*INTERSPEECH 2017*
August 20–24, 2017, Stockholm, Sweden

# Prosody Analysis of L2 English for Naturalness Evaluation through Speech Modification

*Dean Luo*[1], *Ruxin Luo*[2], *Lixin Wang*[3]

[1] School of Electronic Communication Technology, Shenzhen Institute of Information Technology, China,
[2] School of Applied Foreign Languages, Shenzhen Polytechnic, China,
[3] Shenzhen Seaskyland Technologies, China
luoda@sziit.edu.cn, luoruxin@szpt.edu.cn, wlx@seaskylight.com

## Abstract

This study investigates how different prosodic features affect native speakers' naturalness judgement of L2 English speech by Chinese students. Through subjective judgment by native speakers and objectively measured prosodic features, timing and pitch related prosodic features, as well as segmental goodness of pronunciation have been found to play key roles in native speakers' perception of naturalness. In order to eliminate segmental factors, we used accent conversion techniques that modify native reference speech with learners' erroneous prosodic cues without altering segmental properties. Experimental results show that without interference of segmental factors, both timing and pitch features affect naturalness of L2 speech. Timing plays a more crucial role in naturalness than pitch. Accent modification that corrects timing or pitch errors can improve naturalness of the speech.

**Index Terms**: naturalness, pitch, timing, prosodic assessment

## 1. Introduction

Second Language (L2) speakers often speak the target language with a foreign accent. Researchers working on L2 foreign accent have investigated speech properties that affect the perceived degree of foreign accent, including prosodic features [1, 2]. The impact of prosodic features on naturalness has been acknowledged both in teacher belief and pronunciation research [3]. Among world languages (except sign language), two major prosodic features – timing and pitch – are coordinated to constitute the rhythm of languages by their phonological rules. Prosodic features are a key component of natural and intelligible speech, and thus need to be put under examination to find out exactly which features strongly affect native speakers' judgment of L2 speech.

In [4, 5], timing and pitch have been found to play crucial roles in native speakers' perceptual naturalness of L2 Mandarin Chinese and English. Although in the previous experiments, L2 speech data was carefully chosen to contain only prosodic errors, segmental pronunciation properties such as phone-level goodness of pronunciation that can affect native speakers' perception of naturalness [6] were not investigated. In order to examine how different prosodic features affect naturalness of L2 speech, the interference of segmental pronunciation properties need to be eliminated. One possible solution is using voice conversion technology to alter specific segmental or prosodic characteristics while preserving other properties. Recently, accent morphing have been proposed in the domain of L2 intelligibility and prosodic acquisition [7-10]. In [8], the authors modified accents while maintaining speaker identities to improve intelligibility of foreign-accented speech. [9] and [10] used accent reduction techniques to provide L2 learners with their own converted speech as a reference for speaking skills training. We consider the influence of segmental pronunciation quality features on native listeners' naturalness perception of L2 speech could be excluded by using accent conversion techniques that modifies only pitch and timing of the native speech while maintaining other properties including timbre of the voice to represent L2 learners' prosodic error patterns.

In this study, we first examine how segmental and prosodic features affect native listeners' naturalness judgment of L2 English. An accent conversion technique that modifies timing and pitch is used to replace the prosodic cues of the reference native speech with learners' erroneous ones. We then investigate the roles of different prosodic features in the naturalness of English L2 speech both through subjective judgment by native speakers and through objectively measured prosodic features. Finally, we use the same speech modification technique to correct learners' prosodically erroneous speech with prosodic cues from the native reference speech and investigate the improvements in naturalness.

## 2. Data

We use L2 English read-aloud speech spoken by Chinese students using the dubbing practice software as in our previous study [5].

### 2.1. Speakers

Thirty high school and college students who are native Mandarin speakers participate as speakers in our experiments. There are 15 females and 15 males with different degrees of proficiency.

### 2.2. Material and recording procedure

Before recording, a two-minute long clip of a male native English speaker telling a story was presented to the participants twice. The content was carefully chosen so that it contained words that Chinese students tend to make stress errors according to [11], but the degree of vocabulary difficulty did not exceed the level of a typical high school student according to the curriculum. The text information was added to the video as subtitles. During the recording, the subtitled video with no sound tracks was presented to the learners and they were required to read aloud the subtitles and try to match the lip movements of the native speaker on the video.

http://dx.doi.org/10.21437/Interspeech.2017-332

### 2.3. Data selection and categorization

Utterances that contain pitch related stress errors or timing errors or both, with no obvious segmental errors, were chosen, together with a correct utterance from the recorded data. The two error types are defined as below:

1) Stress error: incorrect stress considering both word stress and sentence stress.
2) Timing error: includes untimely pauses between syllables or words, and unnaturally lengthening or shortening of vowels.

The judgment of errors was made by two phonetically trained English instructors and acoustic analysis.

The following four patterns were considered.

1. incorrect stress, correct timing -- SxTo
2. correct stress, incorrect timing -- SoTx
3. correct stress, correct timing -- SoTo
4. incorrect stress, incorrect timing -- SxTx

Thus, six sentences that contain the listed error patterns were chosen and formed 24 stimuli for our perceptual experiment.

## 3. Prosodic modification

The time domain pitch synchronized overlap-add algorithm (TD-PSOLA) [12] is a well-known speech synthesis method for high quality pitch and time scale modification. The sound quality of TD-PSOLA modified speech is very sensitive to a proper positioning of the pitch marks that delimit the individual pitch epochs [13]. In [14], a high precision pitch marking algorithm for TD-PSOLA has been proposed. We adopted this algorithm for pitch and timing modification in our experiments. Since the modification method is based on decomposition of the speech signal into overlapping pitch synchronous frames and the modification of pitch or duration is obtained by duplication or decimation of some frames from the original speech without destroying the coherence of the signal for each frame, the segmental goodness of pronunciation and the timbre of the voice can be preserved.

In our experiments, we first used HMM-based forced alignment and the pitch marking method mentioned above to analyze timing and pitch of the reference speech and L2 speech. Timing and pitch marks were carefully examined and corrected manually by a phonetically trained expert for accuracy. The prosodic modification method based on TD-PSOLA was then applied on the native reference speech to replace timing and pitch cues with those of the 24 L2 utterances mentioned in section 2.3. The modified native speech was also used to form another 24 stimuli. For 18 L2 utterances that contain prosodic errors (i.e. 6 from each of SxTx, SoTx and SxTo), we applied the same technique to modify erroneous pitch or timing patterns with the correct ones from the native speech and generated another 18 prosody-corrected speech. Altogether, there are 66 stimuli.

## 4. Perceptual experiment

### 4.1. Listeners

Altogether 24 native speakers of English (12 males and 12 females) living in Shenzhen (China) without any hearing impairments were recruited as the subjects in the perceptual experiment.

### 4.2. Procedure

The subjects were asked to listen to each stimulus and score the naturalness of the stimulus on a 7-point Likert scale.

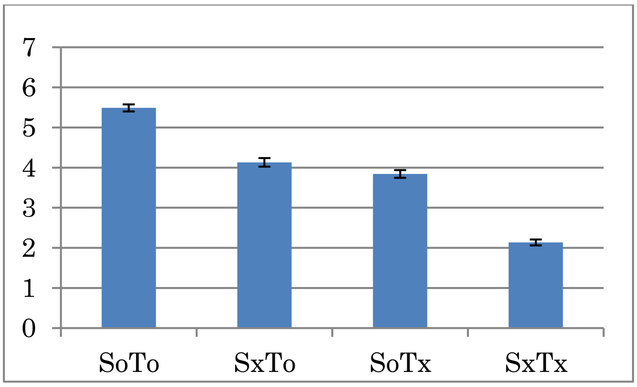

Figure 1: *The average naturalness score (+/- standard error) of L2 speech with four types*

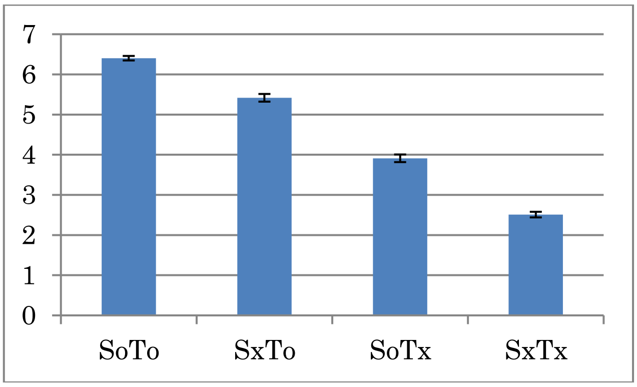

Figure 2: *The average score (+/- standard error) of prosody-modified native speech with four types*

The task was conducted online with the following prompt: "You will listen to 66 utterances. Please judge the naturalness of the utterances by choosing the appropriate scores from 1-7. Make sure to give a higher score to an utterance of higher quality (thehighest score is 7, indicating native-like quality) and a lower score to an utterance of poorer quality. There is no correct or wrong answer. You are allowed to listen to each speech sample only once. Just follow your intuition as a native speaker. Before the experiment begins, four training utterances will be played for you to get familiar to the material."

Before the perceptual experiment, a training session was conducted using four training utterances until the subjects got familiar with the task. Then the 66 stimuli were presented in a randomized order to the subjects for their subjective scoring. Each stimulus could be listened to only once.

### 4.3. Results

The means and the standard errors of the scores of subjective judgment on original L2 data were calculated from 144 stimuli (i.e., 6 utterances × 24 subjects) for each error pattern. As shown in Fig. 1, the mean scores for the four patterns are in the ordering of SoTo > SxTo > SoTx > SxTx. This suggests that the utterances with only timing errors tend to receive worse evaluation than those with only stress errors. Timing is more crucial a factor on native speakers' naturalness judgement of the utterances.

For comparison, the scores of subjective judgment on modified native speech generated by modifying prosodic features of reference native speech to represent the same L2

Table 1: *Average naturalness score of L2 speech and Prosody-corrected speech*

| Error patterns | *L2 speech* | *Prosody-corrected speech* | *Increases* |
|---|---|---|---|
| SxTx | 2.0 | 3.9 | 1.9 |
| SoTx | 3.8 | 5.3 | 1.5 |
| SxTo | 4.1 | 5.0 | 0.9 |

prosodic error patterns are shown in Figure 2. Although the average score of each group is higher than the original L2 speech, the same ordering (i.e SoTo > SxTo > SoTx > SxTx) of the mean scores has been observed and the differences between timing-related and pitch-related groups are more significant than those of the L2 speech. This suggests that without interference of segmental pronunciation features, timing play a more significant role on naturalness judgement.

We also investigated the improvements of correcting prosody errors in L2 speech by modifying erroneous pitch and timing cues of the L2 speech with those of the reference native speech. As shown in Table 1, the average naturalness scores of all the 3 error pattern groups have increased. The increases of SxTx and SoTx are larger than SxTo. This indicates that the correction of timing-related errors improves the naturalness of L2 speech more than the correction of pitch-related stress-only errors.

# 5. Acoustic measures

## 5.1. Measures based on segmental features

The confidence-based pronunciation assessment, which is defined as the Goodness of Pronunciation (GOP), is often used for assessing speakers' articulation and shows good results [15]. GOP scores is defined as follows.

$$GOP(p) = P(p \mid o) = \frac{P(o^{(p)} \mid p)P(p)}{\sum_{q \in Q} P(o^{(p)} \mid q)P(q)} \quad (1),$$

where $P(p \mid o)$ is the posterior probability that the speaker uttered phoneme $p$ given speech $o$ , $Q$ is the full set of phonemes.

[16] proposed a better implementation of GOP by calculating the average frame posteriors of a phone with the output of DNN model:

$$GOP(p) = P(p \mid t_s, t_e; o) = \frac{1}{t_s - t_e} \sum_{t=t_s}^{t_e} P(s_t \mid o_t) \quad (2),$$

where $P(s_t \mid o_t)$ is an output of DNN and $t_s$, $t_e$ are the start and end frame of phone *P*. GOP score calculated in this way is used as one of the objective feature scores to indicate intelligibility.

## 5.2. Measures based on different prosodic features

According to [17], stress is the result of interaction of pitch, intensity, and duration. Other prosodic features such as pauses, articulation rate, start and end time of phones or syllables, are related to timing. We measured these objective features and examined their roles in naturalness judgement of L2 utterances.

### *5.2.1. Distance in F*$_0$

Generally speaking, the more similar the F0 contour of the L2 utterance is to that of the presented native utterance, the more natural the L2 speech tends to be. Hence, we define a measure of F0 distance to characterize the naturalness of L2 speech.

Using Praat, F0 values of the speech were extracted at every 5ms with a time window of 20ms. The F0 values were then smoothed and interpolated to produce a continuous F0 contour. In the present work, F0 was measured in the logarithmic scale and normalized so that speaker differences can be minimized.

The word boundaries in the speech were detected by forced alignment using the HMM monophone acoustic models trained on the WSJ corpus. If a word has altogether *I* samples of F0 in the native utterance and J samples of F0 in the L2 utterance, we can define the Dynamic Time Warping (DTW) distance in F0 for this word between the two utterances as

$$D(native, L2) = \frac{g(I, J)}{I + J} \quad (3).$$

Here g(*I, J*) is calculated in an iterative way:

g(1, 1) = d(1, 1);

$$g(i,j) = \min \begin{Bmatrix} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i,j-1) + d(i,j) \end{Bmatrix}, i > 1 \text{ or } j > 1, \quad (4)$$

where d(*i, j*) indicates the F0 difference between the *i*-th sample of F0 in the native utterance and the *j*-th sample of F0 in the L2 utterance.

The F0 distance between two utterances can then be defined as the average F0 distances for all words in the sentence.

### *5.2.2. Percentage of pauses*

It is recognized that non-native speech tends to have more pauses than native speech. Hence, the percentage of pauses in the utterance can be used to characterize the naturalness of L2 speech. Silent pauses in the speech were detected by hmm forced alignment mentioned in *5.2.1.* Percentage of pauses in the utterance was then calculated by the ratio of the duration of silent pauses in the utterance to the duration of the entire utterance.

### *5.2.3. Average phone duration*

We use the average phone duration as defined below to characterize the naturalness of L2 speech:

$$PhoneDur = \frac{D_{utterance} - D_{pauses}}{N_{phones}} \quad (5),$$

where $N_{phones}$ is the number of phones in the utterance, $D_{utterance}$ is the duration of the entire utterance, and $D_{pauses}$ is the duration of silent pauses in the utterance.

### *5.2.4. Syllable time difference*

Since learners are required to reproduce the original speech by matching the native speaker's lip movements on the video as perfectly as possible, ideally the start and end time of each syllable should be the same as the original speech. Therefore, the time difference between learners' speech and the original native speech can be an indicator of the 'goodness' of matching (or timing). The boundaries of each syllable can be obtained

Table 2. *Correlations between each objective measure and the subjective score of L2 speech*

| Prosodic measure | Absolute correlation (L2 speech) | Absolute correlation (modified speech) |
|---|---|---|
| GOP | 0.59 | 0.02 |
| $F_0$ Distance | 0.43 | 0.60 |
| Duration | 0.10 | 0.12 |
| Pause | 0.36 | 0.39 |
| Syllable time | 0.52 | 0.69 |

through phone-level forced alignment using HMM acoustic models mentioned in *5.2.1*.

The syllable time difference feature is defined as,

$$\mathrm{D}_{\mathrm{syllable}}(native, L2) = \frac{\sum_{i=1}^{N}( |S_i(native)-S_i(L2)|+|E_i(native)-E_i(L2)| )}{N} \quad (6),$$

where $S_i(native)$ and $E_i(native)$ are the start time and end time of *i*-th syllable of the original native speech. $S_i(L2)$ and $E_i(L2)$ are the start time and end time of *i*-th syllable of the L2 speech, and *N* is the number of syllables in an utterance.

### 5.3. Objective acoustic feature analysis

Using the same 48 L2 and prosody-modified native utterances as in the perceptual experiment, we calculated the correlations between each aforementioned segmental and prosodic measures and the scores of subjective judgment.

As shown in Table 2, with original L2 speech, GOP shows the highest correlation, which indicates that segmental goodness of pronunciation also plays a significant role in native listeners' perception of naturalness on L2 speech. When using prosody-modified native speech that represents L2 learners' different prosodic errors, GOP does not correlate with naturalness perception, which indicates that the influence of segmental goodness of pronunciation has been excluded. "Syllable time", which indicates the timing difference between L2 speech and the original speech, gives higher absolute correlation with the subjective judgment than the other four prosodic measures, suggesting that timing plays the primary role in the naturalness of the utterances. At the same time, the absolute correlation of F0 is the second highest. This suggests that F0, which is closely related to perception of pitch and stress, is also a crucial factor. This further confirm our conclusion in previous perceptual experiment: timing improvements contribute more to the improvement of naturalness of L2 utterances perceived by native speakers.

## 6. Conclusion

The roles of different prosodic features in the naturalness of English L2 speech have been investigated with original L2 speech and prosodically modified speech, both through subjective judgment by native speakers and through objectively measured prosodic features. By comparing the utterances with four different patterns related to prosodic errors, we found that pitch related stress and timing play crucial roles in native speakers' judgement of naturalness and the influence of timing is more significant than pitch. Through prosodic modification, the interference of segmental goodness of pronunciation has been excluded. The improvements of naturalness that the prosody-corrected L2 speech shows further confirm our conclusion that timing play more crucial role than pitch in naturalness judgement.

Future works include analysing more data and implementing an automatic prosodic scoring and diagnosis system to improve naturalness of L2 English Learners' speech.

## 7. Acknowledgement

This work is supported jointly by the National Natural Science Foundation of China 81401539 and the National Social Science Fund of China 10CYY009 and 13&ZD189.

## 8. References

[1] Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness, System, 38(2), 301 - 315

[2] Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, 28, 1 - 30.

[3] Tsurutani, C. (2010). Foreign accent matters most when timing is wrong, *Interspeech 2010* 1854-57.

[4] Tsurutani, C. & Luo, D. (2013). Naturalness judgement of L2 Mandarin Chinese — does timing matter? In INTERSPEECH-2013, 239-242.

[5] Luo, D., Luo, R., Wang, L. (2016) Naturalness Judgement of L2 English Through Dubbing Practice. Proc. Interspeech 2016, 200-203.

[6] Smith, R., (2004). The Role of Fine Phonetic Detail in Word Segmentation. Doctoral dissertation University of Cambridge.

[7] Cao, C. (2008). Utilize English Dubbing in English Teaching, *Journal of Liaoning Economic Management Cadre Institute,* 4, 136-137 (in Chinese).

[8] Yanagisawa, K. & Huckvale, M. (2007). Accent morphing as a technique to improve the intelligibility of foreign-accented speech, Proceedings of the International Congress of Phonetics Sciences, Saarbrücken, Germany.

[9] Anne Bonneau, Vincent Colotte. (2011). Automatic Feedback for L2 Prosody Learning. Ivo Ip-sic. Speech and Language Technologies, Intech, pp.55-70.

[10] Zhao, S., Koh, S.N., Luke, K. (2012) Accent Reduction for Computer-Aided Language Learning. Proc. EUSSIP 2012, 335-339.

[11] Zhang, Y., Nissen, S. L., & Francis, A. L. (2008). Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *The Journal of the Acoustical Society of America*, 123(6), 4498 - 4513.

[12] Moulines, E. & Charpentier, F. (1990). Pitch synchronous waveform processing techniques for a text-to-speech synthesis using diphones. Speech Communication, Vol.9, No.5-6, pp. 453-467. Bian, F. (2013). The Influence of Chinese Stress on English Pronunciation Teaching and Learning, *English Language Teaching*, Vol. 6, No. 11, 199 - 211.

[13] W. Mattheyses, W. Verhelst, and P. Verhoeve (2006). Robust pitch marking for prosodic modification of speech using td-psola, in Proceedings of the 2nd Annual IEEE Benelux/DSP Valley Signal Processing Symposium (SPS-DARTS '06), pp.43–46.

[14] Colotte, V. & Laprie, Y. (2002). Higher pitch marking precision for TD-PSOLA, Proceedings of European Signal Processing Conference (EUSIPCO), Toulouse.

[15] S.M. Witt and S.J. Young.(2000). Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning,"Speech Communications, 30 (2–3): pp.95-108.

[16] W. Hu, et al.(2012), A New DNN-based High Quality Pronunciation Evaluation for Computer-Aided Language Learning (CALL) , Proc. INTERSPEECH 2012, 1886-1890

[17] Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1, 126-152.