors might lead to inaccurate and accented production, but one important factor among others is that learners seem to have different sensitivity to the acoustic cues of nonnative sounds (Escudero and Boersma, 2004). Cebrian (2006) shows that Catalan learners of English exhibit overreliance on the temporal information in distinguishing English lax-tense vowel contrasts, despite of the learners' varied experience with English and the fact that duration is not used in signaling L1 phonological contrasts. Kondaurova and Francis (2010) report similar results that native English listeners employ spectral features, while inexperienced Spanish listeners rely exclusively on vowel duration in identifying contrasting English tense and lax vowels. They suggest that L2 phonetic learning might be more effective with "cue-specific" training in sound contrast distinction. The different acoustic reliance on vowel identification is also discussed in Yang (2011). His analysis of the spectral features of English produced by Mandarin speakers shows their failure in using differences in formant frequencies for the tense-lax vowel distinction. Escudero et al. (2009) reports that naïve German listeners weight spectrum more heavily than duration in perceiving Dutch vowels, in the same manner as native Dutch speakers. However, the German listeners are still less accurate in discriminating the /a:/-/a/ contrast in Dutch, suggesting that the auditory dimensions integrated by nonnative listeners might not resemble those of native listeners.

1.3. L2 intelligibility and accentedness

Munro and Derwing (1995a,b, 2011) study the difference of cue weighting employed by native speakers and nonnative speakers from a pedagogical perspective, and focus on the relationship between foreign accent, production accuracy and intelligibility. They observe that L2 speech with a strong accent might not preclude full intelligibility. Therefore, they point out the need for focusing on intelligibility for successful communication, rather than striving to achieve the goal of native-like pronunciation, which may be impossible for late L2 learners. Munro (1998) proposes the concept of "error gravity", and notes that some L2 pronunciation errors are more serious than others, since they could affect speech intelligibility. Cook (2002) and Hansen (2008) propose that L2 speakers should be viewed as language "users" rather than language "learners". The "deviations" of L2 from the target language might not necessarily be errors, but normal language variations. Hayes-Harb and Watzinger-Tharp (2012) insists that the assessment of L2 production should be centered on its intelligibility instead of the negative comparison between L2 performance and the standard target-language model.

1.4. Research goal of the study

Previous studies suggest that successful acquisition of L2 sounds needs to take into account the important cues that native speakers employ in perception and production. In this paper, we seek to investigate the relative weighting of spectral and temporal acoustic cues used by native English speakers in the perception of vowels produced by nonnative speakers. Acoustic measurements will be taken of the spectral and temporal properties of vowels produced by Chinese L2 speakers of English, and by native speakers of American and British English. In the perception experiment, native English speakers will provide evaluations of vowel intelligibility and ratings of accentedness on the recorded data. A correlation analysis will then be conducted between the native speakers' perception results and the acoustic measurements of the production data. We will also explore variations in Chinese L2 speakers' vowel production in an attempt to offer pedagogical recommendations.

Specially, we look to addressing the following three research questions:

1. Are there differences in the perception of L2 vowels by native speakers of American and British English, as assessed in the intelligibility evaluation and accentedness rating tasks?
2. Is there any correlation between the acoustic measurements of L2 production and perceptual evaluations by native speakers?
3. Which acoustic cues significantly relate to the perception of L2 vowels by native English speakers?

Furthermore, many L2 speakers who study English in a non-English-speaking country usually do not have a fixed accent as model. Their pronunciation is greatly influenced by that of their native or nonnative English teachers in class, who may speak with different English accents. L2 speakers who have constrained naturalistic language settings often have to imitate the pronunciation of native English speakers from various textbook recordings, popular songs, or films from

different English-speaking countries. As native speakers of American English and British English mainly differ in their pronunciation of vowels, most L2 speakers do not present a consistent English accent. Some of their English vowels may sound like American pronunciation, while others are more similar to British vowel patterns. However, most previous studies do not take the issue into account, and employ only one group of native English speakers, either American or British to assess the production of vowels by L2 speakers. In the current study we will have two groups of native English speakers to evaluate the L2 speakers' pronunciation.

2. Perception experiment

2.1. Stimuli and recording

The stimuli include 18 English vowels, specifically, /ɑ:/, /i:/, /ɪ/, /æ/, /e/, /ɔ:/, /ɒ/, /u:/, /ʊ/, /ɜ:/, /ʌ/, /eɪ/, /aɪ/, /ɔɪ/, /aʊ/, /oʊ/, /ju:/, /jɔ:/. Each vowel is embedded in a real monosyllabic word, with a stop or a fricative in the syllabic onset and coda, such as but. The 18 words are given in Table 1.

Table 2 includes a list of Mandarin Chinese vowels embedded in real words for comparison purposes. Mandarin has fewer monophthongs and diphthongs than English, and furthermore, it does not contrast tense and lax vowels such as /i:/-/ɪ/, /u:/-/ʊ/, /æ/-/e/, which are distinguished by both spectral and temporal properties in English.

We recorded 18 English vowels pronounced individually by 8 native Mandarin speakers, and 16 native English speakers (8 American and 8 British) in a sound-proof booth. The speakers were on average 26 years old. They were seated and asked to read the words presented on the screen of a laptop computer in front of them. Each wore a headset with a microphone, and read each word three times in a normal voice and at a normal speech rate. The Mandarin speakers were born and raised in Beijing, and have been learning English in China since primary school. None of them had any experience of living in an English-speaking country. In their English classes, they were exposed to both American and British English accents, since their English instructors were from different English-speaking countries, and the listening materials were recorded by speakers of different accents. The American speakers in the study were from the Midwest of USA, and the British speakers were from Southern England. See Table 3 for a summary. None of them reported speaking or hearing problems.

2.2. Participants

25 native English speakers (16 American and 9 British) took part in the perception experiment. See Table 4. They were exchange students studying Chinese for a few months at a university in Beijing. The average age was 28 years old. All reported normal hearing.

Table 1
English word list.

	Vowel type	Vowels	Words
1	monophthong	/ɑ:/	pa <u>s</u> s
2	monophthong	/i:/	bea <u>t</u>
3	monophthong	/ɪ/	bi <u>t</u>
4	monophthong	/æ/	ba <u>ck</u>
5	monophthong	/e/	be <u>t</u>
6	monophthong	/ɔ:/	bo <u>u</u> ght
7	monophthong	/ɒ/	po <u>t</u>
8	monophthong	/u:/	ju <u>i</u> ce
9	monophthong	/ʊ/	bo <u>o</u> k
10	monophthong	/ɜ:/	bi <u>r</u> d
11	monophthong	/ʌ/	bu <u>t</u>
12	diphthong	/eɪ/	pa <u>i</u> d
13	diphthong	/aɪ/	bi <u>t</u> e
14	diphthong	/ɔɪ/	cho <u>i</u> ce
15	diphthong	/aʊ/	hou <u>s</u> e
16	diphthong	/oʊ/	bo <u>a</u> t
17	diphthong	/ju:/	du <u>k</u> e
18	diphthong	/jɔ:/	yo <u>r</u> k

Table 2
Mandarin Chinese vowels.

	Vowel type	Vowels	Words (in Chinese Pinyin and Characters)
1	monophthong	/a/	ba 八
2	monophthong	/ə-/	er 而
3	monophthong	/o/	o 播
4	monophthong	/ɤ/	de 德
5	monophthong	/i/	bi 逼
6	monophthong	/u/	du 督
7	monophthong	/y/	ju 拘
8	diphthong	/ei/	bei 杯
9	diphthong	/ai/	bai 掰
10	diphthong	/ao/	bao 包
11	diphthong	/ou/	dou 都
12	diphthong	/iu/	diu 丢

Table 3
Speaker profiles.

	Chinese L2 speakers (Beijing)	British (Southern England)	American (the Midwest of USA)
Female	4	4	4
Male	4	4	4
Total (N)	8	8	8

Table 4
Participants.

	British (Southern England)	American (the Midwest of USA)
Female	4	6
Male	5	10
Total (N)	9	16

2.3. The procedure

Each participant listened to the stimuli played from a laptop computer in front of them via a headphone in a sound-proof booth. The whole experiment, administered with Eprime 2.0, lasted approximately one and a half hours, and consisted of pre-experiment practices, the intelligibility evaluation task, a 25-min break and the accentedness rating task. The same set of stimuli of NNS vowel production was used for the two tasks.

In order to make sure we had the full attention of participants in the perception tasks, we controlled the time of the whole experiment, and employed only a subset of the recorded stimuli. We selected a total of 108 English monosyllabic words (=18 words*6 speakers), 72 of which were produced by Chinese L2 speakers, 18 by American speakers (each word from a randomly selected American speaker), and 18 by British speakers (each word from a randomly selected British speaker). The selected stimuli had no obvious mispronunciation in consonants.

Before each task started, 4 practice stimuli were provided to participants in order to familiarize them with the Eprime procedure on the computer. For the intelligibility evaluation, participants were asked to identify the word they heard in a four-alternative forced-choice identification task. For example, when the stimulus was “back”, four choice words would appear on the screen “a. beck”, “b. bike”, “c. back”, “d. buck”. The choices appeared on the screen 500 ms after each stimulus was played, and participants would click on the word they identified. In the second task, native English listeners were asked to assess the accentedness of L2 speakers’ vowel production and rate the perceived accent (based on goodness of pronunciation) on a 5-point Likert scale with a number from 1 to 5, where 1 = heavily accented (very poor pronunciation) and 5 = no perceived foreign accent (native-like pronunciation).

Table 5
NS listeners' evaluations of the NS vowel production (in percentage).

	Vowels in words	Listeners (Ame) -Speakers (Ame)	Listeners (Bri) -Speakers (Bri)	Listeners (Ame) -Speakers (Bri)	Listeners (Bri) -Speakers (Ame)
1	pass	100%	100%	63%	100%
2	back	100%	100%	100%	100%
3	bought	100%	100%	88%	67%
4	bit	100%	100%	100%	100%
5	book	100%	100%	94%	100%
6	bird	100%	100%	100%	89%
7	juice	100%	89%	94%	100%
8	pot	94%	100%	100%	22%
9	bet	94%	78%	94%	89%
10	beat	94%	100%	100%	100%
11	but	88%	56%	25%	33%
12	paid	100%	100%	100%	100%
13	bite	100%	100%	100%	100%
14	choice	100%	100%	100%	100%
15	house	100%	100%	100%	100%
16	boat	100%	100%	100%	100%
17	duke	100%	100%	100%	89%
18	york	100%	100%	100%	100%

Ame (Americans); Bri (British).

3. Results of the perception experiment

3.1. Intelligibility task

3.1.1. NS listeners' evaluations of NS pronunciation

To ensure that listeners understood the task and were reliable in their response, we first check the NS listeners' evaluations of the NS vowel production in the study.¹

Table 5 presents results of NS listeners' perception of the vowels produced by NS speakers, in four possible combinations.

Both American and British listeners are highly consistent in their evaluations of the vowels produced by speakers of the same accent. The percentage of vowel intelligibility ranges from 78% to 100%. However, the British production of the lax vowel /ʌ/ in 'but' seems an exception, with the intelligibility rate of only 56% among British listeners.

The NS perception of the vowels produced by speakers of a different accent shows lower intelligibility, especially in 'bought', 'pot' and 'but'. American listeners' intelligibility percentage is much lower in perceiving British vowels, such as 'pass' (63%) and 'but' (25%), while British listeners' perception of American vowels also receives lower intelligibility in 'bought' (67%), 'pot' (22%) and 'but' (33%). The results point to the articulatory difference between American English and British English in *bought*, *pot* and *pass*, which will be further discussed in §4.1. Therefore, it seems necessary to analyze results from the two groups of NS listeners separately in their assessment of the NNS vowel production.

In addition, due to the participants' strong inconsistency in evaluating the lax vowel in 'but', we have decided to eliminate their perceptual evaluation on the same word produced by Chinese L2 speakers from our analysis.

3.1.2. NS listeners' evaluations of NNS vowel production

Table 6 presents NS listeners' perception of the vowel pronunciation by four Chinese L2 speakers. The difference in intelligibility evaluations between American listeners and British listeners was computed with a paired samples *t*-test (Table 7). The statistical results show that the intelligibility rate for British listeners ($M = 76.6$, $SE = 3.8$) is on average higher than that for American listeners ($M = 72.2$, $SE = 3.5$), and the difference between the two groups is marginally significant, $t(67) = -1.806$, $p = 0.075$ (<0.1).

To interpret the results more clearly, we employed a hierarchical-clustering analysis and grouped the data into four categories.

¹ The listeners and speakers in the study were different groups of people.

Table 6
NS listeners' evaluations of the NNS vowel production (in percentage).

	Vowels in words	Listeners (Ame) –				Listeners (Bri) –			
		L2 speakers (C, S, Z, L)				L2 speakers (C, S, Z, L)			
		C	S	Z	L	C	S	Z	L
1	pass	100%	75%	44%	94%	100%	100%	100%	100%
2	back	69%	50%	31%	94%	78%	67%	22%	100%
3	bought	6%	31%	94%	0%	0%	44%	33%	11%
4	bit	75%	75%	56%	6%	33%	89%	78%	11%
5	book	81%	81%	50%	69%	100%	100%	100%	100%
6	bird	100%	25%	100%	94%	100%	67%	100%	100%
7	juice	94%	81%	88%	81%	100%	100%	89%	100%
8	pot	63%	44%	94%	81%	22%	22%	78%	44%
9	bet	6%	50%	81%	44%	22%	56%	100%	56%
10	beat	94%	31%	100%	88%	89%	11%	100%	100%
11	paid	100%	75%	100%	100%	100%	44%	100%	100%
12	bite	100%	38%	100%	100%	100%	22%	89%	100%
13	choice	94%	63%	100%	94%	100%	67%	100%	100%
14	house	100%	81%	94%	100%	100%	89%	100%	100%
15	boat	100%	81%	94%	94%	100%	100%	78%	89%
16	duke	94%	63%	69%	88%	100%	89%	100%	100%
17	york	38%	13%	56%	69%	22%	22%	89%	89%

Table 7
Paired samples statistics.

		Mean (E)	N	Std. Deviation	Std. Error Mean
Pair 1	Intelligibility (Ame listeners)	72.243	68	28.516	3.458
	Intelligibility (Bri listeners)	76.634	68	31.727	3.847

The hierarchical clustering of the data is given in Table 8. Category (rank) 1 has the lowest values of the intelligibility rate, ranging from 0 to 12.50% ($M = 6.25\%$) for American listeners, and from 0 to 33.33% ($M = 19.66\%$) for British listeners, while Category 4 has the highest values of the intelligibility rate, from 87.5% to 100% ($M = 96.04\%$) for American listeners, and from 88.89% to 100% ($M = 97.67\%$) for British listeners.

According to NS listeners' perception, the production of lax vowels, tense vowels and diphthongs by Chinese L2 speakers are different in intelligibility. As shown in Fig. 1, the intelligibility ratings of diphthongs are higher than those of tense vowels and lax vowels for both American and British listeners. The intelligibility ratings of lax vowels are the lowest, suggesting that the production of the lax vowels by L2 speakers are not readily perceived by NS listeners. In the following section, we will present NS listeners' accentedness ratings of the three types of vowels produced by the Chinese L2 speakers.

Table 8
Hierarchical clustering of the intelligibility results.

Categories	Intelligibility results (Ame listeners)					Intelligibility results (Bri listeners)				
	Mean (E)	Std Dev	Min (M)	Max (X)	N	Mean (E)	Std Dev	Min (M)	Max (X)	N
1	6.25%	4.42	0	12.50%	5	19.66%	9.25	0	33.33%	14
2	41.96%	9.94	25%	56.25%	14	48.89%	6.09	44.44%	55.56%	4
3	74.34%	7.19	62.50%	81.25%	19	73.02%	5.94	66.67%	77.78%	7
4	96.04%	4.18	87.50%	100%	30	97.67%	4.57	88.89%	100%	43

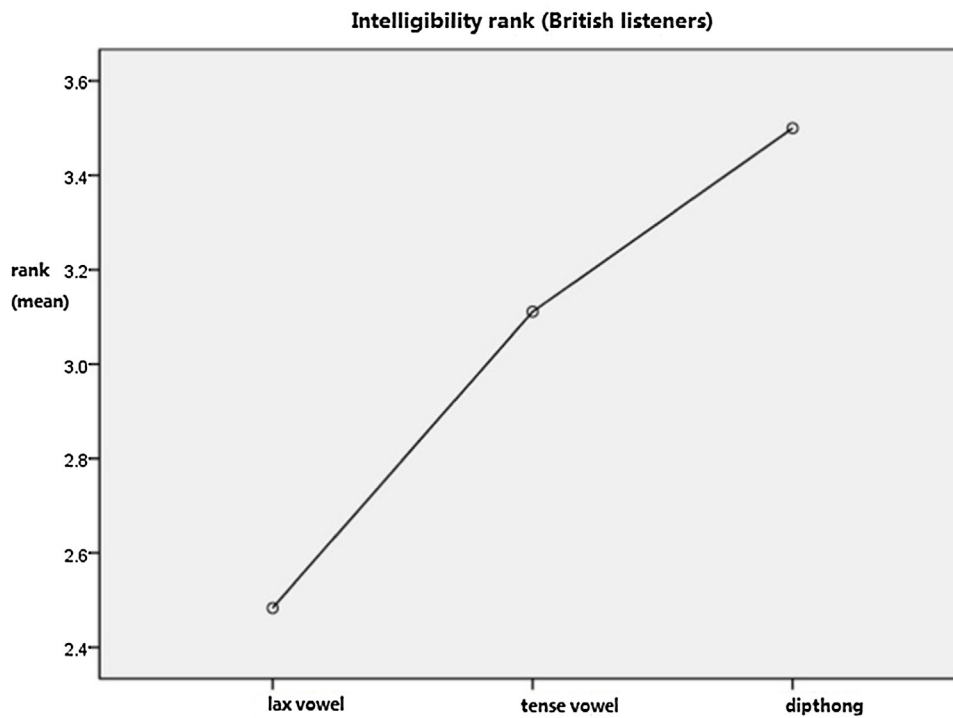
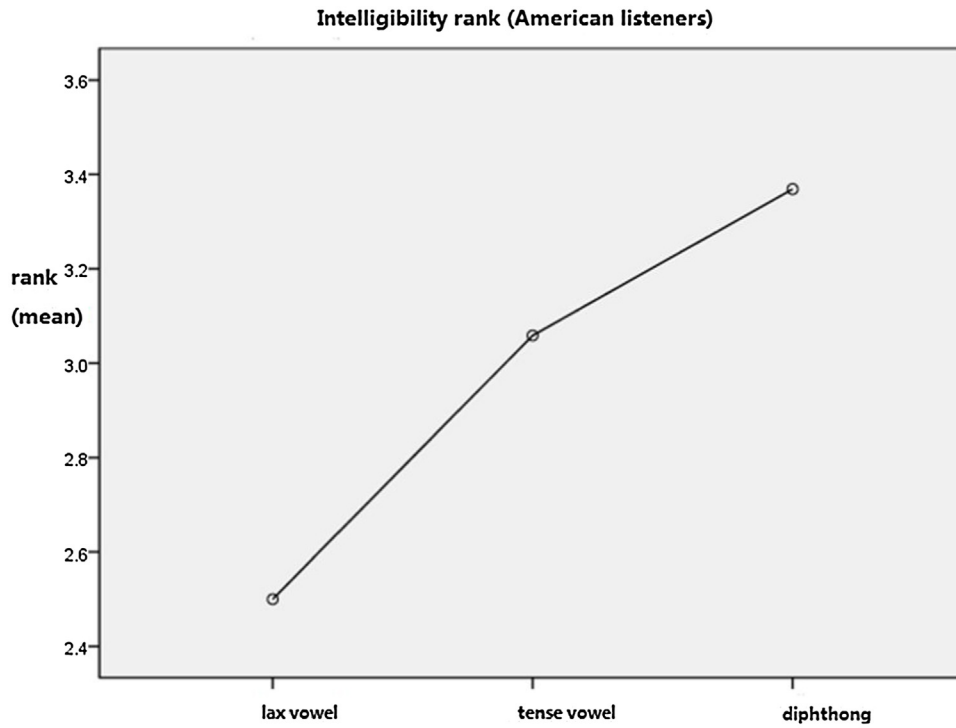


Fig. 1. The intelligibility rankings of different vowel types.

Table 9
Results of accentedness ratings.

	Vowels in words	Listeners (Ame) –				Listeners (Bri) –			
		L2 speakers (C, S, Z, L)				L2 speakers (C, S, Z, L)			
		C	S	Z	L	C	S	Z	L
1	pass	4.8	3.2	3.4	3.1	4.4	4.3	4.6	3.9
2	back	2.8	2.6	3.2	3.9	2.9	2.4	2.6	3.6
3	bought	1.7	1.7	3.7	1.8	1.2	2.3	3.7	1.4
4	bɪt	3.8	3.6	1.9	1.4	4	3.6	2.9	1.8
5	book	2.9	3.1	2.5	2.3	4	3.7	3.3	2.9
6	bird	3.6	1.9	4.3	3.1	4	2.3	4.3	3.2
7	juice	3.8	3.2	4	3.6	3.9	3.7	3.9	3.8
8	pot	3.3	2.8	3.8	3.3	2	2.8	3.9	2.4
9	bet	1.8	2.2	3.9	2.4	2.4	2.3	4.1	2.6
10	beat	3.9	2.4	4.3	3.9	3.8	2.4	4.3	3.6
11	paid	3.6	3.3	3.8	4	3.1	2.7	3.7	4.4
12	bite	3.4	1.9	4	3.5	3.9	2.2	4.4	3.6
13	choice	3.5	3.2	4.3	3.6	3.3	2.7	4.3	3.8
14	house	4.4	3.1	3.9	3.7	4.7	3.4	3.9	4.4
15	boat	3.6	2.8	3.7	3.9	3.7	2.8	3.8	3.8
16	duke	3.7	2.8	3.5	3.1	4.2	3.3	4.1	3.1
17	york	2.3	1.8	3.3	3.3	2.1	1.6	4.1	3.8

Table 10
Paired samples statistics.

		Mean (E)	N	Std Dev	Std. Error Mean
Pair 1	Accentedness ratings (Ame listeners)	3.198	68	0.777	0.094
	Accentedness ratings (Bri listeners)	3.355	68	0.843	0.102

3.2. Accentedness rating task

3.2.1. NS listeners' accentedness ratings of the NNS vowels

Table 9 shows the accentedness ratings of the NNS vowels based on the perceived goodness of production by NS listeners on a 5-point Likert scale, where 1 = heavily accented (very poor pronunciation) and 5 = no perceived foreign accent (native-like pronunciation).

The comparison of rating results between American listeners and British listeners was computed by means of a paired samples *t*-test. The statistics in Table 10 show that accentedness ratings by the British group ($M = 3.355$, $SE = 0.102$) are on average slightly higher than that by the American group ($M = 3.198$, $SE = 0.094$). The difference between the two groups is significant, $t(67) = -2.739$, $p = 0.008$ (** <0.01).

We also used a hierarchical-clustering analysis to group the rating scores into 4 categories in Table 11. Category (rank) 1 includes the lowest accentedness ratings of production, ranging from 1.38 to 1.94 ($M = 1.76$) for American listeners, and from 1.22 to 1.78 ($M = 1.5$) for British listeners; while Category 4 has the highest range of 3.5–4.81 ($M = 3.85$) for American listeners, and the range of 3.78–4.67 ($M = 4.1$) for British listeners.

The rating results are presented according to the different vowel types in Fig. 2. From high to low, the accentedness ratings follow the order of diphthongs > tense vowels > lax vowels for both the American and the British groups, which

Table 11
Hierarchical clustering of accentedness ratings.

Categories	Accentedness ratings (Ame listeners)					Accentedness ratings (Bri listeners)				
	Mean (E)	Std Dev	Min (M)	Max (X)	N	Mean (E)	Std Dev	Min (M)	Max (X)	N
1	1.76	0.17	1.38	1.94	9	1.5	0.23	1.22	1.78	4
2	2.56	0.24	2.19	2.94	12	2.51	0.26	2	2.89	19
3	3.22	0.1	3.13	3.44	16	3.47	0.2	3.11	3.67	16
4	3.85	0.3	3.5	4.81	31	4.1	0.28	3.78	4.67	29

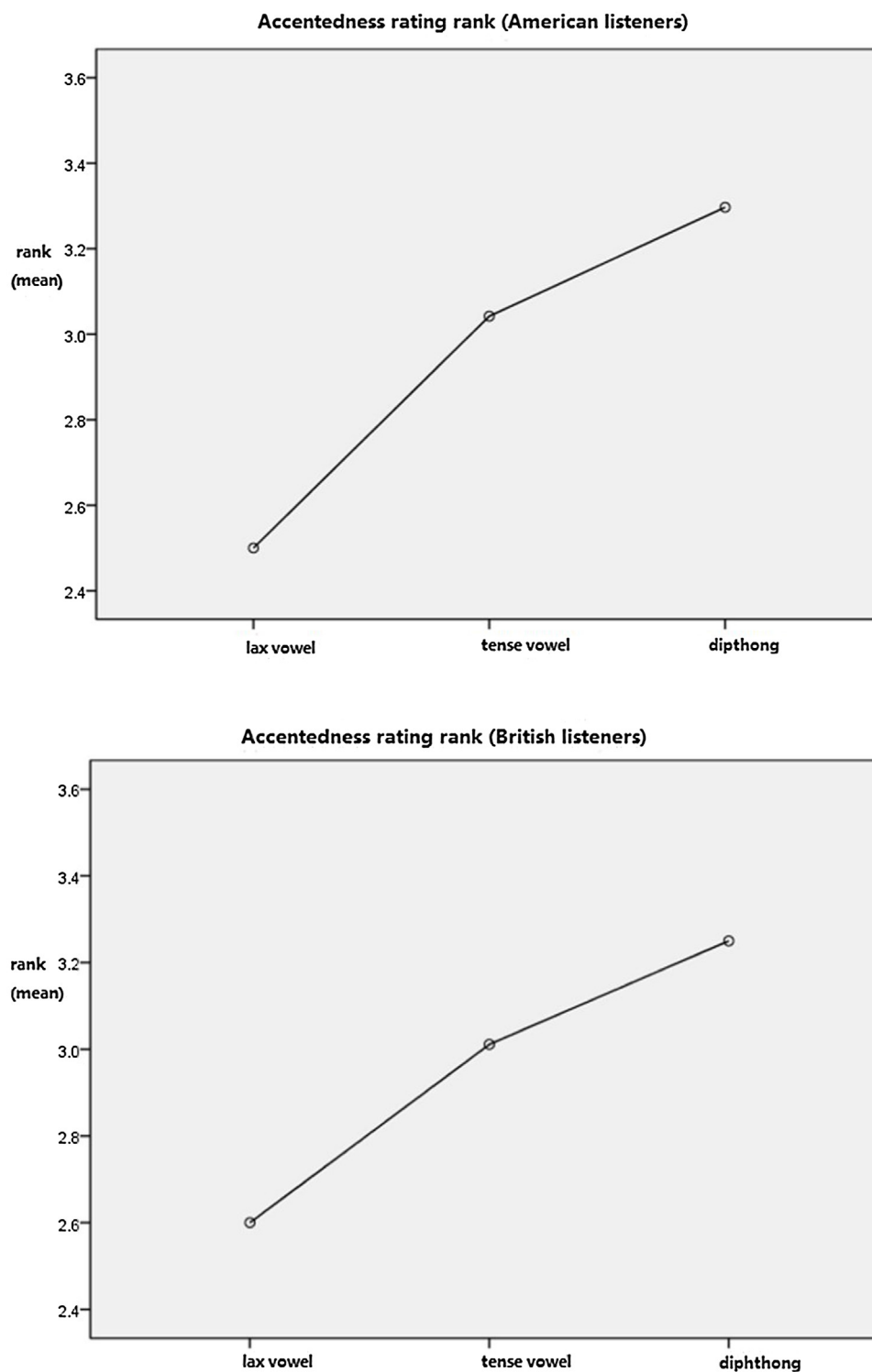


Fig. 2. Accentedness ratings of different vowel types.

turns out to match the intelligibility results. In both tasks, the production of lax vowels such as 'bet', 'bit' and 'pot' seems to present a huge challenge to Chinese L2 speakers, as it has received the lowest intelligibility percentage and the poorest accentedness rating, whereas the production of diphthongs appears to be least problematic to the NS listeners.

The difficulty in the lax vowel distinction could be attributed to the lack of the lax-tense vowel contrast in the native Chinese phonology. PAM-L2 proposes that learners' discrimination of nonnative contrasts could be explicitly predicted depending on how the contrasting sounds are assimilated. The model also posits that listeners would perceptually assimilate nonnative phones to native categories whenever possible, based on their detection of commonalities in articulatory aspects (Best et al., 2001). It is possible that the production of lax vowels in the L2 speech is strongly influenced by the native L1 phonology and the poor perceptual discrimination of the lax-tense vowel contrast, which we intend to further explore in future studies.

3.3. Summary

We conducted a perception experiment that consisted of two tasks with 16 American and 9 British listeners. In the first task they evaluated the intelligibility of vowels uttered in words by Chinese L2 speakers and native English speakers with American and British accents. In the second task they rated the foreign accentedness of the pronunciation of L2 speakers on a 5-point Likert scale. We found that American and British listeners show lower intelligibility in their perception of each other's vowels in words such as 'bought', 'pass' and 'pot', due to the articulatory difference of the vowels in the two accents. As Chinese L2 speakers are exposed to both American and British accents in their English classes, it is necessary to have two groups of native English speakers to evaluate the L2 production data separately. The comparison of the perception of the L2 speakers' pronunciation by the two native groups was carried out by means of a paired samples *t*-test. The results show that the American and British listeners behaved significantly different in their accentedness ratings of the NNS vowels, while their difference in evaluating vowel intelligibility of the L2 speakers is marginally significant. We also find the evaluation results by the native speakers varied according to different vowel types: diphthongs are most acceptable to native speakers with the highest intelligibility and accentedness rating, whereas the production of lax vowels by L2 speakers is most problematic, as shown by their lowest intelligibility and accentedness rating.

4. Acoustic measurement

4.1. Formant frequencies

Each vowel in the production data was annotated using Praat, with F1 and F2 values extracted at 10 equidistant temporal locations (10%, 20%, 30%, 40%...100%) over the duration of the vowel. To exclude the influence of formant transitions from adjacent consonants in the CVC structure, average F1 and F2 values were computed from 20% to 80% of the duration of each monophthong. For diphthongs, due to dynamic spectral change, formant frequency values at 20% and 80% of the vowel period were extracted, corresponding to the diphthong onset and offset respectively.

The formant values were then normalized using Lobanov's z-score transformation method (1971), which is a vowel-extrinsic normalization approach to eliminate physiological difference across speakers, while preserving the phonemic variation of an individual speaker (Adank et al., 2004; Thomas and Tyler, 2007). The transformation of formant values into z-score values is done using the following formula,

$$F_n[V]^N = \frac{F_n[V] - \text{MEAN}_n}{S_n}$$

where V represents the vowel, MEAN_n and S_n are the mean and standard deviation of formant *n* of the speaker, and $F_n[V]^N$ is the resulting normalized value. The normalized data of monophthongs and diphthongs produced by L2 speakers and native English speakers are presented in Tables 12, 13, 14 and 15.

4.1.1. American vowels vs. British vowels

The vowel plots of American speakers (upper-left of Fig. 3) and British speakers (upper-right) were drawn based on the vowel data of 8 native speakers from each accent (Table 14), with the vowel space established by connecting points with lines between the peripheral vowels. The vowel difference between American English and British English could easily be observed when the two plots are overlapped in one graph. American and British speakers produce these vowels quite differently, such as 'pass', 'bought' and 'pot', accounting for the lower intelligibility percentage in each other's perception.

Table 12
Normalized F1 and F2 values of monophthongs by Chinese L2 speakers.

monophthong	Chinese L2 speakers			
	nF1	Std Dev (nF1)	nF2	Std Dev (nF2)
beat	-1.382	0.217	1.846	0.242
juice	-1.217	0.117	-0.672	0.229
bit	-0.990	0.427	1.523	0.454
book	-1.272	0.255	-1.061	0.301
bird	0.285	0.503	-0.094	0.313
bet	0.958	0.454	0.508	0.086
bought	0.274	0.317	-0.786	0.176
pot	0.645	0.379	-0.825	0.134
back	1.039	0.182	0.481	0.080
pass	1.483	0.150	-0.241	0.246

nF1 and nF2 refer to the mean normalized F1 and F2 values respectively.

Table 13
Normalized F1 and F2 values of diphthongs by Chinese L2 speakers.

Diphthong	Chinese L2 speakers			
	nF1	nF2	nF1 glide	nF2 glide
york	-0.510	0.203	0.544	-0.698
duke	-1.274	1.164	-0.930	-0.463
choice	0.077	-1.040	-0.202	0.683
paid	-0.135	0.983	-0.977	1.639
boat	0.243	-0.949	-0.613	-1.364
bite	0.999	-0.167	0.612	0.924
house	0.961	-0.604	-0.031	-0.943

nF1 and nF2 refer to the mean normalized F1 and F2 values respectively.

Table 14
Normalized F1 and F2 values of monophthongs by American and British speakers.

	American speakers				British Speakers			
	nF1	Std Dev (nF1)	nF2	Std Dev (nF2)	nF1	Std Dev (nF1)	nF2	Std Dev (nF2)
beat	-1.484	0.273	1.856	0.088	-1.425	0.109	1.663	0.119
juice	-1.128	0.099	-0.341	0.285	-1.233	0.253	0.081	0.340
bit	-0.386	0.154	0.812	0.169	-0.574	0.098	1.139	0.115
book	-0.280	0.300	-0.608	0.253	-0.515	0.191	-0.467	0.264
bird	-0.138	0.318	-0.427	0.150	0.453	0.383	-0.119	0.200
bet	0.570	0.240	0.418	0.151	0.495	0.217	0.600	0.130
bought	0.937	0.224	-1.107	0.187	-0.643	0.196	-1.639	0.092
pot	1.202	0.219	-0.952	0.156	0.358	0.162	-1.176	0.155
back	1.369	0.234	0.250	0.194	1.756	0.111	-0.119	0.255
pass	1.412	0.241	0.118	0.145	1.070	0.152	-0.872	0.125

4.1.2. Mahalanobis distance

To further illustrate the vowel difference in American English and British English, the Mahalanobis distance (DS) was computed with normalized F1 and F2 values of vowels. The Mahalanobis DS outperforms the Euclidean distance in estimating the distance between a test sample and reference samples, because the former measures the distance between a point and a distribution, and takes into account the natural variability in speech production, such as the variance of each variable as well as the covariance between variables (Kartushina and Frauenfelder, 2014; Kartushina et al.,

Table 15
Normalized F1 and F2 values of diphthongs by American and British speakers.

Diphthong	American speakers				British speakers			
	nF1	nF2	nF1 glide	nF2 glide	nF1	nF2	nF1 glide	nF2 glide
york	-1.290	0.366	-0.149	-0.828	-1.354	1.264	-0.243	-0.732
duke	-1.154	1.187	-1.293	-0.462	-1.257	1.376	-1.332	0.799
choice	-0.517	-1.760	-0.614	0.900	-0.413	-1.303	-0.428	1.028
paid	-0.419	1.157	-1.322	1.797	0.520	0.534	-1.328	1.626
boat	-0.037	-1.145	-0.581	-1.495	0.510	-0.515	-0.874	-0.156
bite	0.803	-0.127	-0.660	1.249	1.055	-0.657	-0.104	0.797
house	1.228	-0.170	0.245	-1.139	1.529	-0.183	0.111	-0.932

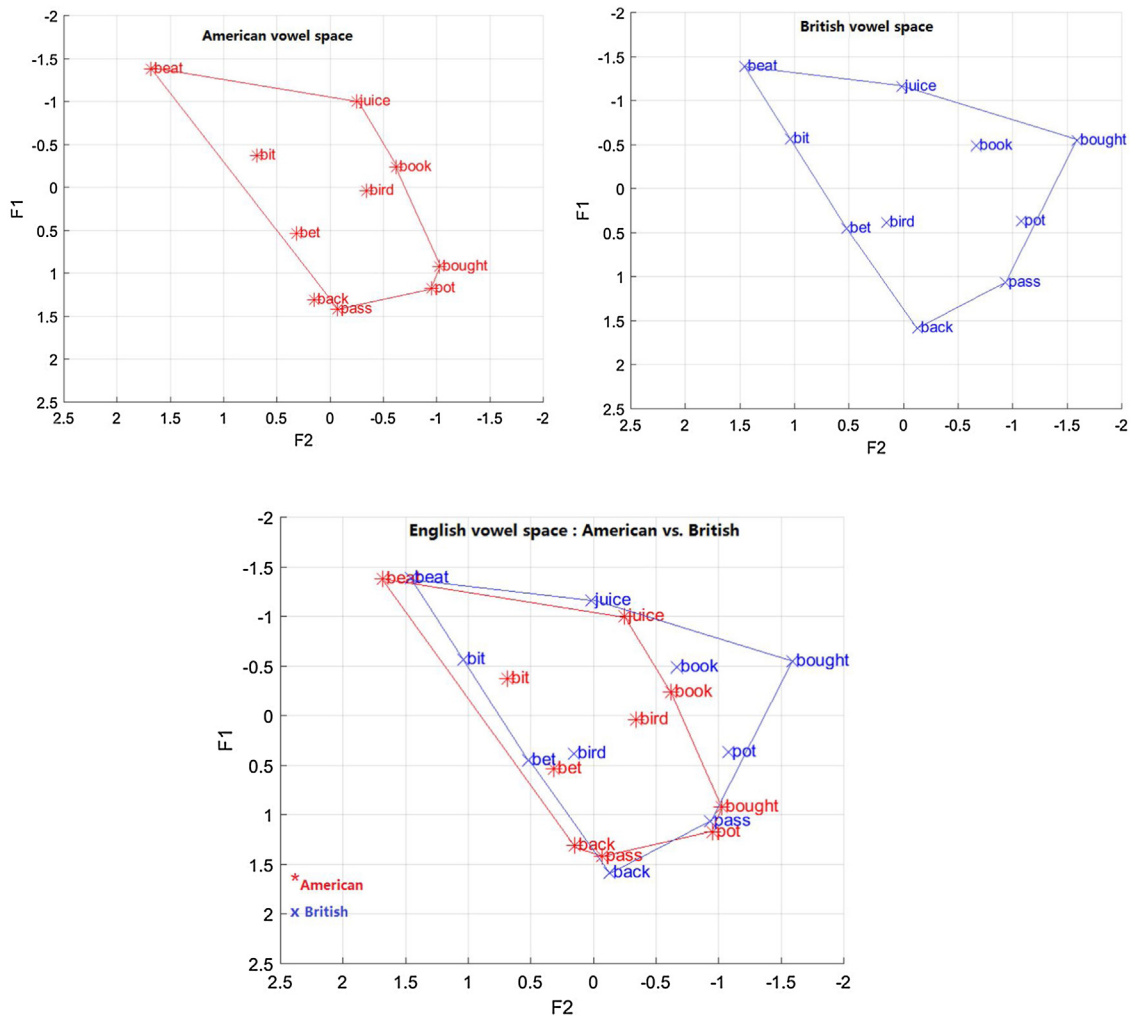


Fig. 3. American English vowel space (upper-left), British English vowel space (upper-right) and the overlapped vowel space (lower).

Table 16
 Normalized duration.

Vowels in words	C	S	Z	L	American	British
<u>pass</u>	0.548	-2.423	-0.524	-0.92	1.095	0.849
<u>back</u>	-0.519	-0.01	-0.273	-0.948	0.006	-0.414
<u>bought</u>	-0.238	-0.734	-0.524	1.005	0.771	0.738
<u>bit</u>	-1.773	-0.734	-1.275	-1.333	-1.23	-0.974
<u>book</u>	-1.474	0.852	-0.228	-0.52	-1.186	-0.974
<u>bird</u>	1.316	1.025	2.299	1.433	1.918	1.811
<u>juice</u>	0.829	1.542	0.433	1.176	-0.289	0.424
<u>pot</u>	-0.594	-0.596	-1.093	-0.421	-0.627	-1.231
<u>bet</u>	-0.332	-1.32	-0.546	-0.977	-0.362	-0.817
<u>beat</u>	-1.567	-0.32	-0.501	-0.991	-0.377	-0.459
<u>paid</u>	1.259	-0.079	1.366	0.52	2.08	1.364
<u>bite</u>	0.155	0.197	-0.159	1.29	0.035	-0.135
<u>choice</u>	1.409	1.232	1.639	1.148	-0.48	-0.202
<u>house</u>	1.166	-0.561	0.979	0.692	-0.274	-0.347
<u>boat</u>	0.567	0.611	-0.501	-0.92	-0.612	-0.403
<u>duke</u>	-0.65	1.197	0.319	0.478	-0.951	-0.325
<u>york</u>	-0.145	-0.01	-0.159	0.535	1.197	2.113
<u>pass</u>	0.548	-2.423	-0.524	-0.92	1.095	0.849

2015), while the latter only calculates the straight-line distance between two target points. In this paper we compute the vowel distance between speakers of different groups (e.g., L2 vs. American, L2 vs. British), so we use the Mahalanobis DS approach in assessing the production accuracy of L2 speakers.

The computed Mahalanobis DS data will be analyzed in §5.1 by a correlation analysis, which probes the relevant spectral cue weighting in NS listeners' perception of NNS vowel intelligibility and accentedness.

4.2. Vowel duration

Another acoustic property, duration, was also measured and extracted between the vowel onset and offset by means of a Praat script. The temporal data of vowels was then transformed to Z-scores for the purpose of normalization using the following formula,

$$Z = \frac{x - \mu}{\sigma}$$

where x is the vowel duration, μ the mean duration of the vowel, and s the standard deviation of the duration. The normalized vowel duration of L2 speakers, American speakers and British speakers was computed with the above formula respectively (Table 16). The durational difference of each vowel between L2 speakers and native speakers was calculated as the L2 speakers' duration minus the native speakers' duration. In the following section we will use a correlation analysis between the duration data and the perception results to explore the role of the temporal cues in vowel perception.

4.3. Summary

Two acoustic properties of vowels, formant frequencies and duration were measured using Praat and computed with normalization methods. The Mahalanobis DS was used to calculate the spectral differences in vowels produced by Chinese L2 speakers and the two groups of native English speakers. The spectral and temporal data is analyzed in the correlation study in the next section to investigate the relative weighting of the two acoustic cues that NS listeners use in the perception of vowels produced by L2 speakers.

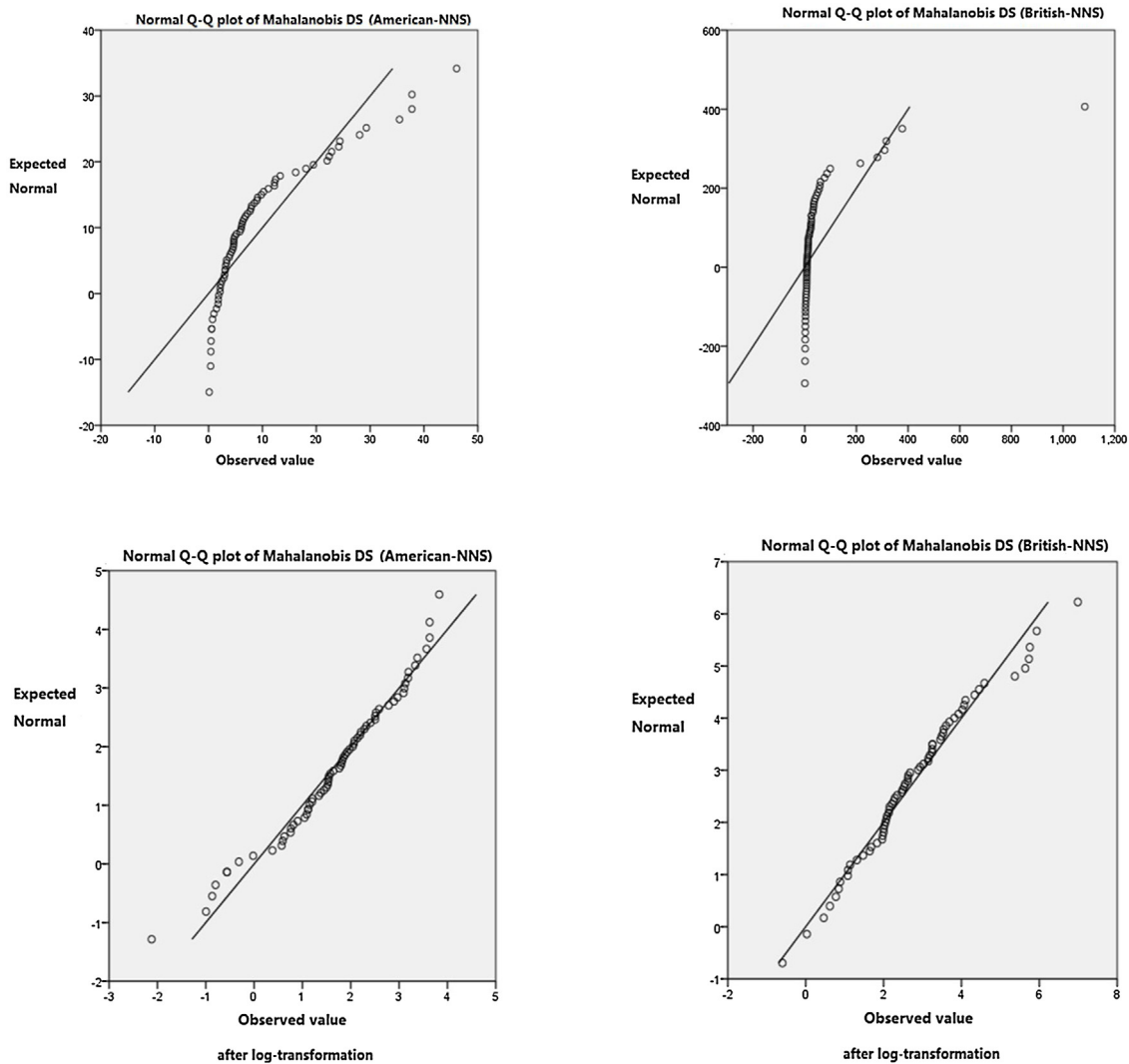


Fig. 4. Q-Q plots of data distribution before (in upper figures) and after (in lower figures) the log transformation.

5. Cue weighting in NS listeners' perception

5.1. Spectral cues and perception results

The Mahalanobis DS between the vowels produced by L2 speakers and American speakers is correlated with the perception results of American listeners, while the DS between the vowels produced by L2 speakers and British speakers is correlated to the perception results of British listeners. To achieve the normal distribution of data required for the correlation analysis, all the Mahalanobis data is log transformed. A Q-Q plot is used to present the data distribution before and after the log transformation. The DS values (the upper two of Fig. 4) show a non-normal distribution before the transform, which is inappropriate for statistical analysis, whereas after the log transform all the data is normally distributed (the lower two). Furthermore, we have carefully checked the Mahalanobis values and eliminated extreme values of the data (e.g., values above 200) from our further analyses.

From the results of the Pearson correlation analysis (Table 17), we find that the American listeners' intelligibility evaluation is significantly correlated with the spectral Mahalanobis DS in the vowels produced by the L2 speakers and the American speakers, $r(68) = -.322, p = 0.007 (** < 0.01)$. The perceptual accentedness rating is also significantly

Table 17
Correlation statistics (American listeners' perception results and the L2-American Mahalanobis DS).

		Intelligibility	Accentedness	Mahalanobis DS
Intelligibility	Pearson Correlation	1	.860**	-.322**
	Sig. (2-tailed)		0	0.007
	N	68	68	68
Accentedness	Pearson Correlation	.860**	1	-.351**
	Sig. (2-tailed)	0		0.003
	N	68	68	68
Mahalanobis DS	Pearson Correlation	-.322**	-.351**	1
	Sig. (2-tailed)	0.007	0.003	
	N	68	68	68

** Correlation is significant at the 0.01 level (2-tailed), *Correlation is significant at the 0.05 level (2-tailed).

Table 18
Correlation statistics (British listeners' perception results and the L2-British Mahalanobis DS).

		Intelligibility	Accentedness	Mahalanobis DS
Intelligibility	Pearson Correlation	1	.716**	-.347**
	Sig. (2-tailed)		0	0.004
	N	68	68	68
Accentedness	Pearson Correlation	.716**	1	-.389**
	Sig. (2-tailed)	0		0.001
	N	68	68	68
Mahalanobis DS	Pearson Correlation	-.347**	-.389**	1
	Sig. (2-tailed)	0.004	0.001	
	N	68	68	68

** Correlation is significant at the 0.01 level (2-tailed), *Correlation is significant at the 0.05 level (2-tailed).

correlated with the spectral distance, $r(68) = -.351, p = 0.003$ (**<0.01). The negative correlation between the perception results and the Mahalanobis DS reveals that the closer the spectral distance, the more acceptable the NNS vowels become to NS listeners.

The correlation statistics in Table 18 between the British listeners' perceptual intelligibility and the L2-British Mahalanobis DS show that there is a significant correlation between the two items, $r(68) = -.347, p = 0.004$ (**<0.01). The accentedness rating and the spectral distance are also found to be significantly correlated, $r(68) = -.389, p = 0.001$ (**<0.01). The negative correlation could be interpreted in the similar way when American listeners are involved.

5.2. Temporal cues and perception results

Analysis of the temporal feature of vowels begins with categorizing the vowels into three types, lax vowels, tense vowels and diphthongs. The durational difference between vowels produced by L2 speakers and native speakers was computed by subtracting the duration of the vowel in the L2 data from that of the vowel in the native data. A two-way ANOVA was then conducted with vowel type and temporal difference as between-subject factors, to examine the effect of the temporal information on the American listeners' perception results (Table 19).

The ANOVA statistics indicates that the temporal cue does not significantly affect the American listeners' perception results in terms of either intelligibility evaluation or accentedness rating. The vowel type, however, has a significant main effect on the American listeners' intelligibility evaluation results, $F = 4.46, p = 0.016$ (*<0.05), and a weak effect on the accentedness rating results, $F = 2.76, p = 0.071$. The Bonferroni post hoc test shows that diphthongs have higher intelligibility ($M = 3.43, SD = 0.84$) than tense vowels ($M = 3.08, SD = 1.06$) and lax vowels ($M = 2.5, SD = 0.82$). The accentedness ratings of diphthongs ($M = 3.32, SD = 0.95$) are also higher than those of tense vowels ($M = 3.00, SD = 1.14$) and lax vowels ($M = 2.5, SD = 1.1$). The result reconfirms our analysis in §3.1 that the production of diphthongs by Chinese L2 speakers is the least problematic as far as the NS listeners' perception is concerned.

Table 19
Between-subject factors of ANOVA.

		Value label	<i>N</i>
Vowel type	1	lax vowel	16
	2	tense vowel	24
	3	diphthong	28
Durational difference (L2– American)	1	longer	32
	2	shorter	36

Table 20
Between-subject factors of ANOVA.

		Value label	<i>N</i>
Vowel type	1	lax vowel	16
	2	tense vowel	24
	3	diphthong	28
Durational difference (L2 – British)	1	longer	35
	2	shorter	33

Similar results are derived from the ANOVA test on the temporal cue and the British listeners' perception results (Table 20). No significant main effect of durational difference is found on the British listeners' perception results in terms of intelligibility evaluation and accentedness rating. Neither the interaction effect between vowel type nor the durational difference on the perception results is significant. The ANOVA result reveals that duration does not play a significant role in shaping the British listeners' intelligibility evaluation and accentedness rating. The vowel type has a significant main effect on British listeners' intelligibility evaluation, $F = 3.70$, $p = 0.03$ ($* < 0.05$), but a non-significant effect on the accentedness rating, $F = 2.14$, $p = 0.13$. Diphthongs are higher in the intelligibility evaluation ($M = 3.54$, $SD = 1$) than tense vowels ($M = 3.17$, $SD = 1.24$) and lax vowels ($M = 2.56$, $SD = 1.32$), and also higher in the accentedness rating ($M = 3.25$, $SD = 0.89$) than tense ($M = 3.04$, $SD = 1.04$) and lax vowels ($M = 2.63$, $SD = 0.96$).

5.3. Summary

In a correlation study we have examined the relevance of acoustic cues in the perception by American and British listeners of non-native vowel production. The statistical results show that the Mahalanobis DS between Chinese L2 speakers and two groups of native speakers show a significant negative correlation with the NS perception in terms of the intelligibility evaluation and the accentedness rating. It indicates that the spectral cues of vowels are significantly related to the NS listeners' perception of vowels. However, duration, the other acoustic feature of vowels that we have also examined, is found to have no significant correlation with the NS listeners' perception results. Our study confirms the relative weighting of acoustic cues in the English vowel perception, as reported in previous studies (Bradlow, 1995; Fox et al., 1995). Since NS speakers rely primarily on spectral cues for vowel distinctions, it makes sense for L2 speakers to be provided with more cue-specific training on the perception and production of English vowel contrasts.

6. Discussion and conclusion

This study started out with three research questions. The first question is concerned with the perceptual difference between the American and British speakers in their evaluation of the vowels produced by nonnative speakers. Considering the fact that many L2 speakers in a non-English-speaking country are influenced by different English accents, particularly from their native or nonnative English instructors, textbook recordings and multi-media sources, they have developed mixed accents in their L2 speech. Some vowels are produced in an American accent, and other vowels could sound closer to the British pronunciation. It is therefore necessary to assess the L2 speakers' vowel production with the accent variability in mind. In the two perception tasks, intelligibility evaluation and accentedness rating on the vowels produced by the L2 speakers were acquired from 16 American and 9 British speakers of English. We have found the following interesting results. (1) The two groups of native English speakers differ significantly in their accentedness ratings of Chinese L2 speakers' vowel production, and they show only marginal difference in their intelligibility evaluations. The

results imply that American and British speakers could agree with each other in the intelligibility evaluations of NNS vowels, but differ greatly in perceiving nonnative foreign accent. The result is understandable once we compare the vowel plots between the two English accents, which exhibit considerable articulatory variations in vowels such as 'bought', 'pot' and 'pass'. We suggest that the production by L2 speakers should be evaluated in a comprehensive way, taking into account both production intelligibility and accentedness, but also the perception of native English speakers with different accents. (2) When considering NS listeners' perception of different vowel types (diphthongs, tense vowels and lax vowels), Chinese L2 speakers' production of diphthongs are most acceptable to the two groups of NS listeners, as seen in the highest intelligibility percentage and accentedness rating, whereas their production of lax vowels is most problematic with the lowest intelligibility percentage and accentedness rating. The results reconfirm previous findings (Chen et al., 2001, Yang, 2011) that Chinese L2 speakers have difficulty in producing English lax vowels and in the distinction of English tense-lax vowel contrasts, due to the lack of such minimal pairs in the Chinese phonology. According to the PAM-L2 model, speakers' L1 phonology could shape their discrimination of nonnative contrasts. Since Chinese has a smaller vowel inventory than English, Chinese L2 speakers might perceptually assimilate some English vowel contrasts as instances of the same native category, resulting in the unsuccessful perception and production by L2 speakers of English lax and tense vowels. In future studies Chinese L2 speakers' perception of English vowels will be used to supplement the current data in exploring the relation between nonnative production and perception.

Our second research question centers on the relationship between acoustic measurements of L2 vowel pronunciation and its perceptual evaluations provided by native speakers. We compared the acoustic data of formant frequencies and duration in vowels produced by the Chinese L2 speakers and the native English speakers with different accents. The Mahalanobis DS was employed to compute the spectral difference in vowels produced by the L2 speakers and the native speakers, while the temporal difference was also calculated. An correlation analysis shows that the spectral Mahalanobis DS has a significant negative correlation with the perceptual intelligibility and accentedness results, but the temporal data has no significant correlation with the NS perceptual evaluations. The correlation between acoustic measurements and perceptual evaluations reveal the relative difference of cue weighting in the perception of NNS vowels by native speakers.

For our third research question, this study confirms that the most significant acoustic cue of English vowels is spectral rather than temporal information. Chinese L2 speakers' different uses of acoustic cues could be attributed to the lack of such vowel contrasts in their L1 phonology. They perceptually assimilate the contrasting sounds to the same native category, instead of establishing two new categories in their phonological space, which might lead to their primary reliance on temporal information for nonnative vowel distinction, as found in previous studies (Cebrian, 2006; Kondaurova and Francis, 2010; Rallo Fabra and Romero, 2012).

The current study also has pedagogical implications. We suggest a 'cue-weighting' strategy in which L2 instructors could focus on important acoustic cues for pairwise distinction in learners' acquisition of L2 phonological contrasts. Learners' struggles in distinguishing nonnative contrasting sounds could also be alleviated by following specific training on important cues employed by native speakers in perception and production.

7. Limitations and directions for future research

We acknowledge several limitations of this study, which will need to be taken into account in future research. Our current study employed 108 English monosyllabic words as stimuli in the perception experiment, which is only a small part of our recorded data, and included only 4 Chinese L2 speakers' English vowel production data. A larger data set from L2 speakers and native English speakers could be included to validate findings in this paper. Another point which was not addressed in the current study is the NS listeners' accentedness rating of NS pronunciation across different English accents, which could be used to compare with our results in L2 accentedness ratings. In future studies investigation on L2 production and perception relation should be carried out in further discussions of nonnative production of vowels.

Acknowledgements

The research was supported by the following funds: The Youth Project of Humanities and Social Science Research Fund of Ministry of Education of China (No. 18YJC740150, "An acoustic and articulatory study of L2 English pronunciation by dialectal Chinese speakers and the use of a visual system in English phonetic instruction"), Beijing Municipal Social Science Fund Program (No. 17YYC023, "A study on English pronunciation training with visual-articulatory methods"), and the Key Program of the National Social Science Fund of China (No. 15ZDB103, "An Interdisciplinary study of L2 English phonetic acquisition by dialectal Chinese speakers"). We are deeply grateful to Zhiqiang Li and Daniel Hirst for their tremendous help and valuable suggestions on the final draft. We also want to express our great gratitude to the editor and anonymous reviewers of the journal for their insightful comments on early and revised versions of the paper.

References

- Adank, P., Smits, R., van Hout, R., 2004. A comparison of vowel normalization procedures for language variation research. *J. Acoust. Soc. Am.* 116, 3099–3107.
- Best, C.T., 1995. A direct realist view of cross-language speech perception. In: Strange, W. (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Experience: Issues in Cross-language Research*. York Press, Timonium, MD, pp. 171–204.
- Best, C.T., McRoberts, G.W., Goodell, E., 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *J. Acoust. Soc. Am.* 109 (2), 775–794.
- Best, C.T., Tyler, M.D., 2007. Nonnative and second-language speech perception: commonalities and complementarities. In: Bohn, O.S., Munro, M.J. (Eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*. John Benjamins Publishing Company, pp. 13–34.
- Bradlow, A.R., 1995. A comparative acoustic study of English and Spanish vowels. *J. Acoust. Soc. Am.* 97 (3), 1916–1925.
- Cebrian, J., 2006. Experience and the use of duration in the categorization of L2 vowels. *J. Phonet.* 34, 372–387.
- Chen, Y., Robb, C., Gilbert, H., Lerman, J., 2001. Vowel production by Mandarin speakers of English. *Clin. Linguist. Phonet.* 15 (6), 427–440.
- Cook, V., 2002. *Portraits of the L2 User*. Multilingual Matters Ltd..
- Edwards, M.L., 1995. Developmental phonology. In: Winitz, H. (Ed.), *Human Communication and its Disorders: A Review*, vol. IV. York Press, Timonium, MD.
- Escudero, P., 2000. Developmental Patterns in the Adult L2 Acquisition of New Contrasts: The Acoustic Cue Weighting in the Perception of Scottish Tense/Lax Vowels in Spanish Speakers. Unpublished Master's Thesis. University of Edinburgh, Edinburgh, Scotland.
- Escudero, P., Boersma, P., 2004. Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition* 26, 551–585.
- Escudero, P., Benders, T., Lipski, S.C., 2009. Native, non-native and L2 perceptual cue weighting for Dutch vowels: the case of Dutch, German and Spanish listeners. *J. Phonet.* 37, 452–465.
- Flege, J.E., 1984. The detection of French accent by American listeners. *J. Acoust. Soc. Am.* 76, 692–707.
- Flege, J.E., 1995. Second-language speech learning: theory, findings, and problems. In: Strange, W. (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research*. York Press, Timonium, MD, pp. 229–273.
- Fox, R.A., Flege, J.E., Munro, M.J., 1995. The perception of English and Spanish vowels by native English and Spanish listeners: a multidimensional scaling analysis. *J. Acoust. Soc. Am.* 97 (4), 2540–2550.
- Gottfried, T., Beddor, P., 1988. Perception of temporal and spectral information in French vowels. *Lang. Speech* 31, 57–75.
- Hansen, J.G., 2008. Social factors and variation in production in L2 phonology. In: Hansen, J.G., Zampini, M.L. (Eds.), *Phonology and Second Language Acquisition*. John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 251–279.
- Hayes-Harb, R., Watzinger-Tharp, J., 2012. Accent, intelligibility, and the role of the listener: perceptions of English-accented German by native German speakers. *Foreign Lang. Ann.* 45 (2), 260–282.
- Hillenbrand, J.M., Clark, M.J., Houde, R.A., 2000. Some effects of duration on vowel recognition. *J. Acoust. Soc. Am.* 108 (6), 3013–3022.
- Kartushina, N., Frauenfelder, U.H., 2014. On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Front. Psychol.* 5, 1–17.
- Kartushina, N., Hervais-Adelman, A., Frauenfelder, U.H., Golestani, N., 2015. The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *J. Acoust. Soc. Am.* 138 (2), 817–831.
- Kondaurova, M.V., Francis, A.L., 2010. The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: comparison of three training methods. *J. Phonet.* 38 (4), 569–587.
- Lobanov, B.M., 1971. Classification of Russian vowels spoken by different listeners. *J. Acoust. Soc. Am.* 49, 606–608.
- Major, R.C., 2014. *Foreign Accent – the Ontogeny and Phylogeny of Second Language Phonology*. Routledge, New York.
- Munro, M., Derwing, T., 1995a. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Lang. Learn.* 45, 73–97.
- Munro, M., Derwing, T., 1995b. Processing time, accent and comprehensibility in the perception of native and foreign accented speech. *Lang. Speech* 28, 289–306.
- Munro, M., Derwing, T., 2011. The foundations of accent and intelligibility in pronunciation research. *Lang. Teach.* 44 (3), 316–327.
- Munro, M., 1998. The effects of noise on the intelligibility of foreign-accented speech. *Stud. Second Lang. Acquis.* 20 (2), 139–154.
- Rallo Fabra, L., Romero, J., 2012. Native Catalan learners' perception and production of English vowels. *J. Phonet.* 40 (3), 491–508.
- Thomas, E.R., Tyler, K., 2007. NORM: The Vowel Normalization and Plotting Suite. [Online Resource: <http://ncslaap.lib.ncsu.edu/tools/norm/>]
- Yang, C.L., 2011. Vowel undershoot in production of English tense and lax vowels by Mandarin and American speakers. *J. Acoust. Soc. Am.* 130 (4), 2522.

ZHI Na is a researcher and lecturer at College of Foreign Languages, Capital Normal University in Beijing, China. She finished her postdoc at Chinese Academy of Social Sciences. Her research focuses on second-language acquisition and phonetics.

LI Aijun is a professor at Institute of Linguistics, Chinese Academy of Social Sciences in Beijing, China. Her research interests are phonetics, phonology and language acquisition.

[This paper was published at *Lingua*, Volume 256, 2021]