INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

F_0 Patterns of L2 English Speech by Mandarin Chinese Learners

Hongwei Ding[1], *Binghuai Lin*[2], *Liyuan Wang*[2]

[1]Speech-Language-Hearing Center, School of Foreign Languages
Shanghai Jiao Tong University, China
[2] Smart Platform Product Department, Tencent Technology Co., Ltd, China
hwding@sjtu.edu.cn, {binghuailin, sumerlywang}@tencent.com

Abstract

Prosodic speech characteristics are important in the evaluation of both intelligibility and naturalness of oral proficiency for learners of English as a Second Language (ESL). Different f_0 movement patterns between native and Mandarin Chinese learners have been an important research topic for second-language (L2) English speech learning. However, previous studies have seldom examined f_0 movement patterns between lower-level and higher-level Mandarin ESL learners. The current study compared f_0 change patterns extracted from the same 20 English sentences read by 20 lower- and 20 higher- level Mandarin ESL learners, and 20 native English speakers from a speech database. Appropriate procedures were applied to ensure a more accurate estimation of f_0 values and to catch characteristic deviation in f_0 movement patterns of ESL learners. The results showed that lower-level Mandarin speakers displayed more frequent f_0 fluctuations and smaller standard deviation of intervals between f_0 peaks than both native speakers and higher-level learners. The special characteristic of many smaller "ripples" on pitch contours of lower-level L2 English speech resembles Mandarin Chinese f_0 movements, which suggests a negative transfer from the first language (L1) Mandarin. The findings can shed light on the assessment and learning of L2 English prosody by Mandarin ESL learners.

Index Terms: F_0 pattern, L2 English speech, Mandarin Chinese Learners

1. Introduction

Mandarin Chinese and English are representative cases of tone and stress languages. The f_0 pattern in Mandarin is determined mainly by tone contours of all the lexical items, but the f_0 pattern in English is determined mainly by the phrasal stresses (i.e. pitch accents) on only a few of the lexical items in a sentence [4, 1, 2]. The pitch contour in Mandarin is an interaction of lexical tones and sentence intonation, which is compared to "small ripples on large waves" by Chao [3]. The investigations in the different f_0 patterns between Mandarin Chinese and English have been an interesting topic for experts. In the early 1980's, Eady found that Chinese speech was characterized by more f_0 fluctuations as a function of time and as a function of the number of syllables, which reflected the difference between a tone language like Chinese and a stress language like English [4]. In recent years, Keating and Kuo also showed that Mandarin speakers had higher f_0 maximums and means, and larger f_0 ranges, and they argued that the choice of speech materials to compare could be critical [5].

Moreover, it is suggested that the prosodic pattern of an adult's L2 can be characterized by that of acquired L1 [6, 7]. The rhythmic features of syllable-based Mandarin Chinese have been observed to be transferred in the L2 speech of a stress-timed language such as English [8]. However, not only rhythmic patterns of syllabic timing, but also patterns of fundamental frequency (f_0) and intensity can demonstrate the prosodic differences between L1 and L2 speech. But because of the complexities of pitch calculation, few investigations have been devoted to analyze the deviance in f_0 change patterns of L2 English by Chinese learners. Though Hirst and Ding [9] employed 18 acoustic metrics to compare 40 English passages and showed a correct prediction of 95.88% of L2 English speech by Chinese from that of native English, they did not show whether these metrics could distinguish higher- and lower-level learners.

Other investigations mainly compared f_0 profiles such as f_0 means and f_0 ranges, and mixed results have been found. Some studies have found compressed f_0 ranges and fewer pitch variations in L2 speech [10, 11], which could be attributed to less confidence in L2 production. Other studies reported larger f_0 ranges in L2 due to the influence of L1 prosody [12, 13]. As to f_0 means, the results were more various between L1 and L2 speakers due to different groups of speakers and different styles of speech (e.g. spontaneous or read speech). In the current study, we focused on the more robust effect of f_0 change patterns to provide relatively reliable measures for distinguishing different prosodic performances in L2 English production of Mandarin speakers. Instead of comparing f_0 means and ranges, we studied f_0 fluctuation patterns as described by Eady [4].

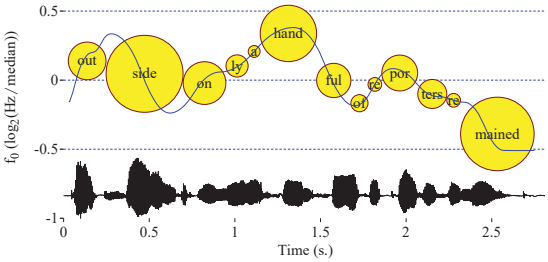
Normally stressed syllables are associated with longer duration and more or larger pitch movements [14]. Based on the rhythmic findings in timing that Mandarin L2 English learners display a smaller normalised pairwise variability index for vocalic intervals than native English speakers [15], we may hypothesize that L2 English produced by Chinese learners may demonstrate more f_0 peaks (small ripples) and lower variations in f_0 peak recurrence intervals, which could be transferred from the lexical tones in the native tone language.

2. Method

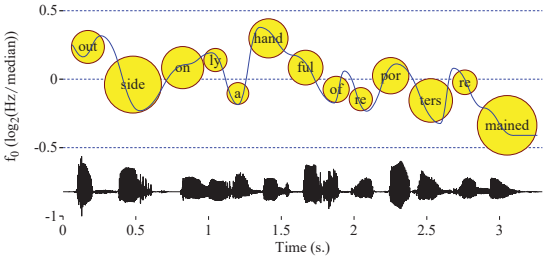
2.1. Speech database

The speech database was taken from the Global TIMIT Learner Simple English [16], which was part of the Global TIMIT project designed for acoustic-phonetic studies [17]. The learner database consists of two separate sets of 50 speakers reading 120 sentences from the original TIMIT [18]. These sentences were selected as *simple* to read by Chinese learners of English. Among the 120 sentences, 20 sentences were read by all speakers, 40 sentences were read by 10 speakers, and 60 sentences were read by five speakers. L1 Simple English was recorded at the University of Pennsylvania from 25 female and 25 male native American English speakers; L2 Simple English was recorded at Shanghai Jiao Tong University from 25 female and 25 male Chinese learners of English.

http://dx.doi.org/10.21437/Interspeech.2021-581



(a) Spoken by one L1 American English speaker



(b) Spoken by one L2 Mandarin Chinese speaker

Figure 1: *Prosody display of sentence "Outside only a handful of reporters remained."*

In order to divide the L2 group into higher and lower levels in prosodic performance, we asked five experienced English teaching assistants to rate the speakers. Based on the official criteria of fluency and coherence used in national standard tests, we made more detailed criteria for the evaluation of prosodic performance. We used 1-5 scales with 1 indicating very Mandarin Chinese-like prosody (i.e., very strong Chinese accent) and 5 indicating near-native English prosody (i.e., near-native accent). After several rounds of trials, the raters could achieve a minimal correlation of 0.8 in their assessments of randomly selected five sentences. For the best and worst 20 speakers, the raters could reach a total agreement. Then each rater should rate four different sentences of each speaker, and we averaged their scores and selected 20 speakers with the highest scores as *En*glish learners of *Hi*gher level (*EnHi*) and 20 speakers with the lowest scores as *En*glish learners of *Lo*wer level (*EnLo*), and left the other 10 speakers in the middle out. We also selected the 10 best L1 English speakers into *En*glish *Na*tive (*EnNa*) group according to the evaluations of our raters. Finally, we obtained three groups, the sex ratios were not exactly 1:1, but they were still balanced, with male:female ratio of 13:7, 9:11, and 12:8 in *EnNa*, *EnHi*, and *EnLo*, respectively.

2.2. Analysis procedure

2.2.1. Display of prosody

A visual inspection of f_0 movement patterns is important to gain an overall impression. With the help of ProZed [19], differences of f_0 patterns between L1 native English and L2 English by Mandarin Chinese are displayed in Figure 1.

The pitch contour is demonstrated in a continuous line with each circle corresponding to one syllable, with the vertical level and diameter of the circle representing the pitch and duration of the syllable respectively. The unit of pitch has already been normalized to the logarithmic scale $\log_2$*(Hz/median)*. A declarative intonation contour with phrasal stresses on a few syllables (circles with larger diameters) in L1 English is shown in Figure 1 (a); while a pitch contour with many fluctuations of circles in medium comparable diameters is displayed in Figure 1 (b). It can be observed that compared with L1 English, L2 English produced by the Chinese learner exhibited more minor stresses not only in syllable timing (e.g. many syllables with comparable duration) but also in f_0 movements (e.g., more fluctuations of f_0 contour). The f_0 movement of the L2 English resembles the Mandarin Chinese intonation of "smaller ripples" on the "large waves". We employed appropriate processing techniques to illustrate the differences between L1 and L2 groups, and to further separate the lower- from higher- level for L2 learners.

2.2.2. Analysis of f_0-related values

We applied different f_0 extraction software packages, and found that unreliable f_0 estimation mainly appeared at two kinds of places: 1) creaky periods, and 2) junction points of voiceless and voiced parts. By comparing the available f_0 extraction techniques, we found that the REAPER (Robust Epoch And Pitch EstimatoR) speech processing system [20] proved to be able to provide more accurate f_0 measurements within the creaky voice at low pitch ranges. To avoid extremely high and low values at junction points, we carried out the pitch tracking through a two-pass procedure following the strategy proposed by Hirst [21, 22]. First, we inspected our data and set a more accurate search range of 75-300 Hz and 100-400 Hz for male and female speakers respectively, and calculated the first and third quartiles (i.e., q1 and q3) across all f_0 samples for each speaker. Second, we set the f_0 floor and ceiling for each speaker to $0.75 \cdot q_1$ and $1.5 \cdot q_3$, respectively. By using a personalized search range, we greatly reduced the estimation errors of f_0 extraction, and the speakers' f_0 histograms were observed to be more centralized around the mean values. Finally, f_0 values were extracted using REAPER [20] with a five millisecond (ms) frame rate with personalized ranges to ensure optimized accuracy.

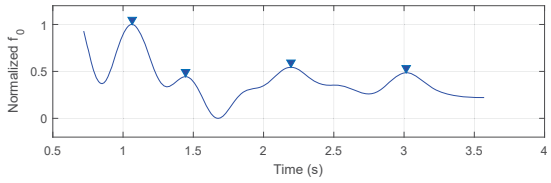
To reduce sex-based differences, f_0 values in Hertz (Hz) were converted to semitones (st). Instead of a fixed base frequently, we employed a speaker-dependent reference calculation (Equation 1) proposed by Yuan and Liberman [23] to further reduce the individual difference.

$$f_0[\text{St}] = 12 \cdot \log_2 \left(\frac{f_0[\text{Hz}]}{f_0\_base} \right) \quad (1)$$

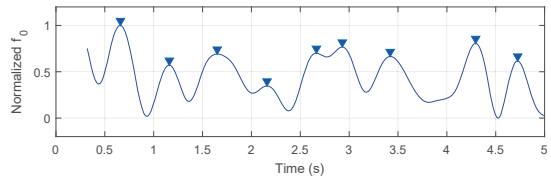
The base frequency used for calculating semitone (f_0_base) was speaker-dependent, which referred to the 5th percentile of all f_0 values of a given speaker.

2.2.3. Analysis of f_0 change patterns

We employed appropriate techniques to plot f_0 curves and identify f_0 peaks. To plot f_0 contours, we intended to compare f_0 change patterns that were related to perceptual prosody in the current study. Unlike the calculation of the variables by Eady in [4], which was based only on the voiced parts, our calculation was determined on the basis of the interpolated and smoothed pitch curves with voiceless parts of signal because the curve continues to evolve during the voiceless portions of speech, as it is explained by Hirst in [22]. Since we also compared the intervals between peaks, we excluded all silent periods longer than 50 ms, and employed the same strategy described by [23] for interpolation and smoothing. The f_0 contours were firstly linearly interpolated to be continuous over the unvoiced segments and creaky periods, and a Butterworth low-pass filter with normalized cutoff frequency at 0.1 with *filtfilt* was employed to smooth



(a) Produced by one L1 American English speaker



(b) Produced by one L2 Mandarin Chinese speaker

Figure 2: *F_0 curves with peaks identified of sentence "His technique is ample and his musical ideas are projected beautifully."*

the f_0 curves. To identify f_0 peaks, we employed topographic prominence analysis, which has been successfully used to detect syllabic peaks in identifying acoustic biomarkers from the repetitive syllable production task [24]. In our current study, a fluctuation was defined to be a point on the f_0 curve at which the slope of the line changes from a positive to a negative value or vice versa, as it was described by Eady [4]. MATLAB tools [25] were used to identify the fluctuation points (local peaks and valleys) in the f_0 contour. Since many micro f_0 movements (micromelody) caused by the segmental nature of the individual speech sounds were not associated with prosodic change patterns [22, 26], some f_0 perturbations due to segmental sounds should not be counted as prosodic fluctuations. The threshold value was manually tuned to set a moderate value of 0.5 semitone as the threshold for all categories of speech. That means in order for the local maximum or minimum point on the curve to be labelled as a fluctuation point, it had to differ from the immediately preceding minimum or maximum point by at least 0.5 semitone. The resulted number of peaks was comparable to those described in [23].

In order to facilitate comparison, we normalized the range of peaks and valleys of each sentence so that all f_0 contours appeared between 0 and 1, which was similar to the prosody display between -0.5 and +0.5 in Figure 1. For each sentence, a graphic display that plotted f_0 variations as a function of time was produced. We obtained altogether 1,200 plots for three groups: 1,200 = $20(speakers) \cdot 20(sentences) \cdot 3(groups)$. One example is shown in Figure 2. It is clear that fluctuations or peaks marked with little triangles in the same sentence occurred more frequently for the Chinese learner in Figure 2 (b) than for the American speaker in Figure 2 (a).

Since all the speakers read the same sentences, we first counted the number of peaks in each sentence, and then divided the peak number by the number of syllables and duration in seconds of each sentence. Thus we obtained two measures: *peak/syllable* and *peak/second*. We also specified two measures of the peaks: 1) *prominence*: how much the peak stands out due to its intrinsic height and its location relative to other peaks; and 2) *distance*: the distance between peaks. The mean and standard deviation of *prominences* and *distance* were also included as f_0-related parameters.

3. Results

3.1. Means of f_0 variables

The means of relevant f_0 information are presented in Table 1.

Table 1: *Means of relevant information of three groups*

Group	EnNa (L1)		EnHi (L2)		EnLo (L2)	
	Male	Female	Male	Female	Male	Female
f_0 (Hz)	121.13	200.02	127.96	216.35	122.37	218.34
f_0 (st)	4.32	5.61	5.33	5.61	4.05	6.61
duration (s)	2.77	2.99	3.65	3.60	3.89	4.24
peak counts	12.01	12.74	14.32	14.62	17.57	17.76
peak/syllable	0.744		0.878		1.078	
peak/second	4.337		4.011		4.416	
mean_prominence	0.251		0.250		0.246	
std_prominence	0.225		0.236		0.237	
mean_distance	0.090		0.078		0.063	
std_distance	0.041		0.039		0.030	

The values in each group were averaged across all sentences and speakers of the same category. As sex difference was no longer significant, the means were averaged across sexes to facilitate comparison among groups. The females had higher f_0 (Hz) means than males, and the Chinese speakers had relatively higher f_0 (Hz) means than the American speakers. Sex-based difference has been greatly reduced in f_0 means expressed in semitones. It is clear that *EnLo* was characterized with the largest values in sentence *duration*, *peak counts*, *peak/syllable*, and *peak/second*; the smallest values in *mean_distance* and *std_distance*; comparable values in *mean_prominence* and *std_prominence*.

3.2. One-way ANOVA across groups of f_0 changes

One way ANOVA and post-hoc test (Tukey's HSD) were run for each variable for f_0 changes separately across three groups and the obtained significance levels are shown in Table 2.

Table 2: *Significance levels for f_0 variables*

Variable	Significance level across groups		
	EnNa-EnHi	EnNa-EnLo	EnHi-EnLo
peak/syllable	***	***	***
peak/second	***	-	***
mean_prominence	-	-	-
std_prominence	***	***	-
mean_distance	***	***	***
std_distance	-	***	***

Note: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

It was found that *peak/syllable* and *mean_distance* displayed significant difference among three groups, and the most prominent difference was the *peak/syllable*. The peak counts in each sentence across three groups are displayed in Figure 3.

3.3. Binominal logistic regression analysis between groups

To further examine the contribution of each variable to the classification of speaker groups, we employed logistic regression analysis, in which all significant variables listed in Table 2 were entered as independent variables and group as binary dependent variable. Two binominal logistic regressions were performed, one for comparison between *EnHi-EnLo* and another between

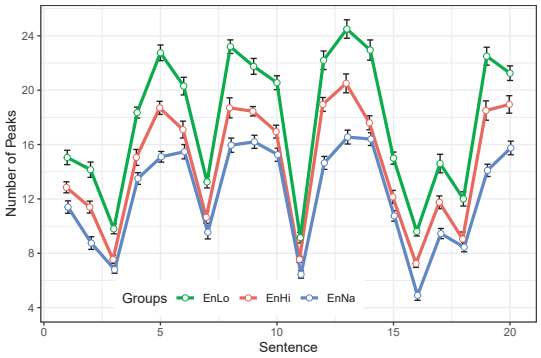


Figure 3: *Comparison of number of peaks among three groups.*

EnNa-EnHi. The number of powerful predictors was further reduced. The results of the logistic regressions are presented in Table 3. It is clear that compared with *EnHi*, *EnLo* was associated with higher *peak/syllable* and lower *std_distance*; while compared with *EnNa*, *EnHi* was related to higher *peak/syllable* and lower *peak/second* and *m_distance*.

Table 3: *Results of logistic regression for two comparisons.*

EnLo (vs. EnHi)			
Predictors	Coefficient	Z-value	p-value
peak/syllable	6.98	11.80	<2e-16 ***
std_distance	- 12.17	-2.35	0.019 *
EnHi (vs. EnNa)			
Predictors	Coefficient	Z-value	p-value
peak/syllable	11.39	12.96	< 2e-16 ***
peak/second	-2.12	-11.93	< 2e-16 ***
m_distance	-11.03	-4.27	1.92e-05 ***

Note: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Though *peak/second* and *mean_distance* could distinguish *EnLo* from *EnHi* when each variable was conducted separately, it did not show a significant difference when combined with other variables. While *std_prominence* could distinguish *EnHi* from *EnNa*, it became less powerful when other variables were added. The number of f_0 fluctuations (*peak counts*) decreased while the sentence duration increased from *EnLo* over *EnHi* to *EnNa*. Therefore, f_0 fluctuations as a function of number of syllables (*peak/syllable*) increased from *EnLo* over *EnHi* to *EnNa*, while f_0 fluctuations as a function of time (*peak/second*) increased from *EnLo* to *EnHi*, but decreased to *EnNa* because of the slow speech rate of L2 speakers.

The results in Table 3 showed that *EnLo* English could be improved by enlarging its deviation of peak recurrence intervals and slightly reducing f_0 fluctuations to resemble the prosody of *EnHi* English, while *EnHi* English could be further improved by reducing its f_0 fluctuations and enlarging the intervals between the peaks to approach the prosody of *EnNa* English.

4. Discussion

In line with our hypothesis, we have demonstrated that more f_0 peaks and lower variations in f_0 peak recurrence intervals are associated with Chinese-accent L2 English, which suggests a negative transfer of the L1 Mandarin language. Besides, some discussions are provided for the results in this study.

It is well acknowledged that f_0 *means (Hz)* of females are significantly higher than those of males. In our current study, L2 Chinese learners also demonstrated higher f_0 *means (Hz)* than those of American speakers, which is in line with the findings of many studies [12, 9] on Chinese learners due to L1 transfer. But there are opposite reports for other L2 learners [10] because they are too cautious to vary more. Actually, f_0 *means (Hz)* reflect individual differences and L1 backgrounds, and also depend on recruited speakers and selected materials. In the current study, short read English sentences instead of long difficult passages or spontaneous speech were selected, so L2 learners were confident enough to show a habitual larger pitch range in L2 as in their L1 language, which also resulted in larger f_0 *means*. By employing speaker-dependent reference frequency in converting f_0 from Hz to st, we maximally reduced cross-sex and cross-group differences, and provided a good basis for comparison of f_0 patterns between groups of different proficiency.

The invariant finding that L2 speech is slower than native L1 speech has been once again proved in the current study. Because of slower speech rate, higher-level learners even displayed fewer f_0 fluctuations as a function of time than L1 native speakers. However, we assume that f_0 patterns of the sentence should not change with the speech rate. Therefore, we could argue Chinese-accented English may be attributed to more f_0 fluctuations as a function of the number of syllables. More and regular small f_0 peaks can distinguish lower- from higher -level learners, and these f_0 patterns mirror the typical lexical tones of "small ripples" of Mandarin Chinese. However, we should be cautious to extend our findings to general contrast between L2 tonal language and L1 stress language. It is reported that L2 English by Vietnamese learners demonstrated fewer f_0 movements than the L1 English native speakers due to flatter pitch contours in the Vietnamese language than the English language [27]. That means more smaller f_0 fluctuations of L2 English may be a special feature for Chinese ESL learners. Keating and Kuo [5] argued that Mandarin high-falling tone alone can differentiate Mandarin Chinese f_0 movements from those of English. And because there are a high percentage of Tone 4 in Chinese texts [28], they can generate so many small ripples of f_0 movements in Mandarin. Moreover, Mandarin speakers apply their accustomed f_0 patterns in their L2 English, especially for lower-level speakers. And this phenomenon can be employed to evaluate the prosodic performance of Mandarin Chinese speakers. Considering that prosodic deviation of L2 speech is a complex interaction of a variety of rhythmic cues (including duration, f_0 and intensity) [29, 30], future work will endeavour to determine the cumulative effect of prosodic properties on the evaluation of L2 English speech by Mandarin Chinese learners.

5. Conclusions

The current study novelly employed the f_0 fluctuation variables to evaluate the L2 prosodic performance for Chinese ESL learners. The findings can provide implications for L2 English speech learning and teaching.

6. Acknowledgements

The work was jointly supported by Shanghai Social Science project (2018BYY003), and the Major Programs of National Social Science Foundation of China (18ZDA293, 15ZDB103 and 13&ZD189).

7. References

[1] M. E. Beckman and J. Edwards, *Articulatory evidence for differentiating stress categories*. Cambridge: Cambridge University Press, 1994, vol. Phonological structure and phonetic form: papers in laboratory phonology III, pp. 7–33.

[2] A. M. C. Sluijter and V. J. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 100, no. 2471, 1996.

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

[4] S. Eady, "Differences in the F0 patterns of speech: Tone language versus stress language," *Language and Speech*, vol. 25, pp. 29–42, 1982.

[5] P. Keating and G. Kuo, "Comparison of speaking fundamental frequency in English and Mandarin," *The Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1050–1060, 2012.

[6] J. Flege and R. D. Davidian, "Transfer and developmental process in adult foreign language and speech production," *Applied Psycholinguistics*, vol. 5, pp. 323–347, 1985.

[7] U. Gut, J. Trouvain, and W. J. Barry, "Bridging research on phonetic descriptions with knowledge from teaching practice – the case of prosody in non-native speech," in *Non-Native Prosody. Phonetic Description and Teaching Practice*. Mouton de Gruyter, 2007.

[8] A. Li and B. Post, "L2 acquisition of prosodic properties of speech rhythm: Evidence from L1 Mandarin and German learners of English," *Studies in Second Language Acquisition*, vol. 36, pp. 223–255, 2014.

[9] D. Hirst and H. Ding, "Using melody metrics to compare English speech read by native speakers and by L2 Chinese speakers from Shanghai," in *Interspeech*, 2015, pp. 1942–1946.

[10] 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," in *Proceedings of Speech Prosody*, 2014, pp. 1037–1041.

[11] J. Yuan, Q. Dong, F. Wu, H. Luan, X. Yang, H. Lin, and Y. Liu, "Pitch characteristics of L2 English speech by Chinese speakers: A large-scale study," in *Interspeech*, Hyderabad, September 2018, pp. 2593–2597.

[12] H. Ding, R. Hoffmann, and D. Hirst, "Prosodic transfer: A comparison study of F0 patterns in L2 English by Chinese speakers," in *Speech Prosody*, 2016, pp. 756–760.

[13] K. Aoyama and S. G. Guion, *Prosody in second language acquisition: Acoustic analyses of duration and F0 range*. John Benjamins, 2007, vol. Language Experience in Second Language Speech Learning: In honor of James Emil Flege, pp. 281–297.

[14] D. R. Ladd, *Intonational Phonology*. Publisher: Cambridge University Press, 2008.

[15] H. Ding, B. Lin, L. Wang, H. Wang, and R. Fang, "A comparison of English rhythm produced by native American speakers and Mandarin ESL primary school learners," in *Interspeech*, 2020, pp. 4481–4485.

[16] H. Ding, S. Liao, Y. Zhan, H. Feng, W. He, X. Hu, Y. Wu, J. Yuan, and M. Liberman. (2020) Global TIMIT learner simple English. Web Download. Philadelphia: Linguistic Data Consortium. [Online]. Available: https://doi.org/10.35111/zf5w-xq73

[17] N. Chanchaochai, C. Cieri, J. Debrah, H. Ding, Y. Jiang, S. Liao, M. Liberman, J. Wright, J. Yuan, J. Zhan, and Y. Zhan, "GlobalTIMIT: Acoustic-phonetic datasets for the world's languages," in *Proceedings of Interspeech*, 2018, pp. 192–196.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. (1993) TIMIT acoustic-phonetic continuous speech corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium. [Online]. Available: https://doi.org/10.35111/17gk-bn40

[19] D. Hirst, "Prozed: A speech prosody analysis-by-synthesis tool for linguists," in *Speech Prosody 2012*, 2012, pp. 15–18.

[20] D. Talkin. (2015) Reaper: Robust epoch and pitch estimator. [Online]. Available: https://github.com/google/REAPER

[21] D. Hirst, "The analysis by synthesis of speech melody: from data to models," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 55–83, 2011.

[22] D. Hirst and C. De Looze, *Fundamental Frequency and Pitch*. Cambridge, 2021, vol. The Cambridge Handbook of Phonetics, ch. 13.

[23] J. Yuan and M. Liberman, "F0 declination in English and Mandarin broadcast news speech," *Speech Communication*, vol. 65, pp. 67–74, 2014.

[24] B. Kashyap, M. Horne, P. N. Pathirana, L. Power, and D. Szmulewicz, "Automated topographic prominence based quantitative assessmentof speech timing in cerebellar ataxia," *Biomedical Signal Processing and Control*, vol. 57, 2020. [Online]. Available: https://doi.org/10.1016/j.bspc.2019.101759

[25] MathWorks, Inc., "MATLAB optimization toolbox," 2017, MATLAB and Statistics Toolbox Release 2017a, The MathWorks, Inc., Natick, Massachusetts, United States.

[26] C. De Looze and D. Hirst, "The ome (octave-median) scale: a natural scale for speech prosody," in *Proceedings of the 7th International Conference on Speech Prosody (SP7)*, N. Campbell, D. Gibbon, and J. Hirst, Eds. Trinity College, Dublin, Ireland, May 2014.

[27] A.-T. T. Nguyen, "F0 patterns of tone versus non-tone languages: The case of Vietnamese speakers of English," *Second Language Research*, vol. 36, no. 1, pp. 97–121, 2020.

[28] R. Hou and C.-R. Huang, "Robust stylometric analysis and author attribution based on tones and rimes," *Natural Language Engineering*, vol. 26, p. 49–71, 2020.

[29] E. Pellegrino, L. He, and V. Dellwo, "Computation of L2 speech rhythm based on duration and fundamental frequency," in *Elektronische Sprachsignalverarbeitung*, J. Trouvain, I. Steiner, and B. Mobius, Eds. Dresden: TUDpress, 2017, pp. 246–253.

[30] L. van Maastricht, T. Zee, E. Krahmer, and M. Swerts, "L1 perceptions of L2 prosody: The interplay between intonation, rhythm, and speech rate and their contribution to accentedness and comprehensibility," in *Interspeech*, Stockholm, Sweden, August 2017, pp. 364–368.