Perceptual Evaluation of Natural and Synthesized Prosody in Chinese Learners' English Production

Ying Chen, Yan Li, Wentao Xiao, Li Liu

School of Foreign Studies, Nanjing University of Science and Technology
ychen@njust.edu.cn

ABSTRACT

The prosody of English production by Mandarin Chinese learners was modeled and synthesized in terms of learners' L2 experience and compared with the production of native speakers of English. Both original speech and synthesized speech were perceptually evaluated by another group of native speakers of English to identify the focus status and rate the speech naturalness. The results reveal that natural speech was recognized and rated better than synthesized speech, more-experienced learners' speech better than less-experienced learners' speech, and narrow-focus sentences better than neutral-focus sentences. Among narrow foci, initial focus and medial focus were identified more accurately than final focus. Lexical stress of focused words showed a minor effect and interacted with focus status and speaker group. The overall results of the two perceptual evaluation tasks suggest a positive correlation between the accuracy of focus status identification and the speech naturalness rating in the English prosody produced by Chinese learners.

Keywords: L2 English, prosodic focus, synthesized prosody, speech naturalness, perceptual evaluation

1. INTRODUCTION

Focus refers to the emphasized word(s) in a sentence. Except for syntactic control, focus can be encoded phonetically with an increase of duration, pitch (F0) and intensity on the focused words, known as in-focus expansion, and a decrease of F0 and intensity, known as post-focus compression (PFC). Studies on bilingual production of prosodic focus have indicated that learners use duration and intensity more than F0 and in-focus expansion more than PFC to code focus [3, 4]. Even though Mandarin and English share the prosodic pattern with in-focus expansion of F0, intensity and duration and PFC of F0 and intensity, the L2 Mandarin production of L1 English learners and the L2 English production of L1 Mandarin learners reveal that learners have more difficulties in using F0 to code focus in L2, especially PFC of F0 [5].

F0, as the most difficult acoustical parameter for L2 learners to code focus, requires a theoretical model to explain its complexity in the prosodic realization of a sentence. The Parallel Encoding and Target Approximation (PENTA) model [6] is plausibly applicable to this. According to PENTA, target approximation parameters, including local pitch target, local pitch range, articulatory strength and duration, encode communicative functions in parallel and are sequentially and asymptotically realized by the surface F0. A quantitative target approximation (qTA) model [8] was proposed to use three model parameters to control the F0 trajectory of each syllable—target slope (m) and target height (b) specify the form of the pitch target and the rate or strength of target approximation (λ) indicates how rapidly a pitch target is approached. The equation of surface F0 is expressed as below:

$$f_0(t) = (mt + b) + (c_1 + c_2t + c_3t^2)\, e^{-\lambda t} \quad (1)$$

The transient coefficients, c_1, c_2, and c_3, are calculated based on the initial F0 dynamic state and the pitch target of the specified syllable:

$$c_1 = f_0(0) - b \quad (2)$$

$$c_2 = f'_0(0) + c_1\lambda - m \quad (3)$$

$$c_3 = (f''_0(0) + 2c_2\lambda - c_1\lambda^2)/2 \quad (4)$$

After a series revision of qTA model, a research tool PENTAtrainer2 [9], as a data-driven system, was developed to learn, model and synthesize the prosody of natural speech on the basis of hierarchically layered functional annotations.

The current study adopts PENTAtrainer2 to model and synthesize English sentences with prosodic focus produced by Chinese learners in terms of their L2 English experience. Native speakers of American English were recruited for a perception experiment to evaluate Chinese learners' original speech and group-synthesized speech in order to explore three research questions: (1) Can learners' acoustic realization of focus be recognized as well as native speakers' production by native listeners? (2) Can native listeners recognize focus in

the synthesized speech modelled by speakers' language experience? (3) How natural do native listeners find learners' original speech and its synthesized version compared with native speakers' production?

2. METHODS

2.1. The corpora and synthesis

Three corpora with 300 sentences in each produced respectively by native speakers of American English (AE), senior Chinese students (SC) and freshman Chinese students (FC), five females and five males in each group as presented in [12], were analysed and modelled. Table 1 shows the target sentences with their prompt questions.

Table 1: Prompt questions and answers.

Neutral focus	Question	What's the news?
	Answer	See initial, medial, final focus sentences below.
Initial focus	Question	Who may marry Ray?
	Answer	LEE/ NIna/ MElanie/ MaRIE/ RaMOna may marry Ray.
Medial focus	Question	What may Lee do to Norman?
	Answer	Lee may leave/ MArry/ NOminate/ reMIND/ reMEMber Norman.
Final focus	Question	Who may Ray marry?
	Answer	Ray may marry Lee/ NIna/ MElanie/ MaRIE/ RaMOna.

The natural speech productions were acoustically analysed by ProsodyPro [10], a Praat script for systematic prosody analysis. Syllable boundary labels were exported to PENTAtrainer2 [9] for annotation. Three functional layers were annotated: word stress (*A*, *Ab*, *aB*, *Abc*, and *aBc*), syllable position (word-final, sentence-final, non-final and monosyllabic) and focus condition (pre-focus, in-focus and post-focus) (cf. [8, 9]). The acoustic modelling was done by normalizing F0 heights across all speakers in each group. The many-to-one synthesis was done by imposing the acoustic model of the ten members in the same speaker group on each individual speaker's original speech.

2.2. The perceptual experiment

The mean standard deviation was calculated for the F0 values of all ten evenly spaced points in each syllable produced by all the 10 speakers in each group. The natural and synthesized speech of one male speaker and one female speaker, who displayed the median standard deviations in each group, was selected as the stimuli for the perceptual experiment. Thus 360 sentences (30 sentences * 2 speech types * 3 speaker groups * 2 speaker genders) were run and randomized by Praat ExperimentMFC. A forced choice identification task requested 24 native speakers of American English to choose the "emphasized" word (S, V and O respectively for subject, verb and object) or to judge that no word (N) was "emphasized" in the sentence. Meanwhile, the listeners were also required to rate the naturalness of the speech with a 1-5 scale (1 for poor and 5 for good). The experiment was conducted in a sound-attenuate booth, using a Lenovo desktop computer and a Sony MDR7506 dynamic headphone.

3. RESULTS

3.1. Focus status identification

Mixed-effects logistic regression model was applied to predict the response of focus status for the 360 sentences taking listener (n = 24) as a random factor, and speaker group [three levels: (1) AE, (2) SC, (3) FC], speech type [two levels: (1) natural, (2) synthesized), focus status [four levels: (1) neutral, (2) initial, (3) medial, (4) final] and word stress [five levels: (1) *A*, (2) *Ab*, (3) *aB*, (4) *Abc*, (5) *aBc*] as fixed factors. The first level of each factor was set as the reference level in the regression analysis. Results are shown in Table 2.

Table 2: Mixed-effects logistic regression results of focus status identification.

	β	***SE***	***z***	***p***
(Intercept)	-0.096	0.113	-0.849	0.396
SC group	0.347	0.078	4.446	0.000***
FC group	-0.283	0.072	-3.918	0.000***
Syn. speech	-0.117	0.052	-2.265	0.024*
Initial focus	1.852	0.077	24.012	0.000***
Medial focus	1.524	0.071	21.263	0.000***
Final focus	0.643	0.064	10.077	0.000***
Ab stress	-0.037	0.075	-0.491	0.624
aB stress	-0.189	0.075	-2.516	0.012*
Abc stress	-0.267	0.075	-3.559	0.000***
aBc stress	-0.125	0.075	-1.656	0.098

Surprisingly, the accuracy of focus status identification for the SC group was higher than that for the AE group; unsurprisingly, the accuracy of focus status identification for the FC group was lower than that for the AE group.

Results of the focus status identification task, by speaker group, are illustrated in conditional inference trees produced using the "party" package in R (version 3.5.0) [11] in Figures 1-3. The numbers on the branches correspond to the numbers of the levels in the fixed factors as in the above-

mentioned text. Bars in Figures 1-3 represent the percent accuracy of focus identification. The more robust the factor, the higher position in the tree branches. Therefore, focus type was the most significant factor in the results of focus status identification across speaker group.

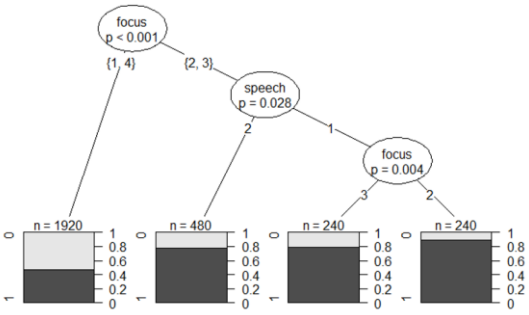

Figure 1: Focus status identification for the AE group.

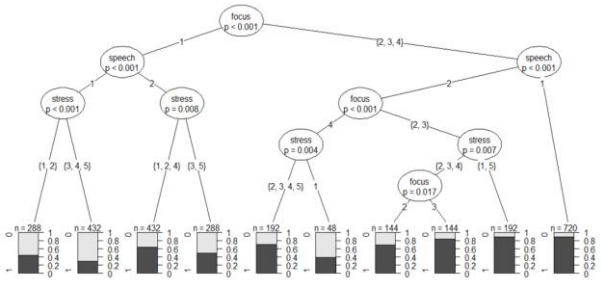

Figure 2: Focus status identification for the SC group.

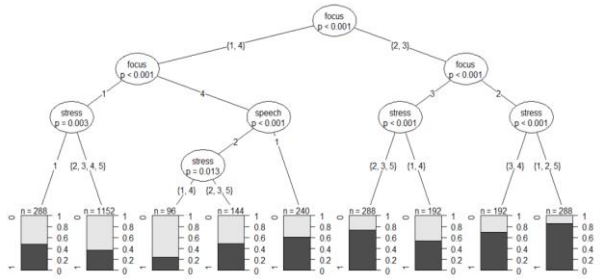

Figure 3: Focus status identification for the FC group.

3.2. Speech naturalness rating

Speech naturalness rating scores in 1-5 scales were transformed into *z* scores in terms of sentence in order to normalize subjects' individual differences. Mixed-effects linear regression model was employed to analyse native English speakers' rating to the learners' original speech and the synthesized speech based on their acoustical models by speaker group. The *z* score of the ratings was set as dependent variable. The random factor and the fixed factors in the independent variables were the same as in the logistic regression analysis for the focus status identification. The first level of each factor was again set as the reference level. The results of mixed-effects linear regression analysis are displayed in Table 3 (*df* = 26.67, 8606).

Table 3: Mixed-effects linear regression results of speech naturalness rating.

	ß	*SE*	*t*	*p*
(Intercept)	0.702	0.089	7.877	0.000***
SC group	-0.641	0.021	-31.069	0.000***
FC group	-0.915	0.021	-44.379	0.000***
Syn. speech	-0.513	0.017	-30.435	0.000***
Initial focus	0.137	0.024	5.733	0.000***
Medial focus	0.056	0.024	2.371	0.018*
Final focus	0.091	0.024	3.812	0.000***
Ab stress	0.051	0.027	1.926	0.054
aB stress	0.081	0.027	3.051	0.002**
Abc stress	-0.092	0.027	-3.443	0.001**
aBc stress	0.086	0.027	3.239	0.001**

Since both SC group and FC group showed lower speech naturalness than the AE group, another mixed-effects linear regression was conducted taking SC group as the reference level and found that the focus status identification result of the FC group was significantly different from that of the SC group [*ß* = -0.35, *SE* = 0.026, *t* = -13.31, *p* < 0.001].

The results of speech naturalness ratings are illustrated in Figure 4. The conditional inference tree shows that speaker group was the most significant factor and followed by speech type.

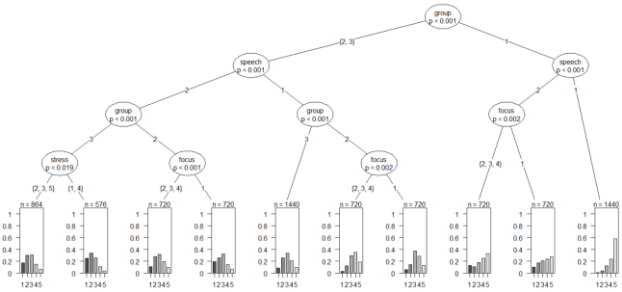

Figure 4: Speech naturalness rating. X- and Y-axes respectively represent the rating scale and percent.

4. FINDINGS AND DISCUSSION

The statistical results of focus status identification indicate that the production of the SC group was recognized better than that of the AE group and the AE group better than the FC group (see Table 2). The overall results answer the first research question that learners' focus production can be recognized by native listeners and more experienced L2 learners' production was recognized more accurately. Nonetheless, the reasons for the surprising result that the SC learners' focus production was recognized more accurately than the AE native speakers' production can be analyzed as follows. The low focus-identification accuracy for the AE group's production mainly lies in the neutral and final foci (see Figure 1). This could be due to

the fact that the boundary tone is located in the last content word of an intonation unit with either neutral focus or final focus [13, 14] and thus final focus did not demonstrate more robust prosodic prominence than neutral focus on the last tonic word in an intonation unit in English. The AE group, as native speakers, might have not intentionally produced final focus more salient than neutral focus, resulting in native listeners' perceptual confusion between final focus and neutral focus. The SC group, as advanced learners, might have learnt to specially emphasize on final focus, made it more prominent than neutral focus, which was at least reflected in duration in [12], and thus final focus was recognized more easily and accurately than neutral focus in their English production (see Figure 2).

The results of identifying focus statuses that were produced by the AE group also indicate that focus status is the most robust factor and followed by speech type (see Figure 1). The identification accuracy of initial and medial foci was significantly higher than that of neutral and final foci. This may be attributed to that initial and medial foci are auditorily more prominent than final and neutral foci. The acoustic analysis showed that both initial and medial foci were produced concomitantly with post-focus compress (PFC) of F0 and intensity in the original speech [12] and PFC was found to facilitate identifying focus status [2, 15]. Regarding the effect of speech type, original speech was identified more accurately than synthesized speech, suggesting no synthesis tool is perfect even though this corpus was native speech. Initial focus was identified more accurately than medial focus in the original speech, suggesting initial focus is more identifiable than medial focus. This could be again due to the effect of more words bearing PFC in the sentence with initial focus than in the sentence with medial focus and PFC works as an acoustic cue for recognizing focused words.

Similar to those of the AE group, the results of identifying focus statuses that were produced by the SC group indicate that focus is the most robust factor and followed by speech (see Figure 2). The identification accuracy of narrow focus, i.e., initial, medial and final foci, was significantly higher than that of broad focus, i.e., neutral focus. This result reflects that the SC group produced more robust final focus than neutral focus compared with the production of the AE group. Neutral focus in original speech was identified less accurately than in synthesized speech whereas narrow foci in synthesized speech less accurately than in original speech. In the synthesized speech with narrow foci, initial focus and medial focus were identified more accurately than final focus. These findings suggest that PENTAtrainer2 may have moderated the surface F0 of L2 learners' speech with certain focus statuses so that they became more identifiable. Lexical stress of the focused words is also a significant factor in focus status identification. However, no clear pattern was found across stress type and its effect interacted with that of speech type and focus status.

In the results of identifying focus status produced by the FC group, focus was the most robust factor again (see Figure 3). Overall, neutral and final foci were clustered together, significantly different from the cluster of initial and medial foci. Although the accuracy was lower in the results of the FC group than that of the AE group, initial and medial foci were again identified more accurately then final and neural foci. However, final focus was identified better than neural focus and initial focus better than medial focus. These results reveal that even less-experienced learners produced narrow focus more prominent than neutral focus and PFC might have also played a role in the identification of initial focus. Speech type only affected the identification of final focus produced by the FC group. Lexical stress of focused words was also a significant factor in focus status identification but again did not demonstrate clear patterns.

The above analyses answer the second research question that native listeners were able to recognize focus in the speech synthesized based on speaker-group modelling but the accuracy was not as high as that in the original speech though the effect of speech type interacted with the effects of focus status and lexical stress of the focused words.

Figure 4 illustrates that speaker group is the most salient factor of the speech naturalness rating. The AE group's production was rated higher than the production of the SC and FC groups and the SC group's production higher than the FC group's production in both original speech and synthesized speech. These results answer the third research question of the comparison between learners' and native speakers' production. The original speech was generally rated more naturally than the synthesized speech, again suggesting no synthesis tool is perfect. Sentences with initial, medial and final foci were rated more naturally than those with neutral focus even though the audio tokens were presented to the raters with no discourse context. Lexical stress was a minor effective factor in the FC group's synthesized speech.

5. ACKNOWLEDGMENT

This work was supported by the National Social Science Foundation of China (approval number 19BYY0430).

6. REFERENCES

[1] W. Wu and Xu. Y, "Prosodic focus in Hong Kong Cantonese without post-focus compression," *Proc Speech Prosody 100040*, pp. 1–4, 2010.

[2] Y. Xu, S. Chen, and B. Wang, "Prosodic focus with and without post-focus compression: A typological divide within the same language family?" *The Linguistic Review*, vol. 29, no. 1, pp. 131–147, 2012.

[3] Y. Chen, Y. Xu, and S. Guion-Anderson, "Prosodic realization of
focus in bilingual production of Southern Min and Mandarin," *Phonetica*, vol. 71, no. 4, pp. 249–270, 2014.

[4] W. Wu and L. Chung, "Post-focus compression in English Cantonese bilingual speakers," *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, pp. 148–151, 2011.

[5] Y. Chen, "The hierarchies of phonetic realization of focus in second language speech," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 2160–2160, 2016.

[6] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, pp. 220–251, 2005.

[7] Y. Xu and Q. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Communication*, vol. 33, pp. 319–337, 2001.

[8] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *Journal of the Acoustical Society of America*, vol. 125, pp. 405–424, 2009.

[9] Y. Xu and S. Prom-on, "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning," *Speech Communication*, vol. 57, pp. 181–208, 2014.

[10] Y. Xu, "ProsodyPro.praat," Version 5.5.2. http://www.phon.ucl. ac.uk/ home/yi/ProsodyPro/, 2014.

[11] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 307, pp. 1–11, 2008.

[12] Y. Chen, "Post-focus compression in English by Mandarin learners," *Proceedings of the 18th International Congress of Phonetic Sciences*, vol. 287, pp. 1–5, 2015.

[13] C. Gussenhoven, *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press, 2004.

[14] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 2006.

[15] Y. Xu, C. Xu, and X. Sun, "On the temporal domain of focus," *Proceedings of International Conference on Speech Prosody*, pp. 81–84, 2004.

[16] Y. Xu, A. Lee, S. Prom-on, and F. Liu, "Explaining the PENTA model: A reply to Arvaniti and Ladd," *Phonology* 32: 505-535, 2015.

[17] Y. Xu, S. Prom-on, and F. Liu, The PENTA model: Concepts, use and implications. In *Prosodic Theory and Practice*. S. Shattuck-Hufnagel and J. Barnes (eds.). Cambridge: The MIT Press. pp. 377-407, 2022.