Learners’ Prosodic Control in the Task of Expressive Storytelling and Predicted Native Listeners’ Impressions of the Learners’ Speech

Chihiro Shoda†, Yingxiang Gao‡, Yurun He†, Nobuaki Minematsu†, Noriko Nakanishi‡, Daisuke Saito†

†The University of Tokyo, Japan, ‡Kobe Gakuin University, Japan
{chihiro_shoda, gyx, heyurun, mine, dsk_saito}@gavo.t.u-tokyo.ac.jp, nakanisi@gc.kobegakuin.ac.jp

Abstract

Various kinds of L1 transfer are found in L2 speech, which often influence transmission of both linguistic and para-linguistic information. In this study, we focus on how adequately Japanese learners of English can transmit para-linguistic expressions in the task of storytelling. For this aim, we analyze 1) how learners control the prosodic features to express specified emotions and 2) what kind of impressions native listeners perceive from the learners’ expressions. For the first analysis, the learners’ prosodic control is compared with the corresponding native control. For the second analysis, with speech emotion recognition technology, L2 speech is converted to its emotion posteriorgram. Then, the intended emotions are compared with the perceived emotions. Experiments show the learners have difficulty in imitating the native intensity control, especially when expressing sadness, and distinction among perceived emotions is remarkably reduced.

Index Terms: Para-linguistic information, L2 prosody, impressions, emotion posteriorgram, storytelling, CALL

1. Introduction

To realize effective and efficient technical support for learners to acquire a good communication skill in a new language, a large number of studies have been conducted [1–4], but a majority of them aimed at accurately detecting segmental errors caused by L1 transfer and assessing the segmental aspect of learners’ pronunciation. However, teachers often claim that segmental errors due to L1 transfer, i.e. accentuatedness, represent learners’ national identity, and that listeners should respect and accept the errors if the learners are intelligible enough [5–7]. Prosodic training is often introduced for language learners to increase the intelligibility of their speech effectively and efficiently to native listeners, and there exist automated systems that can help learners to acquire a good and natural prosodic control to improve their intelligibility and/or comprehensibility [8,9].

In this paper, we focus on learners’ prosodic control, but do not focus on their lexical or phonemic intelligibility, which represent how many words or phonemes intended by the learners are perceived accurately by listeners [10,11]. Our focus on learners’ prosodic control is to analyze the intelligibility of para-linguistic expressions intended by the learners. For this aim, we introduce a special task to learners, that is expressive storytelling of picture books. In regular context of language education, speech training is often given to learners as the task of just repeating sentences read aloud rather inexpressively by model speakers. As far as the authors know, expressiveness training is rare except for drama-based acting training in the target language [12–14]. In this situation, the authors wonder how intelligible learners are when trying to speak expressively.

How to calculate automatically the intelligibility of expressions in learners’ speech? If we take the phonemic intelligibility instead, which represents how many phonemes intended by learners are transmitted correctly to listeners, phoneme-based posterior probability, \( P(p_n|o) \), is regarded as one technical solution [15], where \( p_n \) is the phonemic class, \( n \) and \( o \) are acoustic observation. If \( P(p_K|o) \) is very high, where \( K \) is the phonemic class intended by the learning speaker when producing \( o \) who will be assessed to be very intelligible in producing phonemes. The average score of \( P(p_K|o) \) over an input utterance is the well-known metric of Goodness Of Pronunciation (GOP) [15]. The lexical intelligibility, however, is a more complex target to calculate, because perception of words depends on various acoustic, prosodic, syntactic, and pragmatic factors [16,17]. To take all these factors into account, automatic prediction of the lexical intelligibility was discussed in [18].

Inspired by use of phonetic posteriorgram (PPG) to calculate the phonemic intelligibility based on \( P(p_K|o) \), in this work, we will use emotion posteriorgram (EPG), namely, \( P(e_n|o) \), where \( e_n \) is the emotion class. \( P(e_n|o) \) is calculated by using a high-performance speech emotion recognizer [19], and if \( P(e_K|o) \) is high enough, the learning speaker is assessed to be skillful enough in speaking expressively. The average score of \( P(e_K|o) \) is regarded as Goodness Of Expression (GOE).

In the following sections, after describing some related works and the fundamental techniques applied, we show the results obtained in two experiments. All the recordings were made by introducing to Japanese learners of English the task of expressive storytelling of picture books. Here, the learners have to imitate a model speaker’s expressive storytelling by overlapping the learners’ speech exactly on the model speech with their special attention paid to the intensity, pitch, and duration control made in the model speech. Here, the emotion class intended by the model speaker was explicitly given to the learners when they tried to imitate each of the model speeches.

In the first experiment, the prosodic control of the learners is compared with the model control to examine what kind of prosodic control to express what kind of emotion is difficult for the learners to realize when trying to imitate the model speaker. In the second experiment, the learners’ recordings as well as the model recordings are taken as input to a high-performance emotion recognizer to derive their EPG. The intelligibility of emotions, or perceptual distinctiveness among different emotions, is compared quantitatively between the learners and the model speakers. To the best of our knowledge, this paper is the first technical attempt to introduce EPG to analyze learners’ speech to examine how intelligible or distinctive they are when trying to speak for transmitting specified para-linguistic expressions.
2. Related works

2.1. Prosody in Japanese-accented English speech

As far as we know, L1 transfer to L2 expressive speaking has not been well discussed so far, and no studies are found to examine acoustically Japanese learners’ performance of English expressive storytelling. Meanwhile, speech prosody observed in their reading aloud has been analyzed in previous studies [20–23], where L1 transfer is well discussed. For example, it was pointed out that 1) word accent is realized mainly by controlling pitch because Japanese word accent is pitch accent, 2) similar syllable magnitude is often assigned to stressed syllables and unstressed ones, because Japanese does not have strong and weak alternation in speech, and 3) phrase intonation often has a very high beginning and a preceding gradual and long downfall in pitch, which is a typical phrase or sentence intonation in Japanese.

2.2. Speech training based on prosodic overlapping

To help L2 learners to acquire a good and natural prosodic control, a gamified interface for visualizing the learners’ prosody and comparing it with the model prosody was introduced to English class for Japanese learners [24]. For a model speech, which is to be imitated by the learners, its waveform and two prosodic features: 1) sequence of normalized syllable magnitude as stress-timed rhythm and 2) sequence of normalized fundamental frequency as intonation, are visually presented on the learners’ laptop. By clicking the start button, a replaying indicator started moving on the model waveform from left to right, and the learner has to imitate the model speech so that his/her speech is overlapped with the model speech. Less than 1 sec after the recording, the degree of prosodic overlapping is visually shown and quantitatively rated as score for each of the two prosodic features. Figure 1 shows two examples of good and poor overlapping. Since the visualization and scoring is performed very immediately after recording, this training was taken by the learners as karaoke-like speech game, some of whom practiced repeatedly until midnight. This interface effectively increased their awareness to prosodic control [24].

However in [24], native speakers’ reading aloud samples with less expression were used as model speech. In the current study, we introduce this training method to the task of imitating expressive storytellings given from a skilled native storyteller.

2.3. Speech emotion recognition

Speech emotion recognition is the task of inferring the emotion state of a speaker only based on low-level acoustic features. Recently, deep learning has improved the performance of automatic speech emotion recognition remarkably, and by integrating LSTM and CNN into transformer, which was implemented as cross-attention transformer, [19] realized a better performance than state-of-the-art models. Acoustic input was converted into mel spectrogram, MFCC, and wav2vec2 representations [25], which were taken as input separately to CNN or BLSTM. Cross-attention transformer was applied to these three outputs in a hierarchical manner, and the emotion class was identified by calculating posterior probability $P(y_{n}|o)$.

The classifier was trained using the IEMOCAP dataset [26]. Since imbalance among the emotion classes was not ignorable, as in previous studies, only four classes of anger, happiness, sadness, and neutral were used for training and testing. After splitting any utterance into a sequence of 3-sec segments, consecutive segments of which were overlapped by 1.5 sec, the emotion class was predicted using the above model.

Speech emotion recognizers are generally trained by using the emotion classes inferred subjectively by listeners, not the classes intended by speakers. They are also trained only by using raw acoustic features, not using lexical and textual features obtained automatically by speech recognizers. These conditions for training are well suited to our aim of analyzing L2 learners’ speaking control for expressing specified emotions in terms of intelligibility, or perceptual distinctiveness.

3. Collection of L2 storytellings

3.1. Three picture books and model speech samples

Three picture books of “Goodnight Moon” [27], “Is Your Mama a Llama?” [28], and “The Spider and the Fly” [29] were used. They are known as favorite picture books of young children in US, and many sentences in the three books rhyme clearly. A skilled adult female was recruited as model storyteller, and we asked her to read aloud the books in the following way.

1) She read aloud a set of sentences selected from the books in two modes: a) reading monotonously and b) reading expressively for young children.
2) She read aloud another set of sentences selected from the books in two modes: c) reading neutrally and d) reading expressively according to the emotion specified by the authors. The specified emotion is one of anger, happiness, and sadness, which were used to train speech emotion recognizers.

The number of sentences in 1) is 20 and that in 2) is 17. Since each of them was read aloud in multiple modes, the total number of model speech samples was 77. These samples were repeated by Japanese learners of English while paying special attention to their prosodic control as in Section 2.2.

3.2. Participants and the procedure for data collection

21 university students participated in data collection, and they were asked to read aloud the 37 sentences four times.

R1) They read aloud the sentences with the same instructions above: a) monotonously, b) expressive, c) neutral, and d) specified emotions. Repeated recording was allowed but the model speech samples were NOT presented.

PO) They read aloud the sentences with the same instructions in the framework of prosodic overlapping. They listened to and imitated the model speech carefully. Repetition was allowed.

R2) Immediately after PO, they read aloud the sentences again with the same instructions, but without the model speech samples presented. Repetition was allowed.

R3) One week after R2, they read aloud the sentences again with the same instructions, but without the model speech samples presented. Repetition was allowed.

R1 was for collecting the learners’ baseline performance of reading expressively, and PO and R2 were for their performance during and immediately after prosodic overlapping train-ing.
ing, respectively. R3 was for examining whether the effect of prosodic training for expressive reading lasted one week or not.

All the recordings were done online with a web interface, and after removing some speech samples corrupted due to technical errors, the total number of recordings from the 21 students was 5,085. It should be noted that data collection was started after the authors received an ethical approval statement from the ethics committee of the authors’ institute and an informed consent from all the participants.

4. Comparison of prosodic control between the model and the learners

4.1. Analysis of the three kinds of prosodic control

Forced alignment was conducted for all the 77 model speech samples with their transcripts, and the starting time indexes and the ending ones of all the vowels were detected automatically. After converting all the model samples and the learners’ samples into their Phonetic PosteriorGram (PPG), each sample of the learners was time-aligned with its corresponding model sample via Dynamic Time Warping (DTW). Then, all the vowels were detected also from the learners’ samples. For PPG conversion, the WSJ-KALDI recipe [30] was used and the DTW was performed with symmetric local path constraints [31].

Similarity of prosodic control between a learner’s sample and its corresponding model sample was quantified along their DTW alignment path. For pitch control similarity, the correlation between the logarithmic values of fundamental frequencies of the learner’s vowel frames and those of the corresponding model frames was calculated. For intensity control similarity, the correlation between the logarithmic values of the speech power of the learner’s vowel frames and those of the corresponding model frames was calculated. Unlike these, for durational control similarity, the correlation between the learner’s vowel durations and the corresponding model vowel durations was calculated. By comparing the three correlations for each case of the specified emotions, we can discuss what kind of prosodic control to express what kind of emotion is difficult to imitate even with visual and corrective feedback in PO.

All the correlations above were calculated by the unit of recording. Using the three kinds of prosodic features along the alignment path, however, for each prosodic feature and each model sample, the averaged imitation gap over the learners can be quantified in sequence along the time axis of the model sample. Here, to suppress speaker-specific biases in fundamental frequency and power, these two features were normalized and used for averaged imitation gap analysis. Through this analysis, it will be possible to discuss quantitatively in which syntactic, lexical, phonemic, phonetic, and prosodic contexts, Japanese learners tend to have difficulty to trace the model prosody.

4.2. Results and discussion

For a pair of a learner’s speech and its corresponding model speech, the three correlations were calculated as similarities in pitch, intensity, and duration control. Table 1 shows the averaged correlations in PO recordings over the 21 learners for each prosodic feature and each mode.

<table>
<thead>
<tr>
<th>mode</th>
<th>pitch</th>
<th>intensity</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.62</td>
<td>0.59</td>
<td>0.80</td>
</tr>
<tr>
<td>b</td>
<td>0.62</td>
<td>0.56</td>
<td>0.81</td>
</tr>
<tr>
<td>c</td>
<td>0.67</td>
<td>0.61</td>
<td>0.84</td>
</tr>
<tr>
<td>d-1</td>
<td>0.72</td>
<td>0.60</td>
<td>0.85</td>
</tr>
<tr>
<td>d-2</td>
<td>0.69</td>
<td>0.57</td>
<td>0.87</td>
</tr>
<tr>
<td>d-3</td>
<td>0.57</td>
<td>0.47</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Table 2: Averaged correlations from M to Rn and PO

<table>
<thead>
<tr>
<th>mode</th>
<th>pitch</th>
<th>intensity</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.37</td>
<td>0.52</td>
<td>0.76</td>
</tr>
<tr>
<td>b</td>
<td>0.64</td>
<td>0.56</td>
<td>0.81</td>
</tr>
<tr>
<td>c</td>
<td>0.59</td>
<td>0.55</td>
<td>0.82</td>
</tr>
<tr>
<td>d-1</td>
<td>0.39</td>
<td>0.52</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Table 3: Averaged correlations from R1 to PO and Rn (n=2,3)

<table>
<thead>
<tr>
<th>mode</th>
<th>pitch</th>
<th>intensity</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.47</td>
<td>0.70</td>
<td>0.74</td>
</tr>
<tr>
<td>b</td>
<td>0.55</td>
<td>0.73</td>
<td>0.79</td>
</tr>
<tr>
<td>c</td>
<td>0.56</td>
<td>0.72</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Minimal pairs of words which differ only in vowel duration: short vs. long. This means that they are also good at making durational contrasts among syllables. However, phonologically speaking, Japanese does not have intensity contrasts to distinguish words. Next, we compare the correlations among the six modes.

4.1. Analysis of the three kinds of prosodic control

For each of the 71 model speech samples, we calculated the averaged imitation gap pattern for each of the three prosodic features on the time axis of the model speech. One example in PO is shown in Figure 2. The three gap patterns correspond to pitch, intensity, and duration, each of which also shows its model trajectory on the top. By inspecting this figure, we can discuss where in this model speech the learners had difficulty to trace which prosodic feature. From the 77 results, some findings were found such that 1) the intensity gap is generally large as indicated in Table 2, and 2) very expressive control in pitch and duration shows large gaps. In Figure 2, the pitch gap is large when an abrupt pitch downfall is produced in M, and the vowel durations are larger when vowel durations are much lengthened in M. Other findings are listed in the first author’s bachelor thesis [32], which interested readers should refer to.
5. Predicted native listeners’ impressions using emotion posteriorgram

5.1. Analysis of emotion posterior probabilities

All the recordings of c (neutral) and d (three emotions) in R1, PO, R2, R3 were segmented into 3-sec segments with 1.5-sec overlap. It should be noted that, in c and d, the students recorded their voices with the emotion class specified explicitly. Each segment was taken as input to the speech emotion recognizer [19] to be converted into its emotion posterior vector. After the averaged vector was obtained for each recording, Principle Component Analysis (PCA) was conducted for all the averaged vectors. Then, the first and second principle components were obtained from each speaker’s recordings to visualize the distinctiveness among the intended emotions.

5.2. Results and discussion

Figure 3 is PCA plots of the first and second components of the model speaker and nine students (S1 to S9) who are available in this analysis. Recordings in R2 are used to be plotted in Figure 3. In the model plot, the four emotions seem to be separated clearly into three classes of neutral (top-right), happy+angry (bottom-right), and sad (top-left). In the students’ plots, however, the four emotions are not well separated and are overlapped with each other. With closer inspection, in S6 and S8, model-like separation is realized to some degree although the area of distribution is much smaller than the model distribution. As explained in Section 1, the degree of separation of the emotions is regarded as intelligibility or perceptual distinctiveness of para-linguistic expressions. Although we did not carry out any subjective tests for the students’ utterances, from the figures, we can say that the emotions intended by the students may not be transmitted distinctively to native listeners. Taking the current situations of speaking training in class into account, these results may be reasonable, because expressiveness training is rare except for drama-based acting training [12–14]. Another possible reason for the results is that expressive speaking is in nature involuntary, but the task of expressive storytelling requires voluntary production of expressions, which may have induced some unnaturalness. In drama-based acting training, learners have to be expressive voluntarily at first, and after training, they will be able to produce natural expressions involuntarily. As shown in Table 2, during prosody overlapping training, the learners were able to reproduce the native-like prosodic control to some degree, but they were not one week later. They may have to keep up these kinds of training for longer to be more expressive in a natural way.

6. Conclusions

This paper conducted two kinds of analysis on the L2 speech samples collected in the task of storytelling of picture books. In the first analysis, by comparing the model prosodic control and the learners’ control, we found that how L1 transfer influenced the learners’ production of expressions. In the next analysis, the distinctiveness of expressions was estimated quantitatively using speech emotion recognition techniques, which showed reduced distinctiveness in the students’ expressions.

\footnote{When we listened to speech samples of IEMOCAP, we found that happy and angry utterances are similar acoustically, indicating that these utterances are supposed to be separated using non-acoustic features such as lexical features as well as acoustic features.}
7. References


