On which side of the Atlantic is Chinese-accented English?
An acoustic comparison of Mandarin, British and American English monophthongs

Wei He and Vincent J. van Heuven
Phonetics Laboratory, Leiden University Centre for Linguistics

1. Introduction

We all know from experience that it is very difficult to acquire a (near-)native pronunciation in a foreign language when the learning process takes place after puberty. Learning to pronounce a foreign language is difficult enough when the learner’s native language (source language, SL) is genealogically close to the foreign language (target language, TL), as is the case for Dutch and English. The task is much more difficult when the learner’s mother tongue is genealogically remote from the target language, as is the case for Chinese (Mandarin) learners of English. The People’s Republic of China (with its 1,250,000,000 population) is the largest reservoir of learners of English as a foreign language. English training begins at primary school and is continued on a compulsory basis until the last year of undergraduate education at the university level. In this paper we address the pronunciation of English vowels by Chinese learners. This is by no means a new topic. Many phoneticians and language pedagogues have compared the English and Chinese vowel systems and observed that the English vowel system is more complicated. There are many vowels in English which do not have direct counterparts in the Chinese vowel system. And these vowels are claimed to be difficult for Chinese learners, who often fail to find the correct place of articulation.

According to Flege’s (1987) Speech Learning Model (SLM), the sounds in a source language can be divided into three categories.¹ When the sounds are transcribed with exactly the same IPA symbol (including diacritics) in SL (source language) and TL (target language), there will be no learning problems
(‘identical sounds’). When two sounds in SL and TL are transcribed with the same IPA base symbol but differ in diacritic marks, the target sound will be a long-term learning problem (‘similar sounds’). When a sound in TL does not occur in SL, learning problems will occur in the initial stages of learning, but the learner will set up a new category in due course that will be quite authentic. Figure 1 presents a comparison of the (Mandarin) Chinese and (British) English vowel inventories. Phonemes in unmarked cells are ‘identical sounds’ in Chinese and English, the phonemes in black cells are lacking in the source language (‘new sounds’). Phonemes in grey cells in Mandarin have no use in English; vowels in grey cells in English are ‘similar sounds’ in English.

Indeed, there are quite a few studies that point out that there is no tense-lax vowel contrast in Mandarin, which is needed to distinguish between the members of English vowel pairs such as /iː ~ ɨ/, /uː ~ ʊ/, /ɔː ~ ɒ/ and /ɛː ~ e/, and people in the north of China generally fail to open their mouth wide enough, so that [e] is pronounced instead of [æ] (Shen 1979: 22; Wang and Van Heuven 2006). The English tense-lax vowel contrast has two phonetic correlates. Of

<table>
<thead>
<tr>
<th>V-height (down)</th>
<th>Place of Constriction (across)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Front</td>
</tr>
<tr>
<td>Source: Mandarin</td>
<td></td>
</tr>
<tr>
<td>−round</td>
<td></td>
</tr>
<tr>
<td>−round</td>
<td>−round</td>
</tr>
<tr>
<td>High</td>
<td>i</td>
</tr>
<tr>
<td>High-mid</td>
<td>e</td>
</tr>
<tr>
<td>Low-mid</td>
<td>e</td>
</tr>
<tr>
<td>Diphthong</td>
<td>ai</td>
</tr>
<tr>
<td>Target: English</td>
<td></td>
</tr>
<tr>
<td>Tense</td>
<td></td>
</tr>
<tr>
<td>Tense</td>
<td></td>
</tr>
<tr>
<td>Tense</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>æː</td>
</tr>
<tr>
<td>Diphthong</td>
<td>ai</td>
</tr>
</tbody>
</table>

**Figure 1.** Contrastive vowel analysis of Mandarin and English. Grey cells in Mandarin denote source sounds are not needed in English. Grey cells in English denote ‘similar sounds’. Black cells in English represent ‘new sounds’ and white cells ‘identical sounds’. (After Wang 2007: 48). Rising diphthongs and triphthongs (which do not occur in English) have been omitted from the chart. Further see text.
these the difference in phonetic quality (peripheral for tense versus centralized for lax) is claimed to be the primary correlate, whilst the duration (long for tense versus short for lax) is held to be secondary. It has been shown before that foreign learners tend to be more sensitive to duration differences in a target language than to small differences in phonetic quality (Van Heuven 1986).

In the present study we ask several questions. First, given that Chinese learners of English are influenced by both American English through the media and by (southern) British English (or Received Pronunciation), RP, in parts of the education system, which pronunciation standard is approximated best by the typical Mandarin Chinese accent? If the resulting pronunciation, in spite of the teachers’ efforts, is American rather than British, we would recommend giving up the British norm and changing the pronunciation standard to American English. Second, since female learners have been generally found to perform better in foreign languages than men, is the gender difference also seen in the acquisition of English as a foreign language with Chinese learners? Third, Flege’s SLM predicts that the new sounds in English will cause learning problems to Chinese speakers in all but the final stages of the acquisition process. Specifically, when it comes to the tense-lax contrast, which strategy will be dominant among the Chinese learners of English? Will they rely primarily on vowel quality as the native speakers do, or will they rely predominantly on the duration difference as has been observed before in foreign language learning?

2. Method

To answer the above questions we recorded English speech from a representative sample of Chinese learners of English as a foreign language and compared the results to native English control data published in the literature.

Sixteen Chinese speakers were selected, 7 males and 9 females. They had been enrolled in practically the same English learning curriculum, with about ten years of training in English at primary and secondary school and two more years at university. Most importantly, they all hailed from the same dialect region, Tianjin, a metropolis 120 km south of Beijing. The Tianjin dialect belongs to the same dialect subgroup as Beijing, and the pronunciation is almost the same as that of standard Mandarin, except for the tones in some words. No further inclusion criteria were applied; students varied considerably in their level of achievement of English.
The subjects were recorded at a quiet language laboratory in Nankai University, Tianjin, China. The vocal output was digitally recorded through a head-worn microphone (Jiahe HTS-860MVI) with Audition 1.5 software (44.1 KHz, 16 bits). The subjects read the materials in individual sessions, which took about two minutes each for the whole material. From the recordings we selected the ten vowel types that have been shown in earlier research to be the most difficult to identify correctly (Wang 2007). Thus we selected the same set of ten monophthongs (two of which are also considered to be semi-diphthongs) for each of our speakers, with each vowel type instantiated by three tokens, viz. one /hVd/ form and two high-frequency key-words, as shown in Table 1.

Table 1. Selection of stimulus material used in the present study.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>/hVd/</th>
<th>Transcription</th>
<th>Key-words</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPA</td>
<td>Fig</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. i:</td>
<td>i:</td>
<td>heed</td>
<td>/hiːd/</td>
</tr>
<tr>
<td>2. i</td>
<td>ih</td>
<td>hid</td>
<td>/hid/</td>
</tr>
<tr>
<td>3. eː</td>
<td>ey</td>
<td>hayed</td>
<td>/heːd/</td>
</tr>
<tr>
<td>4. e</td>
<td>eh</td>
<td>head</td>
<td>/hed/</td>
</tr>
<tr>
<td>5. æ</td>
<td>ae</td>
<td>had</td>
<td>/hæd/</td>
</tr>
<tr>
<td>6. uː</td>
<td>uː</td>
<td>who’d</td>
<td>/huːd/</td>
</tr>
<tr>
<td>7. ʊ</td>
<td>uh</td>
<td>hood</td>
<td>/hud/</td>
</tr>
<tr>
<td>8. ɔː</td>
<td>ɔː</td>
<td>hawed</td>
<td>/hɔːd/</td>
</tr>
<tr>
<td>9. ǝʊ</td>
<td>ǝʊ</td>
<td>hoed</td>
<td>/hɔːd/</td>
</tr>
<tr>
<td>10. ʌ</td>
<td>ʌ</td>
<td>hud</td>
<td>/hʌd/</td>
</tr>
</tbody>
</table>

All the recordings were segmented and labeled by ear and eye using Praat speech analysis software (Boersma 2001). The centre frequencies of the lowest two vowel resonances (F₁, representing vowel height, and F₂ representing vowel constriction place) were measured semi-automatically (in hertz, Hz). Vowel duration (in milliseconds, ms) was extracted automatically from the labeled files. For procedural details see Wang (2007) and Wang and Van Heuven (2006).

Five datasets with native English control materials were selected from published papers. Due to differences in experimental procedures, the data in each of the control sets do not fully match those of our Chinese learners. Two RP English and three American English (AE) datasets were selected. The RP English data (Deterding 1997) only contain the mean formant values of our
ten monophthongs. No data are available on individual speakers. Moreover, the words are not in /hVd/ form, and no vowel durations are specified. Wells’ (1962) data also contain mean formant values of the ten monophthongs recorded in /hVd/ words, but the results were not broken down by gender. For the American English reference materials, Peterson and Barney (1952) and Hillenbrand, Getty, Clark and Wheeler (1995) both report mean formant values of the ten monophthongs recorded in /hVd/ words, but list no vowel duration data. Wang’s (2007) presents formant values of all ten vowels spoken by 20 native English speakers (10 male, 10 female), fully matching our non-native data, including a breakdown by gender and specifying vowel duration.

The formant frequency values (F in Hz) were transformed to an auditory Bark scale using the Zwicker and Terhardt (1980) formula. To abstract away from differences in context between our recordings and the control data, as well as from differences between individual speakers (including the gender difference), all F1 and F2 values (in Bark) as well as vowel durations were Z-normalized within-speakers (see Lobanov 1971, Wang and Van Heuven 2006).

For each foreign-accented vowel, we computed its Euclidean distance in the acoustic vowel space (defined by Z-normalised F1 and F2) from either the RP or AE control vowels. These distances were computed only for /hVd/ tokens (with the exception of Deterding’s (1997) control data — which were collected from spontaneous speech).

Next, we computed the distance of members of tense-lax vowel pairs from the centre of the vowel space (where Z1 = Z2 = 0, which is where we would expect the neutral vowel schwa to be located). Finally, we established a tense-lax duration ratio by dividing the tense vowel duration by the duration of the lax counterpart.

3. Results

3.1 Non-native and native vowel space contrast

The purpose of this experiment is to find the acoustic difference of Chinese-accented English (L2) and native English (L1). Figure 2 plots the L1 and L2 vowels in the Z-normalised vowel space. Each panel shows the contrast between the L2 data and one British L1 control reference data set (Wells 1962 in the left-hand panel, Deterding 1997 in the right-hand panel). The extreme vowels are linked by polygons. The dark polygons delimit the L2 speakers, while the light ones represent the L1 controls. The vowel locations in the two reference data
sets are virtually identical, and so is therefore the difference of the L2 vowels to the reference vowels. The vowel height of Mandarin speakers is close to that of RP English speakers, at least when we consider the high and low vowels only. The English mid vowels /ɛ/ and /ɔː/, however, constitute a problem. Both are articulated too open. This observation is in line with the prediction that ‘new’ vowels are problematic in the beginning and intermediate stages of the foreign language acquisition process but may also reflect an American pronunciation model. Also, the non-natives seem to have a problem with the constriction place of /æ/, which is too far back, and suggests a lack of contrast between /æ/ and /ʌ/. Discrepancies in the realisation of the constriction place are consistently seen in all the ‘similar’ vowels, i.e. front /i:/, /i/, and back /u:/.

Figure 2. Vowel space contrast of non-native (NNE) to RP English reference data. The left panel contains Wells’ (1962) data, the right panel uses Deterding (1997) as the reference. For transcription symbols see Table 1 (‘Fig’).

Figure 3 displays the same L2 English vowels as in Figure 2 but now paired with three reference vowel sets taken from American English. The three reference vowel spaces differ substantially. In each panel of Figure 3 the L2 vowels are rather more distant from their reference location than was the case with the British reference vowels in Figure 2. This seems to indicate that the Mandarin-accented vowels are more different from the American reference locations than from the British targets. The greater discrepancies are observed both in the vowel height dimension and in the constriction place. High vowels /i:/ and /u:/ are not high enough. The low vowel /æ/ seems too low and back in the
comparison with the Peterson & Barney data, but too high and front when we compare with the Hillenbrand et al. data. ‘New’ and ‘similar’ vowels /ɔː/, /ɛ/, /ɪ/ and /ʌ/ all have larger distances to the corresponding reference location.

3.2 Gender effect on the L2 learning

In Section 3.1, the vowel distances were discussed informally, on the basis of visual inspection of graphs. We will now quantify the discrepancies between L2 vowels and their counterparts in the various reference datasets more formally. At the same time we will break the results down for the male and female
speaker groups. Figure 4 presents the mean discrepancy between L2 and L1 vowel counterparts, broken down by British (light bars) versus American (dark bars) reference sets for male and female L2 speakers. In the top row the English reference vowels were taken from Wells (1962), the British reference set in the bottom row was taken from Deterding (1997). The American English datasets are those by Peterson and Barney (left-hand column), Hillenbrand et al. (middle column) and Wang (right-hand column). The top-left panel, for instance, shows the mean spectral distance (after bark-transformation and z-normalisation) between the Mandarin-accented English vowels from the L1 targets in British English (light bars) and in American English (dark bars) for male (left bar cluster) and female (right bar cluster) EFL speakers, where the British targets were taken from Wells and the American targets from Peterson and Barney.

The vowel distance to native norms is consistently larger for females than males, indicating that males did a better job than females in this experiment.

![Figure 4](image_url)
Moreover, the distance is consistently larger for American reference vowels than for learners of English, but the explanation is still unclear.

In addition, another comparison was conducted. Table 2 lists the four most difficult English target vowels (in descending order, and broken down by gender) as evidenced by a comparison between the Mandarin-accented tokens and the reference vowels in each of the five control datasets.

Table 2 can be summarized as follows. The most difficult vowels for male Mandarin learners of English are those in *hawed*, *head*, *hood*, and *had*, and for females those in *hawed*, *head*, *hid*, and *had*. Both genders share the difficult vowels *hawed*, *head*, and *had*, but there is still one difficult vowel that is specific to each gender (*ʊ* for male learners and *ɪ* for the females). Co-incidentally, the most difficult vowels are consistently ‘new’ vowels to the L2 learners (see also Figure 1). This result is (predictably) in line with earlier observations based on visual inspection of vowel graphs (Figures 2 and 3) and supports the claim based on Flege’s Speech Learning Model.

### 3.3 Tense-lax vowel contrast pairs

Since there is no tense-lax vowel contrast in Mandarin Chinese, in contradistinction to English, the contrast it is claimed to be difficult for the L2 learners. The target vowels are either ‘similar’ sounds (differing in the length diacritic only, and therefore most difficult) or ‘new’ sounds (differing in the base transcription symbol, a difficulty that will only disappear at the final stage of the L2 acquisition process). What strategy will be adopted by L2 learners? Will they exploit the vowel quality difference between the English tense versus lax vowels or will they mark the contrast by duration? Among the ten vowels in
this experiment, four vowel pairs are available for comparison, \(/æːɛ/, /ɛːɛ/, /iːɪ/, /uːʊ/\). The target /ɛ/ is used twice since it pairs with both a lower and a higher tense vowel. This is mentioned by other researchers (Shen 1979, Strange et al. 2004, Wang and Van Heuven 2006).

We computed the Euclidean distance in the Z₁ by Z₂ (normalised bark) vowel space between the L2 token and the centre of gravity of the vowel space (where Z₁ = Z₂ = 0, i.e. the theoretical position of the neutral vowel schwa) as an indication of the degree of vowel reduction. Since English lax vowels are spectrally more reduced (closer to schwa) than their tense counterparts, the difference in this degree of reduction will serve as an adequate measure of spectral contrast between the tense and lax members of an opposition. In addition to this we also computed the tense:lax duration ratio. Duration data are missing from all RP English reference data and are therefore shown on the ratio graph only as markers with no error bars around them. The comparison and analysis of duration ratios can be done for the American reference sets only.

Apparently there is no RP /ɛːɛ/ vowel pair in the dataset. The spectral contrast between tense and lax counterparts is clearly smaller for the L2 speakers than for either RP or American English speakers. The smaller tense/lax spectral contrast among L2 learners indicates that these learners are barely aware of the articulatory difference between the pairs. In terms of temporal marking, non-natives have much larger ratios between the tense and lax members of the contrast than the native speakers, showing that duration

![Figure 5](image.png)

**Figure 5.** Difference in Euclidean distance from schwa between tense and lax counterparts (left panel) and mean tense:lax vowel duration ratio (right panel) for four tense–lax vowel pairs, broken down by Mandarin learners (‘Tianjin’), British native speakers (‘RP’) and American native speakers (USA). Error bars indicate 95% confidence limits of the mean.
information plays a critical role in the learners’ pronunciation. The learners tend to over-emphasise this feature in their articulation. Moreover, the duration ratio is unevenly distributed over the pairs, showing that non-natives do not apply their duration strategy consistently.

4. Conclusion and discussion

In this experiment, we observed how Chinese-accented English vowels differ from native English tokens in acoustic terms. Since we studied RP English and American English as separate pronunciation standards, we are now in a position to determine whether Chinese-accented English vowels are closer to the one or the other standard. We also wished to test the claim that the female gender has an advantage in language learning. Our results also afford a test of the prediction made by Flege’s Speech Learning Model that ‘new’ sounds constitute a problem in L2 learning. As the tense-lax contrast is distinctive in English but absent in Mandarin, we expected that Mandarin speakers would over-emphasise the duration feature in the tense-lax pairs, and ignore the spectral acoustic cues.

In the vowel space comparison, we found that our Mandarin speakers’ vowel pronunciation was closer to RP English than to Standard American English. Mandarin speakers performed relatively poorly in managing vowel contrasts along the front-back dimension. They performed even more poorly on this dimension when the comparison was made relative the American reference locations.

In the gender advantage test, the results run counter to our expectation that female speakers do a better job (in the comparison with all English reference datasets, showing closer vowel distance to the norms than the male speakers did). The top-four most difficult vowels are those in hawed, head, had, irrespective of the gender of the Mandarin learner. The only gender difference in the top four most difficult vowels is found between hood in males and hid in females. These findings lend further credibility to Flege’s Speech Learning Model that predicts that the ‘new’ sounds /æ/, /iː/, /uː/, and /ɔː/ in English will be difficult for L2 learners of English with a Mandarin Chinese native language background.

The results obtained for the tense-lax vowel pairs support our prediction. Mandarin speakers show smaller distance to schwa than was observed in either the British or American reference datasets in most pairs, indicating smaller
spectral difference in realizing the contrast. Crucially, however, the contrast between the tense and lax members of the oppositions was more clearly expressed in the vowel duration ratio by Mandarin learners of English than by the native speakers themselves. This indicates that Mandarin speakers over-emphasise the duration contrast.

Supported by the acoustic analysis, both RP English and American English have influence on L2 learners. Even though L2 speakers show smaller vowel distances to RP English norms than to American reference vowels, it is still unknown whether the learners were primarily influenced by the RP English norm, or whether RP English vowel targets come more naturally to Chinese learners of English (for instance due to greater inherent similarity between the Mandarin and British English vowel targets). Moreover, the gender difference shown in this experiment runs counter to the common assumption. Is this result an artefact of the present experiment? Could it be the case that women would have been more successful than the men if also consonant targets had been included? Is it possible that the (intuitive) claims of female superiority in language learning are based on better fluency and more vivid facial expressions that would help intelligibility?

The results also support Flege’s SLM in that ‘new’ vowels are still difficult to learners in the intermediate stage. Ways must be found to help the learners shorten the learning process. One possibility would be to use computer-controlled automatic comparison of target vowels (whether British or American) and imitations of these by (Chinese) learners. The learner would get immediate feedback (also visually in a vowel space on a computer screen) how close his imitation was to the target. Such systems have been developed for use with the auditorily handicapped, even several decades ago (e.g. Povel and Wassink 1985), but have not yet found their way to the common classroom.

Notes

1. Best’s Perceptual Assimilation Model (PAM, Best, McRoberts and Goodell 2001) addresses the problem of how non-native vowels are categorized on first confrontation. Our speakers are (certainly by Chinese standards) advanced speakers of English and no longer naive with respect to the sounds of the target language. Flege’s SLM, on the other hand, explicitly targets the intermediate and final stages of language learning. That is why we take our cue from SLM rather than from PAM.

2. Bark = 13 × arctan (0.76 × F) + 3.5 × arctan (F/7.5)^2
3. There is no clear indication that the Mandarin vowels that could be used as substitutes for the English target vowels by Chinese learners, are systematically closer to either the British or the American norm. Formant values for Mandarin vowels have been published by Li, Yu, Chen and Wang (2004). Only three vowels can be matched directly with tense English counterparts (‘identical sounds’): /i, u, ɔ/. Mandarin /e/ and /o/ cannot be matched with English vowels since their closest counterparts are semi-diphthongs and were therefore not included in our study. Four Mandarin vowels can be matched only imperfectly with lax English vowels (‘new sounds’): /æ, ɛ, ɪ, ʊ/. The mean spectral distance between the three Mandarin and English matching vowel pairs is 1.7 bark for British and 2.0 bark for American targets, i.e. favouring British English. The mean distance between the four non-matching vowel pairs is 2.7 and 2.5 bark for British and American English, respectively, i.e. favouring American English. The difference between British and American English is insignificant in either of the two comparisons.

4. The two RP datasets are remarkably similar (though not identical) in spite of the fact that Deterding’s (1997) free-speech recordings were made three decades later than Well’s (1962) tokens. In contrast to this, the American vowels show considerable change over time (and possibly also over place) as especially /u:/ is centralized in the more recent Hillenbrand et al. (1995) and Wang (2007) recordings than in the older recordings by Peterson and Barney (1952). It should be pointed out that the American vowel sets were collected from speakers of quite different local backgrounds. The male speakers in Peterson & Barney (1952) were sampled from all over the USA, while their female speakers were from the Central Atlantic area. Hillenbrand et al.’s speakers hailed from the upper midwest, while Wang’s recordings were made in Southern California. The latter circumstance would be compatible, for instance, with the centralization of /u:/.

References


288. **From traditional phonology to modern speech processing.** *Festschrift for professor Wu Zongji’s 95th birthday*. Beijing: Foreign Language Teaching and Research Press.


*Authors’ addresses*

He Wei
Morsstraat 46A
2312 BN, Leiden
The Netherlands
heweitracy@hotmail.com

Vincent J. van Heuven
Leiden University
Phonetics Laboratory/LUCL
P.O.Box 9515
2300 RA Leiden
The Netherlands
v.j.j.p.van.heuven@hum.leidenuniv.nl