Interspeech 2018
2-6 September 2018, Hyderabad

Pitch Characteristics of L2 English Speech by Chinese Speakers: A Large-scale Study

Jiahong Yuan, Qiusi Dong, Fei Wu, Huan Luan, Xiaofei Yang, Hui Lin, Yang Liu

Liulishuo Inc.

{jiahong.yuan, qiusi.dong, ferris.wu, lisa.luan, xiaofei.yang, hui.lin, yang.liu}@liulishuo.com

Abstract

AI-powered English learning apps are used by hundreds of millions of people across the globe on a daily basis. This presents a great opportunity for the study of L2 speech. On one hand, the amount of data accessible for research is very large and rapidly growing; on the other hand, new theories and understanding of L2 speech can be continually tested and revised through real-life and real-time applications.

This paper presents a study of pitch characteristics of L2 English speech using a large-scale dataset from a language learning app. Our dataset contains 180,000 spoken utterances which amount to 240 hours of speech. The results show that compared to L1, L2 English has narrower pitch range and slower rate of pitch change, but more small "ripples" on the pitch contour. The percentage of F_0 rise time is higher in L2, and the maximum F_0 in an utterance is realized later (with respect to the onset of the word on which the maximum F_0 resides). These results suggest that the influence of L1 on L2 prosody is more complex than previously demonstrated, and they shed light on L2 prosody assessment and learning.

Index Terms: L2 speech, prosody, pitch, large-scale phonetics

1. Introduction

With the explosion of social media, AI, and mobile technologies, second language (L2) teaching and learning is experiencing a revolution. Nowadays, AI-powered English learning apps are used by hundreds of millions of people across the globe on a daily basis. This presents a great opportunity for the study of L2 speech. On one hand, the amount of data accessible for research is very large and rapidly growing; on the other hand, new theories and understanding of L2 speech can be continually tested and revised through real-life and real-time applications. In this paper, we present an analysis of data from Liulishuo's English learning app, which has now more than 60 million registered users. As a first step toward a comprehensive study of L2 English speech by Chinese speakers, the analysis reported in this paper focuses on the pitch aspect of L2 English in comparison to L1.

Languages differ typologically in the way they use pitch. A predominant difference is, for example, between tone languages such as Chinese and stress languages such as English. How the typological differences are reflected in pitch characteristics of speech has been an interesting topic for research [1-5]. A number of studies investigated the difference between English and Mandarin in terms of mean pitch, pitch range, and pitch variation, but the results were inconsistent. [1] reported that the average rate of F_0 change, mean F_0, and F_0 fluctuations (peaks and valleys) were all greater in Mandarin than in English, but F_0 range was the same in the two languages. [4] reported that although the two languages' use of F_0 in single-word utterances was quite different, for a prose passage, they were more similar, differing only in the mean F_0, with Mandarin being higher. [5] reported that in broadcast news speech Mandarin has wider pitch range and more F_0 fluctuations than English.

Pitch characteristics of L2 speech have also been studied in the literature. It is generally agreed that the pitch form of L2 differs to some degree from what is considered the native norm [6-11]. Some studies have found a compressed pitch range and less pitch variation in L2 speech [6,8], which might be attributed to less confidence in L2 production. On the other hand, the influence/transfer of L1 prosody is also apparent in L2 speech. [10] reported that in terms of both pitch range on the phoneme level and pitch change amount on the utterance level, L2 English speech by Chinese speakers displayed a larger value than L1 English. [7] reported that the pitch range of content words is larger in L2 English speech by Japanese speakers than in L1 English. These results were interpreted as a demonstration of the negative transfer of L1 phonology. [11] reported that L1 Spanish speakers were more comparable than L1 Japanese speakers to native English speakers in the choice of pitch accent contour, but they both tended to realize the high tone (H*) significantly later than native speakers. Such results suggest that multiple factors are simultaneously responsible for the pitch characteristics of L2 speech.

Most of the previous studies of L2 speech used only a few hours of data. Because there is great variability in L2 speech, large datasets are desirable [12,13]. There is also a lack of method and validation for automatic analysis of pitch contours. In this study, we attempt to address these issues by increasing the amount data used for analysis by two orders of magnitude over most previously published work, and by exploring new methods for measuring the dynamics of pitch contours.

2. Data

The data were collected through a mobile app developed by Liulishuo to help users learn and practice spoken English. With the app, a user can read a sentence after listening to it from a L1 speaker, and get an automatic assessment of his/her speech.

Our dataset contains approximately 180,000 English utterances which amount to 240 hours of speech. The utterances were read from 4000 sentences. All sentences were statements and consisted of between 5 and 15 words. Each sentence was read by one (but not the same) L1 English speaker, and by 40 to 50 Chinese speakers. The assessment scores of the L2 utterances in the dataset were between 70 and 90 (out of 100), indicating intermediate to advanced degrees of proficiency in spoken English. The L2 speakers were from a variety of dialect regions across China with a total number of more than 44,000 (77% were female).

Word and phone boundaries were automatically obtained through forced alignment using Liulishuo's speech recognition engine, built on state of the art deep learning technology and thousands of hours of L2 English speech by Chinese speakers.

10.21437/Interspeech.2018-1556

3. Method

The analysis of the utterances consisted of two steps: pitch extraction for each utterance, followed by characterizing and measuring the pitch contour.

3.1. Pitch extraction

Pitch extraction was performed using Praat [14] with utterance-dependent pitch range settings, i.e., pitch floor and ceiling values (which are important for accurate pitch analysis in Praat [15]). For each utterance, the 75-500Hz range was first used to extract F_0s, from which we calculated the median value, and then used a pitch range setting of between one half and two times the median for final extraction. The analysis time step was 10 msec. To generate a continuous pitch contour, unvoiced regions were interpolated through. Finally, the pitch contour was smoothed using Praat's smoothing algorithm (frequency band = 10 Hz). Figure 1 illustrates two extracted pitch contours, one for L1 and one for L2.

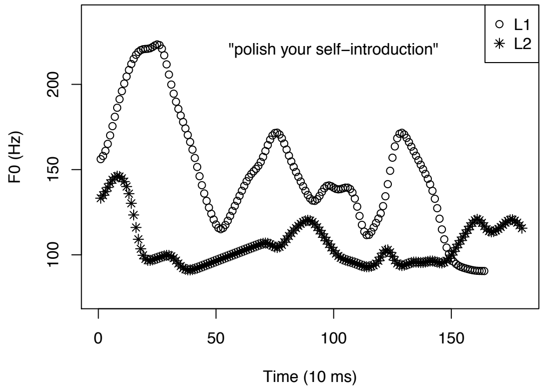

Figure 1: *Pitch contours of the sentence "polish your self-introduction" in L1 and L2.*

The extracted F_0 values were converted to semitones according to equation (1). The base frequency used for calculating semitones, *F_0_base* in the formula, was utterance dependent, which was the mean of F_0 values in the utterance.

$$Semitone = 12 * \log_2(\frac{F_0}{F_0\ base}) \qquad (1)$$

3.2. Detection of peaks and valleys from convex hull

Methodologically, two types of measures have been used to study pitch variation in L2, one based on descriptive statistics such as mean, range, and standard deviation, and the other based on specific landmarks such as pitch peaks and valleys.

We used a "convex hull" algorithm to find peaks and valleys in a pitch contour. The algorithm was proposed for a syllable segmentation task in [16] and was applied in [5] to study pitch characteristics. It finds the peak F_0 in an F_0 contour and constructs a convex envelope over the contour. On each side of the F_0 peak point, the differences between the convex envelope and the F_0 values are computed and the point that has the maximal difference is selected as a boundary, if the difference is larger than a pre-determined threshold value. Each side of the F_0 peak point is then divided into two subsegments at the new boundary, a convex envelope is constructed for each subsegment, and the procedure repeats recursively until no new boundaries can be found given the threshold value. Figure 2 is an example showing the peaks and valleys determined by the algorithm.

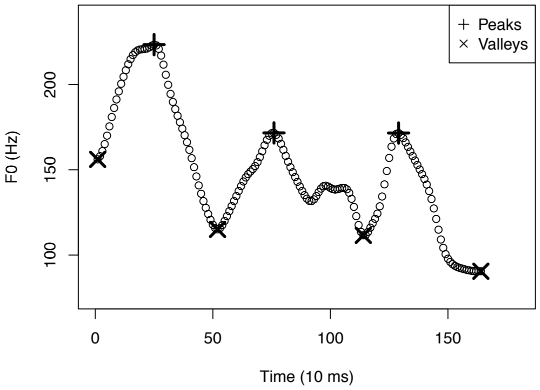

Figure 2: *Pitch peaks and valleys detected by convex-hull.*

3.3. Discrete Cosine Transform

Pitch contours can be decomposed into local and global components, comparable to small ripples riding on big waves [17]. Discrete Cosine Transform (DCT) provides a mathematical technique for such decomposition. In DCT a pitch contour is represented as a sum of cosine functions oscillating at different frequencies. Low-frequency components represent global modulation whereas high frequency components represent local fluctuations. In the literature, DCT has been applied in speech synthesis to model and generate F_0 contours [18,19].

4. Results

4.1. Overall pitch variation

Three variables were calculated for each utterance to characterize the overall pitch variation: 1. Pitch range: the difference between the maximum and minimum pitch values in the utterance; 2. Total amount of pitch change: the sum of the absolute difference between every two pitch values (corresponding to 10-msec interval); 3. Average of pitch change rate: the mean of the absolute difference between every two pitch values divided by 10ms interval. We calculated this for pitch falls (the next F_0 is lower than the current one) and pitch rises (the next F_0 is higher than the current one) separately.

The mean pitch range and mean total amount of pitch change for L1 and L2 are shown in Figures 3 and 4. In the figures utterances are divided into four groups based on the number of syllables in the utterance: 8 syllables or less (_08), 9 to 12 syllables (_12), 13 to 16 syllables (_16) and more than 16 syllables (16_). We can see that L2 English has both narrower pitch range and smaller total amount of pitch change, which means there is less pitch variation in L2 speech. Figure 1 above shows an example of these results.

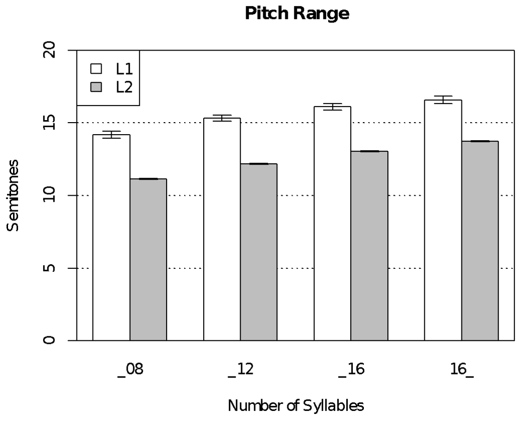

Figure 3: *Pitch range of L1 and L2.*

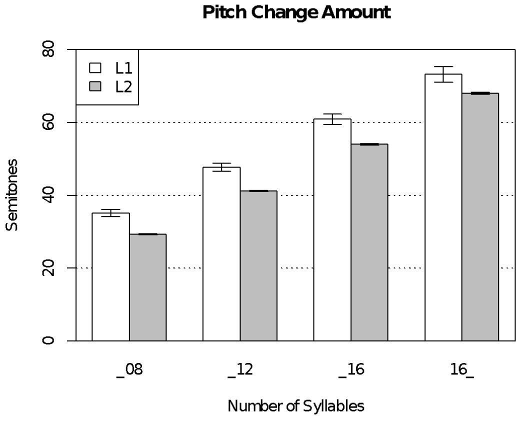

Figure 4: *Total amount of pitch change of L1 and L2.*

Figure 5 shows the average pitch rise and fall rates at different percentiles. Two results can be seen from the figure: First, L2 rises and falls slower than L1. Secondly, pitch falls faster than pitch rises in both L1 and L2, but the difference between the fall rate and the rise rate is smaller in L2.

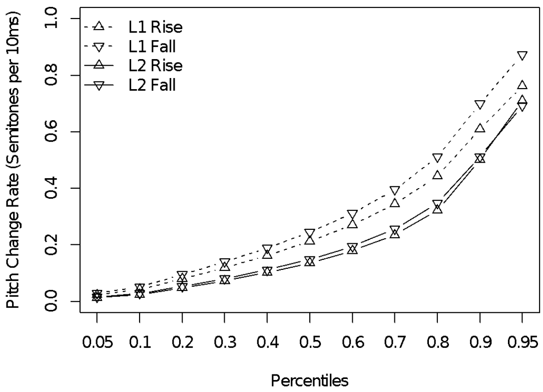

Figure 5: *Average rate of pitch change of L1 and L2.*

4.2. Big waves vs. small ripples

We need to predefine a threshold value when using convex-hull to find peaks and valleys in F_0 contours. Using a large threshold will only find turning points on big waves, whereas small fluctuations can be found with a small threshold. The smaller the threshold is, the more peaks and valleys will be found.

The average number of peaks and valleys per utterance, determined by different thresholds, are shown in Figure 6. Clearly, L2 English has more peaks and valleys when the threshold is small. That is, L2 English speech by Chinese speakers has more small fluctuations than L1 English (which can also be seen from the example in Figure 1).

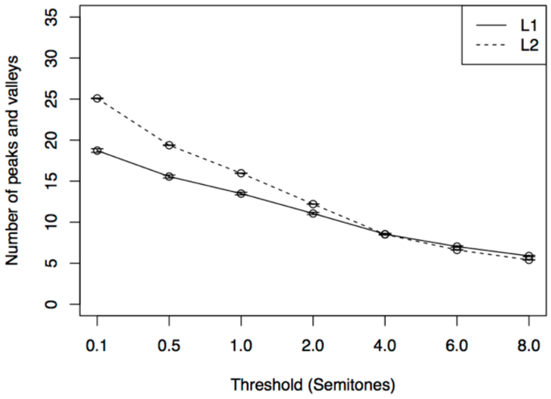

Figure 6: *Number of pitch peaks and valleys per utterance by convex-hull.*

The other method we used to separate big waves and small ripples in a pitch contour was to perform DCT to the contour. The mean log absolute values of the first 20 DCT coefficients are shown in Figure 7, for L1 and L2 respectively. We can see that L1 English has higher values of low-frequency coefficients, which means bigger global modulation, whereas L2 English has higher values of high-frequency coefficients, which means more small fluctuations.

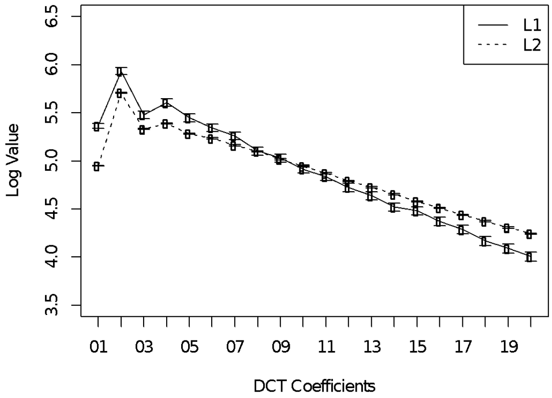

Figure 7: *Average log absolute values of the first 20 DCT coefficients for L1 and L2.*

4.3. Rise and fall time

Besides pitch variation, we also investigated the duration of pitch rise and fall. For every 10-msec interval in vowels, nasals, and glides, if F_0 at the end is higher than that at the beginning, then it's 10-msec of pitch rise; otherwise if F_0 at the end is lower, it's 10-msec of pitch fall.

The results show that within an utterance the time in which F_0 falls is longer than F_0 rises. This is true for both L1 and L2, arguably due to F_0 declination. However, the percentage of F_0 rise time is higher in L2 than in L1. The mean percentage of rise time is 42.59% in L1 and 45.54% in L2. The Q-Q plot in Figure 8 shows that the percentage of F_0 rise time has similar shape of distribution in L1 and L2, but higher in L2 across all samples.

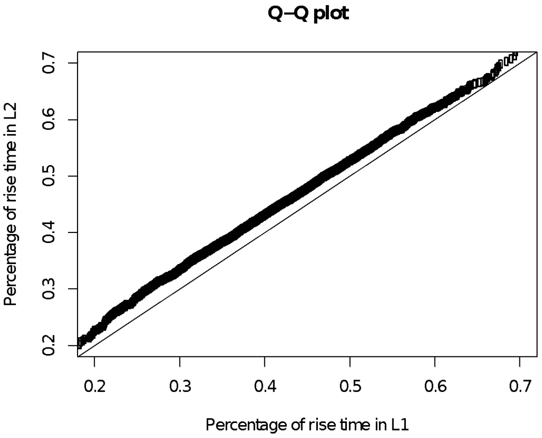

Figure 8: *Percentage of pitch rise time in L1 and L2.*

Finally, we measured how long the F_0 maximum in an utterance was realized from the onset of the word on which the maximum F_0 occurred, both in seconds and in percentage of the word duration. This time is often called "the alignment time". Because for the same sentence the F_0 maximum may appear in different words in L1 and L2 utterances, we calculated the alignment time on two sets of sentences. First, all sentences (and utterances) were used regardless of whether the F_0 maximum occurred on the same word in L1 and L2. Secondly, we used only the sentences for which L1 and at least five L2 utterances had the F_0 maximum on the same word ("shared" sentences) and measured those utterances. Among the 4000 sentences in the dataset, only 1700 were "shared" sentences. The results are listed in Table 1.

Table 1: *Alignment time of F_0 maximum.*

	All sentences	**"Shared" sentences**
L1	0.201s (52.32%)	0.217s (54.44%)
L2	0.259s (55.07%)	0.262s (56.47%)

We can see that it took longer time in L2 to reach the F_0 maximum from the onset of the word on which the F0 maximum resides, in terms of both duration (in seconds) and the percentage of time over the word duration. That is, the F_0 maximum was realized later in L2 than in L1.

5. Discussion

In our dataset, L2 English speech had a smaller pitch range and less amount of pitch change than L1 English. The average pitch change rate (both rise and fall) was also slower in L2. A plausible explanation for these characteristics is "too cautious to vary more", proposed in [8], which is that "[L2] speakers are less confident in their productions, therefore, they concentrate more on segments and words and subsequently refrain from realizing pitch range more native-like."

L2 speech by Chinese speakers had more pitch fluctuations than L1 English, demonstrated by both the greater number of small peaks and valleys and the larger magnitude of high-frequency components in a DCT of the pitch contour. This characteristic can be well explained by the transfer of L1 prosody in L2 production. In Chinese, every syllable has a tone and arguably a pitch target. Therefore, Chinese speakers may tend to produce most syllables with a pitch target in their L2 English, making more fluctuations in the pitch contour. Another possible explanation is that L2 speech is less fluent and has more hesitations and restarts, making the pitch contour less smooth. To test these hypotheses, we need to investigate pitch fluctuations in L2 speech by non-tonal L1 speakers.

A new finding from this study was that the percentage of F_0 rise time was higher in L2. This could be due to that L2 speakers are often unsure about their pronunciation, and therefore use more rising pitch ("frequency code" [20]). We also found that although pitch falls faster than pitch rises in both L1 and L2, the difference between the fall rate and the rise rate was much smaller in L2. Finally, the F_0 maximum in an utterance was realized later (with respect to the onset of the word on which the maximum F_0 resides) in L2 than in L1. These results suggest that the influence of L1 on L2 is more complex than previously demonstrated. As stated in [21], prosodic transfer in L2 learning merits more detailed analysis.

Empirically, the pitch characteristics of L2 speech found in this study can be convenient for assessment of L2 prosody and computer assisted prosody learning. We performed a preliminary study to classify L1 and L2 speakers using three features based on the above study: the percentage of F_0 rise time, the average rate of pitch rise, and the average rate of pitch fall. We selected 3600 L1 speakers and 9000 L2 speakers. Each L2 speaker has five utterances randomly selected from our dataset, and the feature values for each speaker are the average over the five utterances. We then built a SVM classifier using these features to classify L1 and L2 speakers. The classification accuracy (10-fold cross validation) was 92%. Further studies are needed to evaluate these features with other datasets.

6. Conclusions

We present a large scale phonetic study on pitch characteristics of L2 English speech by Chinese speakers. Our study demonstrates that L2 English has narrower pitch range, slower rate of pitch rise and fall, and smaller total amount of pitch change. It has, however, more pitch fluctuations than L1 English. The percentage of F_0 rise time is higher in L2, and the maximum F_0 in an utterance is realized later (with respect to the onset of the word on which the maximum F_0 resides). These results suggest that the influence of L1 on L2 prosody is more complex than previously demonstrated, and they also shed light on L2 prosody assessment and learning.

7. References

[1] S. J. Eady, “Differences in the f0 patterns of speech: Tone language versus stress language,” *Language and Speech*, vol. 25, pp. 29–42, 1982.

[2] M. Dolson, “The pitch of speech as a function of linguistic community,” *Music Perception*, vol. 11, pp. 321–331, 1994.

[3] I. Mennen, F. Schaeffler, and G. Docherty, “Cross-language differences in fundamental frequency range: a comparison of English and German,” *Journal of the Acoustical Society of America*, vol. 131, pp. 2249–2260, 2012.

[4] P. Keating and G. Kuo, “Comparison of speaking fundamental frequency in English and Mandarin,” *Journal of the Acoustical Society of America*, vol. 132, pp. 1050–1060, 2012.

[5] J. Yuan and M. Liberman, “F_0 declination in English and Mandarin Broadcast News Speech,” *Speech Communication*, vol. 65, pp. 67-74, 2014.

[6] U. Gut, “Foreign accent,” In C. Müller (Ed.), *Speaker classification*, pp. 75–87, 2007.

[7] K. Aoyama and S. G. Guion, “Prosody in second language acquisition: Acoustic analyses of duration and F0 range,” In O. S. Bohn and M. J. Munro (Eds.), *Language Experience in Second Language Speech Learning: In honor of James Emil Flege*, pp. 281-297, 2007.

[8] F. Zimmerer, J. Jügler, B. Andreeva, B. Möbius, and J. Trouvain, “Too cautious to vary more? A comparison of pitch variation in native and non-native productions of French and German speakers,” *Proceedings of Speech Prosody 2014*, pp. 1037-1041, 2014.

[9] I. Mennen, F. Schaeffler, and C. Dickie, “Second language acquisition of pitch range in German learners of English,” *Studies in Second Language Acquisition*, vol. 36, pp. 303-329. 2014.

[10] H. Ding, R. Hoffmann, and D. Hirst, “Prosodic Transfer: A Comparison Study of F0 patterns in L2 English by Chinese Speakers,” *Proceedings of Speech Prosody 2016*, pp. 756–760, 2016.

[11] C. Graham and B. Post, “Second language acquisition of intonation: the case of peak alignment,” *Journal of Phonetics*, vol. 66, pp. 1-14, 2018.

[12] H. Meng, C. Y. Tseng, M. Kondo, A. Harrison1, and T. Viscelgia, “Studying L2 Suprasegmental Features in Asian Englishes: A Position Paper,” *Proceedings of Interspeech 2009*, pp. 1715-1718, 2009.

[13] N. F. Chen, D. Wee, R. Tong, B. Ma, and H. Li, “Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL,” *Speech Communication*, vol. 84, pp. 46-56, 2016.

[14] P. Boersma and D. Weenink, *Praat: doing phonetics by computer* [computer program], Version 6.0.37, http://www.praat.org/.

[15] A. P. Vogel, P. Maruff, P. J. Snyder, and J. C. Mundt, “Standardization of pitch-range settings in voice acoustic analysis,” *Behavior Research Methods*, vol. 41, pp. 318–324, 2009.

[16] P. Mermelstein, “Automatic segmentation of speech into syllabic units,” *Journal of the Acoustical Society of America*, vol. 58, pp. 880-883, 1975.

[17] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkeley: University of California Press, 1968.

[18] J. Teutenberg, C. Watson, and P. Riddle, “Modelling and synthesising F0 contours with the discrete cosine transform,” *Proceedings of ICASSP 2008*, pp. 3973–3976, 2008.

[19] X. Yin, M. Lei, M., Y. Qian, F. K. Soong, L. He, Z. H. Ling, and L. R. Dai, “Modeling DCT parameterized F0 trajectory at intonation phrase level with DNN or decision tree,” *Proceedings of Interspeech 2014*, pp. 2273–2277, 2014.

[20] J. J. Ohala, “Cross-language use of pitch: an ethological view,” *Phonetica*, vol. 40, pp. 1-18, 1983.

[21] C. Y. Tseng and C. Y. Su, “Learning L2 Prosody Is More Difficult than You Realize – F0 Characteristics and Chunking Size of L1 English, TW L2 English and TW L1 Mandarin,” *Proceedings of Interspeech 2014*, pp. 1806-1810, 2014.