INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Naturalness Judgement of L2 English through Dubbing Practice

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

[1] Department 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' perception of L2 English spoken by Chinese students through dubbing, or re-voicing practice on video clips. Learning oral foreign language through dubbing on movie or animation clips has become very popular in China. In this practice, learners try to reproduce utterances as closely as possible to the original speech by closely matching lip movements on the clips. The L2 utterances before and after substantial dubbing practices were recorded and categorized according to different prosodic error patterns. Objective acoustic features were extracted and analyzed with naturalness scores based on perceptual experiment. Experimental results show that stress and timing play key roles in native speakers' perception of naturalness. With the practice of dubbing, prosodic features, especially timing, can be considerably improved and thus the naturalness of the reproduced utterances increases.

Index Terms: naturalness, L2 English, stress, timing, foreign accent , dubbing, 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. Among world languages, three major prosodic features – timing, pitch and intensity – 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.

Learning English through dubbing or re-voicing (replacing the sound track with your own recorded voice) has been a boom in China. In this learning practice, learners are required to reproduce the voice of the characters in a video clip as closely as the original speech and match lip movements as perfectly as possible. It is like a kind of voice acting in which learners need to impersonalize the characters by speaking with the same style and speaking rate. One of the most popular educational mobile phone applications, English Fun Dubbing, which enables users to easily dub their own voice into video and post their dubbed works online to share with peer learners or teachers, has attracted millions of downloads in China [3]. Studies show that learning through dubbing not only motivates learners' interests in practicing speaking the target language, by closely observing the lip movements of the characters in the video clips and trying to reproduce the speech in exactly the same way, it can also improve learners' pronunciation by reducing foreign accents [4, 5]. Because of the characteristics of the practice of dubbing that require learners to mimic native speakers' speaking styles, including prosodic features, we consider it a good source to examine learners' English L2 prosody and degree of naturalness through dubbing.

Previous studies on Mandarin speakers' prosodic characteristics have explored how different acoustic cues such as pitch and intensity are used to signal English lexical stress between native English speakers and native mandarin speakers [6, 7]. However, these studies did not measure how different prosodic features are related to naturalness of L2 speech perceived by native speakers. [8] investigated how temporal properties such as speech rate and pause features are related to the degree of foreign accent. However, pitch and intensity, which are the key factors to perceive stress position, were not investigated.

In this study, the influence of prosodic features, in particular, pitch, intensity and timing will be examined, using native listeners' judgment of L2 English utterances. We also investigate how prosodic properties and native speakers' perception of naturalness change with the practice of dubbing.

2. Data collection

2.1. Speakers

36 high school and college students (age 17-22) who are native Mandarin speakers and have previous experiences in learning English through dubbing practice participate as speakers in our experiments. There are 17 females and 19 males with different degrees of English speaking proficiency.

2.2. Material

The video used for dubbing practice is a two-minute long clip of a male native English speaker telling a story. The content was carefully chosen so that it contains words that Chinese students tend to make stress errors according to [9], but the degree of vocabulary difficulty does not exceed the level of a typical high school student according to the curriculum. The text information (transcript) was added to the video as subtitles. Participants can play the video clip and mimic the original native speech at their own pace before dubbing their own

http://dx.doi.org/10.21437/Interspeech.2016-623

voices. During the dubbing practice, the subtitled video with no sound tracks is presented to the learners.

2.3. Recording procedure

In order to compare the prosodic properties of L2 speech before and after substantial dubbing practice, we first recorded the learners' reading-aloud of the transcript of the video. Before recording, the participants were allowed to listen to the sound track of the clip (no video was presented) twice to get a general impression of the rhythm and pronunciation of the original speech. During the recording, only the transcript were presented to the participants. After the recording of reading-aloud finished and the recording quality was checked, we sent the video clip to the participants and they were allowed to practice mimicking the original speech at their own pace. However, they were required to dub their own voice into the clip and upload at least one dubbed work after each day's practice using a mobile phone application specifically designed for dubbing practice. We monitored each upload for recording quality and inform the participant if any problems were found. Thus we ensured each participant had finished at least 10 rounds of dubbing practice.

2.4. Data selection and categorization

The audio data collected from the participants' reading-aloud and the final round of dubbing practice were segmented into sentence utterances (one utterance contains one sentence). From reading-aloud data, utterances that contain stress errors or timing errors or both, with no obvious segmental errors, were chosen, together with a correct utterance from a large sample of 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 from reading-aloud speech recorded before dubbing practice (referred to as "before dubbing" hereafter) data that contained the listed error patterns were chosen and formed 24 stimuli. For comparison, the same sentences spoken by the same speakers from final dubbing practice speech data (referred to as "after substantial dubbing" hereafter) were also chosen as stimuli. Altogether, there are 48 stimuli for our perceptual experiment.

3. Perceptual experiment

3.1. 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. Their ages ranged from 25 to 49.

3.2. Procedure

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

The task was conducted online with the following prompt:

"You will listen to 48 utterances spoken by non-native speakers. 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 (the highest 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 48 stimuli were presented in a randomized order to the subjects for their subjective scoring. Each stimulus could be listened to only once.

3.3. Results

The means and the standard errors of the scores of subjective judgment on "before dubbing" 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 tends to receive worse evaluation than those with only stress errors. Timing is more crucial a factor on native speakers' naturalness judgement of the utterances.

To check the statistical significances of these differences, multiple paired t-tests with Bonferroni adjustment were conducted on the subjective scores of "before dubbing" utterances between each pair of error patterns. The two utterances sharing the same text sentence but differing in error pattern were paired in the t-tests. As shown in Table 1, the differences in the mean of the subjective scores are significant ($p < 0.01$) between any pair of error patterns, further confirming two conclusions: (1) both stress and timing have perceivable contributions to the naturalness of L2 speech; (2) timing errors have bigger impact than stress errors on the naturalness of L2 speech.

In order to examine how dubbing practice affects L2 prosody and native speakers' perceived naturalness, the means and the standard errors of the scores of subjective judgment on "after substantial dubbing" utterances of the same sentences (and spoken by the same speakers) were also calculated for each previous error pattern. Note that the 4 error patterns here are only used for labeling to show that the speakers had such errors with the sentences before dubbing practice (the errors might be corrected after dubbing practices). As shown in Fig. 2, the average naturalness scores of all the four groups of utterances after substantial dubbing practice increase, compared with the data of before dubbing. Especially the utterances previously labeled as SoTx and SxTx show higher increases than the other two groups. This indicates that through substantial dubbing practice, learners with timing errors can improve the naturalness of their speech more significantly than those with stress errors. However, we still need to conduct acoustic measures to confirm that prosodic features related to timing can be improved more significantly than those features that are related to stress.

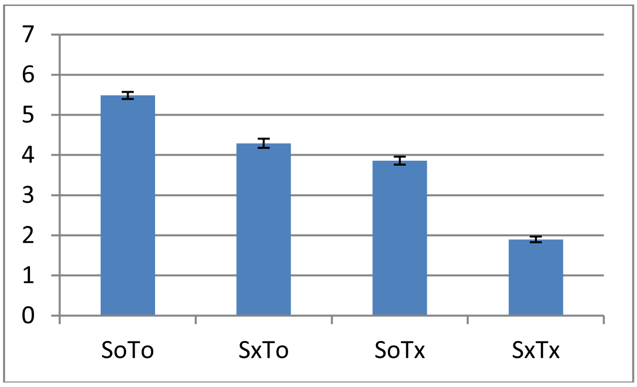

Figure 1: *The average naturalness score (+/- standard error) for four types of errors before dubbing*

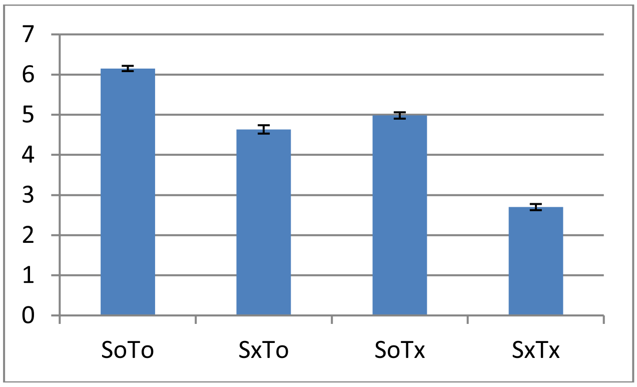

Figure 2: *The average score (+/- standard error) after substantial dubbing practice (four types of errors refer to the errors in the same sentences recorded before dubbing)*

Table 1: *Results of multiple paired t-tests between four error patterns.*

Pair of error patterns	*t*	*p*
SoTo–SoTx	7.591	.000
SoTx–SxTo	5.673	.000
SxTo–SoTo	3.624	.001
SxTx–SoTo	18.615	.000
SxTx–SoTx	14.136	.000
SxTx–SxTo	15.351	.000

4. Acoustic measures

4.1. Measures based on different prosodic features

According to Fry, stress is the result of interaction of pitch, intensity, and duration [10]. 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 examine their roles in naturalness judgement of L2 utterances before and after dubbing.

4.1.1. 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 [11]. 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

$$D(native, L2) = \frac{g(I, J)}{I + J} \quad . \qquad (1)$$

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

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

$$g(i,j) = \min \left\{ \begin{array}{c} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i,j-1) + d(i,j) \end{array} \right\}, i > 1 \text{ or } j > 1, \qquad (2)$$

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.

4.1.2. Distance in intensity

For a similar reason, we can compare the intensity contours between an L2 utterance and a native utterance. The intensity contours of the speech were also extracted using Praat. Like for F0, the DTW distance in intensity between two utterances can be defined in the same way.

4.1.3. 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 *4.1.1.* 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.

4.1.4. Average phone duration

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

$$PhoneDur = \frac{D_{utterance} - D_{pauses}}{N_{phones}} \qquad (3),$$

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.

4.1.5. 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 in dubbing practice, 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 through phone-level forced alignment using HMM acoustic models mentioned in *4.1.1.*

Table 2. *Correlations between each prosodic measure and the subjective score.*

Prosodic measure	Correlation
F_0 Distance	–0.42
Intensity Distance	–0.31
Duration	–0.01
Pause	–0.32
Syllable time	–0.54

Table 3. *Syllable time difference and F0 distance features of the utterances in three different recordings*

Recordings	Average Syllable Time Difference (seconds)	F_0 Distance (word average)
Before dubbing	3.1	0.18
First dubbing	2.01	0.15
Final dubbing	0.97	0.14

The syllable time difference feature is defined as,

$$\mathrm{D}_{\text{syllable}}(native, L2) = \frac{\sum_{i=1}^{N}(\ |S_i(native)-S_i(L2)|+|E_i(native)-E_i(L2)|\ )}{N} \quad (4),$$

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.

4.2. Objective acoustic feature analysis

Using the same 48 utterances as in the perceptual experiment, we calculated the correlations between each aforementioned prosodic measure and the score of subjective judgment. As shown in Table 2, "syllable time", which indicates the timing difference between dubbed 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.

We then examine how these two crucial feature factors change before and after substantial dubbing practice. We extracted syllable time different and F0 distance from utterances recorded before dubbing, in the first round of dubbing practice (referred to as first dubbing) and the final dubbed speech used in perceptual experiment (referred to as final dubbing). As shown in Table 3, the syllable time difference decreases significantly with the practice of dubbing, which indicates that the timing on syllable level of the L2 speech is much closer to the original speech. 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.

5. Conclusion

The roles of different prosodic features in the naturalness of English L2 speech before and after dubbing practice have been investigated, both through subjective judgment by native speakers and through objectively measured prosodic features. By comparing the utterances with four different patterns of prosodic errors, we found that stress and timing play crucial roles in native speakers' judgement of naturalness. The different results of naturalness scores of the utterances recorded before and after substantial dubbing show that learners can improve naturalness of their speech through dubbing practice, especially with timing errors. Objective measures of acoustic features such as F0, intensity, pauses and syllable time, which are related to stress and timing, further confirm that with substantial dubbing practice, the prosodic property of timing can significantly be improved and thus the degree of naturalness increases.

Future works include analysing more data and implementing an automatic prosodic scoring system to predict native speakers' naturalness judgement of L2 English speech.

6. Acknowledgements

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

7. 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] Hangzhou Daily, http://hzdaily.hangzhou.com.cn/hzrb/html/2015-06/18/content_1993494.htm

[4] Duan, C. (2011). Dubbing and Oral English Teaching, *Continue Education Research*, 28, 128 - 129 (in Chinese).

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

[6] Chen, Y., Robb, M. P., Gilbert, H. R., & Lerman, J. W. (2001). A study of sentence stress production in Mandarin speakers of American English. *The Journal of the Acoustical Society of America*, 109(4), 1681 - 1690

[7] 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.

[8] Wang, X. (2014). Mandarin Speakers' English L2 Prosody and Degree of Foreign Accent: Effect of Length of Residence, *Proceedings of the International Symposium on the Acquisition of Second Language Speech*, 737 - 749.

[9] Bian, F. (2013). The Influence of Chinese Stress on English Pronunciation Teaching and Learning, *English Language Teaching*, Vol. 6, No. 11, 199 - 211.

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

[11] Charniak, Eugene, et al. BLLIP 1987-89 WSJ Corpus Release 1 LDC2000T43. DVD. Philadelphia: *Linguistic Data Consortium*, 2000.

[12] Gong, K. J. (1991). Comparison of Chinese and English Intonation and English Intonation Teaching. *Modern Foreign Languages*, 3, 43-45..

[13] Tsurutani, C. (2009) Intonation of Japanese sentences spoken by English speakers *INTERSPEECH 2009*, 692-695.

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