Comparing fundamental frequency of German vowels produced by German native speakers and Mandarin Chinese learners

Yingming Gao,1,a) Hongwei Ding,2,b) Peter Birkholz,1 and Yi Lin2

1Institute of Acoustics and Speech Communication, TU Dresden, Dresden 01069, Germany
2Speech-Language-Hearing Center, School of Foreign Languages, Shanghai Jiao Tong University, Shanghai 200240, China

yingming.gao@mailbox.tu-dresden.de, hwding@sjtu.edu.cn, peter.birkholz@tu-dresden.de, carol.y.lin@sjtu.edu.cn

Abstract: This study compared the \( f_0 \) of 14 German vowels in monosyllabic words (/dVt/) embedded in carrier sentences produced by 30 native speakers and 30 Mandarin Chinese learners. Appropriate techniques were employed to robustly measure \( f_0 \) values and reliably analyze \( f_0 \) profiles. The results showed that Mandarin learners produced the vowels bearing sentence stress with significantly larger \( f_0 \) ranges and steeper \( f_0 \) slopes but comparable \( f_0 \) mean and maximum in comparison to German natives. Moreover, lax vowels produced by both groups demonstrated narrower ranges with faster \( f_0 \) changes than tense vowels, which was stronger for Mandarin learners.

© 2021 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

[Editor: Charles C. Church] https://doi.org/10.1121/10.0005593

Received: 11 February 2021 Accepted: 22 June 2021 Published Online: 14 July 2021

1. Introduction

Compared with the extensive studies on Mandarin learners speaking English (Chen et al., 2001a; Jin and Liu, 2013; Yuan and Liberman, 2014), acoustic analyses of their production of German as a second language (L2) are still limited. Studies on L2 speech learning usually start with the acquisition of vowel segments in the target language. In German, without consideration of schwa /ə/ or long tense /ɛː/, there are 14 monophthongs that can be grouped into seven pairs; the members of which differ exclusively with respect to tenseness. The phonetic-acoustic differences of the German tense-lax opposition in production may manifest through changes in vowel formants, duration, and \( f_0 \) (Schneeberg and Schlußler, 2006). The former two aspects have been compared between German native speakers and Mandarin Chinese learners (Gao et al., 2020), while the last factor (\( f_0 \) difference) remains to be investigated. Though previous studies have revealed that L2 German produced by Mandarin learners has higher \( f_0 \) mean and larger \( f_0 \) range on both sentence and phoneme levels compared with that produced by German natives (Ding et al., 2006), they have rarely concerned the \( f_0 \) profiles associated with the tense-lax vowel contrast. Vowel intrinsic \( f_0 \) (IF0) was proved to be a language universal (Whalen and Levitt, 1995). It has been shown that high vowels have a higher intrinsic \( f_0 \) than low vowels and that intrinsic \( f_0 \) also plays an important role in distinguishing the vowel identity. Therefore, the tense-lax contrast of vowels should be evident not only in formants and duration, but also in \( f_0 \).

Unlike the role of \( f_0 \) in German vowels to signal stress and possibly tenseness, \( f_0 \) in Mandarin vowels is associated with a lexical tone and employed to distinguish lexical meanings. Moreover, Mandarin Chinese monophthongs are usually classified as tense vowels, that is, there are no tense-lax contrasts in Mandarin Chinese. Such \( f_0 \) contrasts are supposed to be employed by native German speakers to distinguish between tense and lax vowels (Schneeberg and Schlußler, 2006), while L2 Mandarin Chinese learners may not use the same \( f_0 \)-related strategy due to the different roles of \( f_0 \) in their native tone language. In addition, vowel intrinsic \( f_0 \) differences are also dependent on the prosodic context in running speech, and the intrinsic \( f_0 \) difference should be maintained when the vowels bear the main phrasal stress (Shadle, 1985). Regarding the interaction between the intrinsic \( f_0 \) and the prosodic environment, we predicted that Mandarin Chinese learners might have some difficulties in using \( f_0 \) properly when they speak the non-tone language German. To address this specific issue, the current study aims to compare the \( f_0 \) profiles of German vowels produced by German native speakers and Mandarin Chinese learners with a particular interest in the tense-lax contrast under sentence stress.

a)ORCID: 0000-0001-5881-3723.
b)Author to whom correspondence should be addressed.
2. Methods

2.1 Participants

Two groups of speakers were recruited for the study, namely, a German native group (DEU) and a Chinese L2 learner group (CHN). The DEU group consisted of 30 German students studying at the TU Dresden with a mean age of 23.6 years (range: 18–38), while the CHN group included 30 Chinese L2 learners of German with an average age of 24.1 years (range: 18–31). Some CHN speakers were students majoring in German at Shanghai Jiao Tong University who had passed the nationwide unified examination for German students at the senior level of PGH (Prüfung für das Germanistik-Hauptstudium), and the others had passed the required German language examination (up to DSH-2 or DAF-16) before taking up their studies in German at TU Dresden. Speakers were exactly gender-balanced, i.e., 15 male and 15 female speakers in each group. Although they were born in different regions of their countries, the speakers in both groups had no strong regional accents. For example, all CHN speakers had achieved Grade Two Level B or above on the national standard Mandarin proficiency test (Putonghua Shuiping Ceshi), and most of them had less than one year of experience living in Germany after the age of 18 years. All participants, according to their self-reports, had normal speech and hearing functions with no history of any communication disorders.

2.2 Data collection

First, we embedded all 14 German vowels in monosyllabic words (/dVt/) to ensure that the speakers could produce the target vowels in a natural way. To create a systematic orthographic contrast, an “h” or an additional “t” was placed after the target vowel to indicate a tense or a lax vowel, respectively. Thus, we obtained 14 words, and most of them were non-sense but legal phoneme strings according to German phonotactic rules. These 14 words consisted of seven pairs with their International Phonetic Alphabet (IPA) transcriptions in parentheses as follows: daht-dutt (/dæt/-/dæt/), deht-dett (/dɛt/-/dɛt/), diht-ditt (/diɛt/-/diɛt/), doht-dott (/dɔt/-/dɔt/), diuht-dutt (/duet/-/duet/), daht-dutt (/dæt/-/dæt/). The regular spelling patterns facilitated the correct grapheme-to-phoneme conversion for the speakers, so that the target vowels could be easily elicited. Moreover, we put each of the target words in a carrier sentence, “Ich habe /dVt/ gesagt (I have said /dVt/),” to ensure a stable prosodic context. By randomizing each set of the 14 sentences five times, we created a reading list of 70 sentences, and thus it was guaranteed that each vowel was produced five times with different intra-group orders by each speaker. The speakers were told that they should read all the sentences as naturally as possible with a short pause between them. After a period of familiarization and practice, all the speakers chose to place a pitch accent on the target word automatically. This way, \( f_0 \) values of target vowels under sentence stress could be elicited in an implicit way with well-controlled prosody. Though several CHN recordings were made in Shanghai and the others were in Germany, we ensured the same instructions and conditions. All recordings took place in a studio equipped with a recording console (Behringer Eurorack MX1602). The microphone (Microtech Gefell M930) was placed at a distance of approximately 20 cm from the speaker’s mouth. All utterances were recorded with a sampling rate of 44.1 kHz and a quantization of 16 bits. The experiment lasted for about 5 min for each speaker, and they were financially compensated for their participation.

2.3 Acoustic measurements

In the first step, an automatic forced-alignment was carried out via the WebMAUS service (Kisler et al., 2017) on both word and phoneme levels, outputting a TextGrid format annotation of Praat (Boersma and Weenink, 2019). Based on the derived word-level boundaries, we segmented all the recordings into 4200 individual sentences (14 vowels × 5 repetitions × 15 speakers × 2 genders × 2 language groups). Inaccurate alignments from the automatic phoneme annotation were manually adjusted by a phonetic expert by taking into account changes in both waveforms and spectrograms as well as perception cues if necessary.

A five-step procedure was applied to achieve a high accuracy of \( f_0 \) estimation. The first step was to extract the fundamental frequencies by the pitch-tracker developed by Shi et al. (2019), the analysis window of which was set at a length of 30 ms with 5 ms shift. A robust \( f_0 \) estimate and a voicing probability for each frame of speech were obtained. In the second step, we carried out the pitch tracking through a two-pass procedure following the strategy proposed by Hirst (2011) by calculating a more accurate \( f_0 \) range for \( f_0 \) estimation. In the first pass, we inspected our data and set a more accurate search range of 150–400 Hz and 75–300 Hz for female and male speakers, respectively, to cover all reasonable \( f_0 \) samples, and we extracted the \( f_0 \) with this range. Then we calculated the first and third quartiles (i.e., \( q_1 \) and \( q_3 \)) across all \( f_0 \) samples for each speaker. In the second pass, the \( f_0 \) floor and ceiling for each speaker were set to 0.75\( q_1 \) and 1.5\( q_3 \), respectively. By using a personalized search range, we greatly reduced the estimation errors of \( f_0 \) extraction. This was confirmed by comparing speakers’ \( f_0 \) histograms, in which long tails disappeared and samples were more centralized around the mean values. In the next step, a frame of speech with a voicing probability smaller than 0.5 was automatically removed from the data for \( f_0 \) extraction because these \( f_0 \) values were considered unreliable according to Shi et al. (2019). The fourth step was incorporated to deal with creaky voice that was frequently produced by several speakers. In glottalized periods, the corresponding \( f_0 \) was estimated by another pitch-tracker (Drugman and Alwan, 2011), which was more robust to
glottalization. Finally, we applied a median filter with a window of seven $f_0$ samples to smooth the $f_0$ contour. Manual corrections were only applied when the values were still wrong during the final check. For example, if the $f_0$ samples with voicing probabilities slightly larger than 0.5 were actually voiceless, we had to make necessary corrections manually.

2.4 Acoustic analysis

Based on the optimized $f_0$ samples, we calculated the $f_0$ mean, $f_0$ range (maximal $f_0$ minus minimal $f_0$), and $f_0$ slope of each target vowel. Following Lehiste and Peterson (1961), we also measured the $f_0$ maximum as a complement of $f_0$ mean. We further measured the positions of the $f_0$ maximum and minimum of each target vowel. Each vowel variable for a specific speaker was the average of his/her five repetitions of this vowel. We adopted two approaches to make the acoustic analysis more precise and robust.

One statistic approach to alleviate the influence of physiological differences efficiently was to convert the physical measurement of $f_0$ to the perceptual variable of $f_0$ using speaker-specific bases, which made the $f_0$ produced by all speakers comparable across gender. In previous studies, $f_0$ was usually converted from the Hz scale to the semitone (ST) scale with a fixed value as a reference, which did not change the relative relationship between them due to the monotonic property of the logarithmic function (Chen et al., 2001b; Ding et al., 2006; Zhang et al., 2008). In the current study, we adopted the reference proposed by Yuan and Liberman (2014), where each $f_0$ in Hz was transformed to a ST value according to Eq. (1), in which the $f_0$base was the speaker-specific 5th percentile of all $f_0$ in Hz scale

$$f_0[St] = 12 \log_2 \left( \frac{f_0[Hz]}{f_0\text{base}} \right).$$

Another statistical approach was to measure the $f_0$ slope by conducting a linear regression with time as an independent variable and $f_0$ as the dependent variable, which was more robust than the usual practice of dividing the absolute $f_0$ range (difference between the maximal and minimal $f_0$ values in ST) by the duration of the vowel. The slope we obtained could thus characterize the dynamic movements of $f_0$ contours, where a positive slope represented an overall rising pattern, and a negative slope indicated an overall falling one. The absolute value of the slope reflected the steepness of the rising or falling. In the case of an $f_0$ contour containing two parts (LH plus HL or HL plus LH), the value of the slope was dominated by the longer part or the relatively steeper part, i.e., the dominant part contributed more to the direction of the estimated slopes.

3. Results

A series of linear mixed-effects (LME) models were run in MATLAB (MathWorks, 2019), where SpeakerGroup, Gender, Voweldentity, or Tenseness and their interactions were treated as fixed effects and Subject as a random effect for inter-

3.1 $f_0$ mean and maximum

Average group values for $f_0$ mean (in Hz), maximum (in Hz), and $f_0$ mean (in ST) of the target vowels are shown in the first, second, and third rows, respectively, in Table 1. It can be observed that the female speakers produced higher $f_0$ mean (245 Hz) and maximum (261 Hz) (measured in Hz) than the male speakers (137 Hz; 147 Hz); also, high vowels (199 Hz for /i:/; 1 y; y u u/) were associated with higher $f_0$ than low vowels (179 Hz for /a:/; a/).

The results for $f_0$ mean or maximum (in Hz) showed similar patterns: significant effects were found for Gender and Vowelentity but not for SpeakerGroup. The LME regression for $f_0$ in ST revealed significant effects of Vowelentity $F(13,784) = 92.25, p < 0.001$ but non-significant effects of both SpeakerGroup $F(1,784) = 0.02, p = 0.877$ and Gender $F(1,784) = 0.28, p = 0.599$. However, the effect of Gender was statistically significant for $f_0$ mean in Hz $F(1,784) = 282.62, p < 0.001$. In other words, by transforming $f_0$ from Hz to ST with Eq. (1), we preserved the difference of $f_0$ mean due to vowel intrinsic $f_0$ effects but reduced the difference due to the gender effect. A significant interaction effect on $f_0$ mean in ST was found between SpeakerGroup and Vowelentity $F(13,784) = 10.85, p < 0.001$. For example, the $f_0$ mean (in ST) of vowels /u/ and /u/ ranked as the first and seventh among 14 monophthongs for the DEU speakers, respectively, while they ranked as the sixth and fourth for the CHN speakers, respectively. Moreover, the effect of Tenseness was significant $F(1,838) = 45.11, p < 0.001$, with lax vowels generally having a higher $f_0$ than their tense counterparts (6.06 St versus 5.6 St). This test was conducted for group mean differences between tense and lax vowels, and no post hoc test was applied to each individual pair.

The results further showed that, compared to the DEU speakers, the CHN speakers produced vowels with comparable $f_0$ mean (both in Hz and ST) and maximum (in Hz), and they demonstrated similar intrinsic $f_0$ values among different vowels. As $f_0$ expressed in ST could neutralize anatomy-based acoustic differences while retaining phonemic differences, we analyzed $f_0$-related parameters in ST hereafter.
3.2 f0 range

The f0 ranges of each vowel produced by the DEU and CHN speakers are shown in Fig. 1. The effect of SpeakerGroup \([F(1, 784) = 12.71, p < 0.001]\) was significant, with the CHN speakers' target vowels having larger f0 ranges than those of the DEU speakers (3 St versus 1.75 St), which reflected the same trend as with the Hz scale (32 Hz versus 20 Hz). Furthermore, the f0 difference between maximum and mean in Hz for each vowel was consistently larger for the CHN speakers than that for the DEU speakers (Table 1). When f0 added range was represented in St, there was a significant effect for VowelIdentity \([F(13, 784) = 63.87, p < 0.001]\) but no significant effect for Gender \([F(1, 784) = 0.40, p = 0.526]\). Also, the f0 ranges of the tense vowels were considerably larger than their lax counterparts (3.05 St versus 20.15 St), indicating the significant effect of Tenseness \([F(1, 832) = 77.36, p < 0.001]\). In each German tense-lax vowel pair, the tense vowel has an inherently longer duration than its lax counterpart (215 ms versus 105 ms across pairs in this study). The smaller f0 ranges of lax vowels were deemed related to their shorter duration. However, whether the tense and lax vowels had the same rate of f0 change was still unclear. Therefore, we further compared the slope of f0 contours between tense and lax vowels.

3.3 f0 slope

The f0 slope was used to represent the dynamic changes of the vowel f0 contour, including the direction (rising or falling) and the steepness (i.e., f0 change rate). To analyze the directions of f0 contours, we measured the positions of the f0 maximum and minimum for each vowel. Fig. 2 depicts the f0 maximum/minimum position relative to vowel onset in percent,

![Fig. 1. Bar plots of average f0 range (in St) of German vowels produced by DEU and CHN speakers. F, female; M, male.](image-url)
where 0% and 100% correspond to the onset and offset positions of vowels, respectively. As can be seen from the figure, the patterns of $f_0$ contours are similar between the DEU-F and DEU-M speakers, that is, minimums preceded maximums, suggesting a rising trend in general. The CHN-F speakers exaggerated this pattern, since their $f_0$ minimums and maximums were closer to the onset and offset of vowels, respectively, compared to the DEU speakers. The vowels produced by the CHN-M speakers showed a reverse pattern, in which $f_0$ maximums occurred generally earlier than $f_0$ minimums, resulting in a roughly overall falling direction.

We further examined the steepness of the $f_0$ slope. The proportions of negative slopes were 31.24% and 24.76% for tense and lax vowels produced by the CHN female speakers, respectively. The CHN male speakers produced even more negative slopes with 54.67% and 58.1% of tense and lax vowels, respectively. Averaging these slopes resulted in a “cancellation” effect in which the negative and positive slopes nullified each other. Therefore, we plotted the absolute values of slopes in Fig. 3(a), in which the average duration was also included to reflect the interactions between the $f_0$ slope and duration of the vowels. All tense or lax tokens were averaged over seven vowels of each group. All lines were normalized to the same starting point for convenient comparisons, and the end points represent the average duration and slope. The LME regression for the absolute $f_0$ slope revealed that there were significant effects of SpeakerGroup [$F(1,832) = 8.64$, $p = 0.003$] and Tenseness [$F(1,832) = 51.63$, $p < 0.001$]. As can be seen from Fig. 3(a), the CHN speakers used a greater rate of $f_0$ change for the vowels bearing sentence stress than the DEU speakers ($20.8 \times 10^{-3}$ St/ms versus $11.8 \times 10^{-3}$ St/ms). Also, the lax vowels generally had shorter duration but steeper slopes ($18 \times 10^{-3}$ St/ms versus $14.6 \times 10^{-3}$ St/ms) than their tense counterparts. However, the effect of Gender [$F(1,832) = 1.5$, $p = 0.222$] was not significant, although the male speakers produced vowels with generally steeper $f_0$ ($18.2 \times 10^{-3}$ St/ms versus $14.4 \times 10^{-3}$ St/ms).
St/ms) contours than the female speakers (with the exception of lax vowels for the DEU speakers). There was also a significant interaction effect of SpeakerGroup × Tenseness [$F(1, 832) = 13.21$, $p < 0.001$]. As shown in Fig. 3(b), both speaker groups had greater rates of $f_0$ change for lax vowels than for tense vowels, and the tendency was stronger for the CHN speakers than for the DEU speakers.

4. Discussion

In our current study of $f_0$ profiles of German vowels under sentence stress, apart from the general comparison of $f_0$ mean and maximum, range, and slope, a particular focus was the tense-lax contrast of German vowels produced by Mandarin learners of German compared to native German speakers. The following new findings emerged.

First, after transforming $f_0$ from Hz to St with the speaker-specific base, the effect of Gender on $f_0$ mean was no longer significant, whereas the effect of VowelIdentity remained significant. It is also clear that CHN learners demonstrate similar intrinsic $f_0$ patterns as German native speakers under sentence stress, namely, high vowels are produced with an intrinsically higher $f_0$ than low vowels. We have thus provided further evidence to support the universality of intrinsic $f_0$ pattern in L2 speech. Furthermore, we have found that lax vowels are also associated with a higher $f_0$ mean than their tense counterparts in the same stressed context, especially for peripheral vowels. However, Schneeberg and Schlüffer (2006) examined 14 German vowels produced by six speakers and found that only one tense-lax vowel pair showed a significant difference with the lax vowel having a higher $f_0$, while other pairs showed no significant differences or tense vowels had significantly higher $f_0$ than lax vowels. Other studies suggested that lax vowels had about the same/similar (i.e., statistically non-significant) $f_0$ as their tense counterparts, for example, in Fischer-Jorgensen (1990) examining six vowels (/iː tː eː əː aː/) produced by six speakers and Pape and Mooshammer (2006) examining six vowels (/iː tː uː oː aː/) produced by three speakers. Whether and how the mixed findings result from different reading materials, individual speakers, or measuring approaches of $f_0$ remains to be examined in the future.

Moreover, we have shown that both CHN learners and DEU speakers produce a larger $f_0$ range for tense vowels than for lax vowels, which probably results from the longer inherent duration of tense vowels. We have also found that CHN learners produce vowels with a larger $f_0$ range than DEU speakers, which echoes the findings in the previous studies for Chinese learners speaking German or English (Ding et al., 2006; Zhang et al., 2008). Due to negative first language (L1) transfer, many CHN learners may attach a lexical rising or falling tone to the vowel, which may enlarge the $f_0$ range at the syllable level. More specifically, male CHN learners tend to produce more vowels with negative $f_0$ slopes, while female CHN learners tend to produce more vowels with positive $f_0$ slopes, which are more likely realized as lexical falling and rising tones in their L1 language, respectively. This difference may also result in much larger $f_0$ ranges of vowels for male CHN learners than for female CHN learners (see the bar heights in Fig. 1), which could be explained by the fact that the Mandarin high-falling tone has the largest $f_0$ range among lexical tones in general. In addition, the CHN learners produce 14 vowels with longer duration than the DEU speakers, 10 of which are statistically significant (Gao et al., 2020). The longer duration together with steeper slope may also contribute to the larger $f_0$ range of CHN learners’ vowels.

Finally, we have found that both DEU and CHN speakers produce lax vowels with greater steepness than tense ones, and CHN speakers increase steepness more than DEU speakers when they produce lax vowels. Having inherently shorter duration than their tense counterparts, lax vowels may require a larger $f_0$ change rate to achieve sentence prominence. Besides, we have shown that CHN learners produce the target vowels bearing sentence stress with different directions of $f_0$ slope. Like DEU speakers, CHN female speakers produce target vowels with an overall rising $f_0$ contour, while CHN male speakers produce those with an overall falling $f_0$ contour, which could be attributed to the negative L1 transfer of Mandarin Chinese. Though all CHN learners recruited for the study had comparable L2 German proficiency, we observed that the L2 German speech produced by the males was more Chinese-accented than that produced by the females. Their Chinese accent may result from their frequent use of high-falling tones to achieve the prominence of the target vowel. This is in line with the previous finding that Chinese students tend to use a falling tone to signal an English stressed syllable (Juffs, 1990), which supports Ohala’s argument that falling tones are more perceptually salient and can be accomplished quicker (Ohala, 1978). Similar explanations are also found in previous studies of L2 English, e.g., Mandarin learners use a sharply falling $f_0$ contour for strongly emphatic stress (Zhang et al., 2008).

Acknowledgments

This research work is partially sponsored by the China Scholarship Council and Shanghai Social Science Project (Grant No. 2018BYY003). We would like to express our gratitude to the anonymous reviewers and the editor for their constructive comments and suggestions to improve this work.

References and links


