L2 LEARNERS' EMOTION PRODUCTION IN VIDEO DUBBING PRACTICES

Lei Chen†, Cheng Chang, Cheng Zhang, Huan Luan, Jiawen Luo, Ge Guo, Xiaofei Yang, Yang Liu†

LAIX Inc.
† Silicon Valley AI Lab, San Mateo, CA
Shanghai, China

ABSTRACT

Video dubbing is a new type of language learning practice. Because of the fun it brings into learning, video dubbing mobile applications have become quite popular. During video dubbing, learners not only mimic characters' pronunciations but also other voicing characteristics, e.g., emotions. In this study, we collected a corpus from Liulishuo English learning App's video dubbing practices and investigated three research questions. We find that L2 learners do try to mimic emotions via their voices, and human raters tend to focus more on the emotion production aspect when evaluating dubbing performance. Furthermore, we develop state-of-the-art speech emotion recognition technology and show it can be used for evaluating emotion production automatically.

Index Terms— video dubbing, language learning, emotion recognition, pronunciation evaluation, automatic assessment

1. INTRODUCTION

As mentioned in [1], "Learning English through dubbing or re-voicing (replacing the sound track with your own recorded voice) has been a boom in China". In a video dubbing task, learners are expected to mimic the voice of the character in a video clip on both linguistics and para-linguistics aspects. For example, a successful dubbing consists of re-matching pronunciation, accent, and emotions to the original characters' voices. Previous studies (e.g., [2]) have suggested that the dubbing practice helps attract students and keep their studying enthusiasm, because of its high entertainment value. As a result, the dubbing practices have been found to be useful to improve language learners' pronunciation.

In conventional sentence repetition practices, L1 stimulus mostly is spoken in a neutral emotion condition. However, stimulus in video dubbing practices generally is extracted from movies and contains a considerable amount of emotional voices. Consequently, using emotional L1 stimulus brings more challenges to the existing automatic pronunciation assessment methods. Some acoustic features that have been found useful for rating neutral speeches may not work well for emotional speeches. New human rating guidelines as well as the development of new automatic evaluation systems may be needed for emotional speeches. On the other hand, using emotional L1 stimulus also brings a new opportunity to study more comprehensive aspects required in language competence. For example, in China's new English skill level metric [3], properly expressing emotions has been required to be one of pragmatic skills. Therefore, video dubbing practices using emotional L1 stimulus may provide an opportunity for training this kind of skills.

In this study we leverage the video dubbing data from Liulishuo language learning App, and aim to answer three research questions: (i) do users mimic emotion? (ii) do human raters pay attention to emotion mimic? (iii) can we automatically recognize learner's emotion in video dubbing?

The remainder of the paper is organized as follows: Section 2 reviews previous research. Section 3 lists the three major research questions in this paper. Section 4 describes the corpus we built. Section 5 describes the experiments we conducted to answer the research questions. Section 6 concludes and discusses future work directions.

2. RELATED WORK

Previous studies [2, 4] have investigated the usefulness of video dubbing for improving learners' pronunciation and intonations in the Second Language Acquisition (SLA) domain. [2] used film dubbing to improve intonation skills and found that the majority of participants considered the approach to be effective. [4] tracked 34 Chinese English learners' progresses in four weeks, and found that the learners' oral proficiency in comprehensibility, fluency, and accent reduction, improved. In addition, these learners enjoyed the task and had a high perception of their progresses and positive attitude toward the task. [5] clearly mentioned "mastery of paralinguistic elements can be practiced". It suggested that learners can practice multifaceted, high-level language production tasks in video dubbing practices.

[1] compared L2 learners' prosody skill developments before and after substantial video dubbing practices on a mobile App. Both subjective ratings on naturalness and objective acoustic evaluations (i.e., correct or wrong timing and stress) were used. Their experimental results showed that stress and

timing play key roles in native speakers' perception of naturalness. With the practice of dubbing, L2 learners' prosodic performance, especially timing, can be improved and thus the naturalness of the reproduced utterances increases.

3. RESEARCH QUESTIONS

Previous studies using video dubbing practices have not considered the fact that L1 video stimulus may be emotional. Therefore, in this work we focus on the emotion aspect of speech in movie dubbing. Specifically, we aim to answer the following three research questions.

- RQ1: Do L2 learners mimic emotions expressed in video stimulus in movie dubbing?
- RQ2: For emotional stimulus, when rating the dubbed voices, which aspects (e.g., pronunciation, intonation, or emotion production) do human raters focus on?
- RQ3: Is it possible to use existing emotion recognition technology to automatically evaluate learners' emotion production capability?

4. CORPUS

Liulishuo (on both iOS and Android platforms) is one of the most popular English learning Apps in China, with more than 80 million registered learners. It focuses on training speaking abilities and provides various types of practices, including sentence repeating, situation-based chatting, and so on. Learners' voice inputs are automatically evaluated on multiple aspects, i.e., pronunciation, fluency, pacing, and so on. The video dubbing was a component in the App. A learner first watches video clips with native English voices and then use his/her L2 speech to replace original audio track. The learner can try as many times as possible until he/she is satisfied with the dubbed video.

Given the size of all the L1 video clips used in the App, it would be very time consuming to manually annotate each clip's emotion category. Therefore, we utilized a pre-trained movie clip emotion recognition model [6][1], which was released for the 2018 Emotion in Wild (emotiW) grand challenge, to select the clips with a high probability of being emotional. These selected clips were then used in human annotation. Two human raters annotated $3,141$ pre-selected clips using the following emotion categories: *neutral, angry, happy, sad, surprised, fear, disgust*, and *other*. A clip can be labeled with multiple categories[2]. In addition, regarding the strength a character expresses his or her emotion (denoted as effect), raters provided their perceptions on three levels, from the lowest one (A) to the highest one (C). Among the two raters' emotion category ratings, the Cohen's κ value is 0.60, showing an acceptable rating consistency. From all of the pre-selected L1 clips, we obtained $2,177$ clips that have agreed human ratings on emotion categories.

To prepare the L2 learners' data, we focus on three most common emotion categories, i.e., happy, angry, and sad. From the annotated clips, we selected 138 whose effect levels are either B or C levels for sampling L2 audio files. For each of these L1 clips, from all of its corresponding L2 dubbing voices, we randomly selected 30 speech files. One selection criterion we used is pronunciation score (a value from 0 to 100 computed by our own automatic pronunciation evaluation service, which will be described in Section 5.2). Their pronunciation scores need to be at least 80, so that in this work we focus on learners with a reasonable pronunciation capability and study if and how well they mimic emotions. For this initial study, we use 20 L1 clips and manually annotated 600 L2 voices (30 sampled speakers for each clip) from two aspects. One is related to user's emotion production, ranging from 0 (no emotion production at all) to 3 (mimicking well). The other aspect is a holistic rating, from 0 (the lowest level) to 4 (the highest level), which reflects raters' overall perception of the dubbing quality, considering speaking quality, e.g., pronunciation, intonation, and emotion. Two human raters performed annotation in parallel. Their κ values are 0.639 on the emotion-production aspect and 0.634 on the holistic rating aspect.

5. EXPERIMENTS

5.1. Do learners mimic emotions?

According to the rating guideline, score 0 means that L2 learners has no attempt of mimicking emotions. score 1 means emotion production can be sensed, while score 2 and 3 show noticeable emotions. On the 600 L2 dubbing speech files, we averaged the two raters' scores for the user's emotion production, and calculated score percentages for the three bins: $[0.0, 0.5]$ (no emotion production), $[1.0, 1.5]$ (showing the sign of emotion production), and $[2.0, 3.0]$ (noticeable emotion production). From Table 1, we can find that 9.67% have clear emotion productions and 56.0% dubbing voices in our study show learners tried mimicking emotions. Note that in the App, there is no explicit instruction to ask L2 learners to produce emotions. However, even in such a condition, we still see a high percentage of users that do mimic the emotions shown in the stimulus.

score	[0.0, 0.5]	[1.0, 1.5]	[2.0, 3.0]
percentage	44.00	31.33	9.67

Table 1. Distribution of emotion production scores.

[1] https://github.com/zeroQiaoba/EmotiW2018

[2] We only used the first category appearing in the annotation string in follow-up data analysis.

5.2. Do human annotators consider emotion when rating dubbing voices?

To answer this question, we measure the correlation between the overall holistic score of the dubbing voice and two factors raters may consider in their judgment: emotion production and pronunciation quality. For pronunciation, we use an in-house automatic measurement system. Goodness of Pronunciation (GOP) [7] measurement method is used for phone level pronunciation performance. Then the average of these measurements on sentence level is used to provide an overall pronunciation score. The ASR system used for recognizing learners dubbing voices is a DNN-HMM hybrid ASR system built with the Kaldi open-source toolkit. This model is a 9-layer TDNN using acoustic features from the current frame plus the previous and following 5 context frames. The ASR model is trained on from Liulishuo's internal corpus containing about 2500 hours of native speech and non-native speech. The ASR system achieved a word error rate (WER) of 9% on learners speech files.

Table 2 reports the correlation analysis results. The average of the two raters on both holistic and emotion-production aspects is used. We show the correlation between holistic scores and emotion production, holistic and pronunciation, and emotion production and pronunciation measurements. Pronunciation measure does not show a high correlation to human rated holistic scores. One possible reason may be that the pronunciation measurement developed for speech with a neutral emotion may not work well on emotional speeches. We will investigate this issue in our future work. We can see from the table that Emotion scores have a quite high correlation ($r = 0.641$) with Holistic scores, showing that human raters' perceptions of emotion productions play an important role when forming holistic judgments on learners' dubbing voices.

	Emotion	Pronunciation (auto.)
Holistic	0.641	0.128
Emotion	–	−0.074

Table 2. Correlations between human rated holistic scores, emotion production, and automated-rated pronunciation measurement.

5.3. Can we automatically recognize learners' emotions?

5.3.1. Automatic speech emotion recognition method

Speech emotion recognition (SER) has been actively investigated for decades. [8] reviewed a lot of work before deep learning technology has become a mainstream method in various research areas. [9] created the IEMOCAP multimodal emotional conversation corpus that has been widely used in the following SER research. In recent years, deep neural network (DNN) methods have been widely used to recognize emotions. For example, [10] used an Long Short-Time Memory (LSTM) recurrent neural network (RNN) model and found that using an attention weighting approach gives improved performance than simply averaging low-level acoustics features.

In this paper, we propose a Bidirectional Gated Recurrent Unit [11] (BGRU)-based NN model with attention for emotion recognition. Figure 1 shows the model architecture. All the implementations are based on TensorFlow [12] and Keras [13]. We first use the pyAudioAnalysis [14] toolkit to extract short-term features, 34-dimension[3]. The short-term window size and window step were set to $0.2 * Fs$ and $0.1 * Fs$ respectively, where Fs was the sampling frequency, i.e. 16000 Hz. Then we truncate the features to only the first 20 dimensions by removing the features related to chroma. Lastly the $1st$ order delta coefficients were computed and concatenated with the 20-dimensional features. We denote the resulting 40-dimensional feature vectors as segment-level features. Then each segment's features were fed to a FC-BN-RELU structure, i.e., a Fully Connected (FC) layer with 64 neurons and no activation function, followed by a Batch Normalization (BN) layer [15] and a rectified linear unit (RELU) activation layer. This will result in a new encoding for each segment. The BN layer here was used for faster convergence and better generalization. Then all the segments' new encoding vectors went into a BGRU with 64 hidden states to do another encoding along both the forward and backward temporal directions. The hidden-state sequences were then averaged using an attention pooling layer to form a fixed-dimension representation of the entire speech. The attention pooling layer implemented in this work followed the attention mechanism with context vector described in [16]. At last, another FC-BN-RELU structure followed by a FC layer with softmax activation was used to classify the emotion types. Here the first FC layer contains 64 neurons, and the second one contains the same number of neurons as the emotion types. Dropout of 0.2 was used for the linear transformation of the inputs in the BGRU layer and after all RELU activation.

We first conducted an emotion recognition experiment using the IEMOCAP data set to compare with published results to validate our model performance. We used the same experimental setup as in [10]. We run a 5-fold cross validation (session based split in the corpus) to categorize a speech sample to be one of four emotion types, i.e., neutral, happy, angry, and sad. In training, the model was trained for 80 epochs using AdamW [17] optimizer. The batch size was set to 64 and the learning rate was set to 0.001 initially and reduced by a factor of 0.5 if the validation loss is not improved for 5 epochs. The top two rows in Table 3 show the emotion recognition results on IEMOCAP. Following previous evaluations, both

[3]The feature list can be found at `https://github.com/tyiannak/pyAudioAnalysis/wiki/3.-Feature-Extraction`

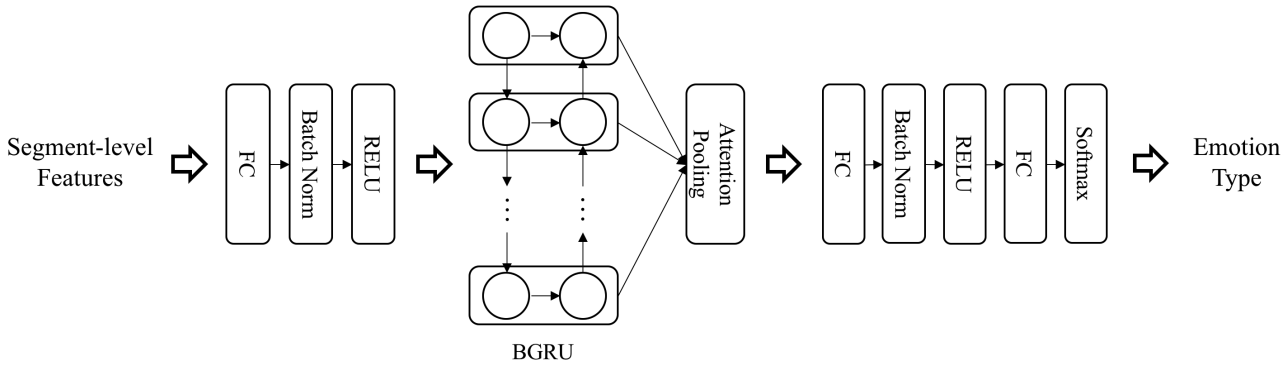

Fig. 1. The proposed NN model. First, the segment-level features are fed to a FC-BN-RELU structure to get new encodings. Second, BGRU followed by attention pooling is employed to get an audio-level encoding. Finally another FC-BN-RELU structure followed by a FC layer with softmax activation is used to predict the emotion type for the given audio.

weighted average recall (WA) and un-weighted average recall (UA) were reported. We can see that our attention-weighted BGRU model achieves better performance than the state-of-the-art results in [10].

Task	Data	WA (%)	UA (%)
[10]	IEMOCAP	63.5	58.8
neutral,happy,angry,sad		65.0	59.9
neutral vs. happy	Dubbing	77.9	65.0
neutral vs. angry		86.3	78.8
neutral vs. sad		88.9	67.5

Table 3. Top section: Emotion recognition performance on IEMOCAP data set for a 4-way classification task, compared to the previous result in [10]; Bottom section: SER performance when distinguishing neutral and other emotion types in our own dubbing data.

5.3.2. Applying the SER model on dubbing voices

We then used the L1 stimulus data described in Section 4 to train models for estimating users' emotion production. $2,177$ L1 speech files were separated into a training set and a testing set. For each emotion type, i.e., happy, angry, and sad, we used the samples with that type of emotion in the training set along with samples with neutral emotion to train a binary classifier. When evaluating emotion production, we use a default threshold of 0.5 to these binary classifiers' posterior probabilities to make the classification decision. Classification performance on the three emotion categories can be seen in the bottom rows of Table 3. These results show that we can apply automatic emotion recognition to detect if users produce the targeted emotion reasonably well. Furthermore, we calculated the correlation between the SER model's output posterior probabilities and human judged emotion scores, $r = 0.583$. This again suggests the usefulness of SER estimation.

Our final analysis aims to better understand the pronunciation and emotion behaviors on holistic ratings. We first perform a z-score transformation (with a zero mean, variance of 1) on Holistic scores, automatic pronunciation and emotion recognition scores. Then, we run an OLS analysis and obtained the formula as follows: $Holistic = 0.45 * emotion + 0.198 * pronunciation$. This shows that emotion behaviors have more than twice impact than pronunciation to the final decisions of holistic scores.

6. CONCLUSIONS

This paper is a first study investigating L2 learners' emotion production behaviors when using video dubbing to practice English speaking, which is now increasingly popular in China's mobile learning market. Through our data annotation work, we observe that a large percentage of L2 learners have emotion production when practicing emotional L1 stimulus. On the dubbing tasks using emotional L1 stimulus, we found that human perception in fact gives higher weights to emotion-related performance aspect than other speaking aspects, e.g., pronunciation, and the widely used automatic pronunciation evaluation method only shows a low correlation to human judged performance scores. This calls for more follow-up research to properly utilize the automatic pronunciation evaluation technology in more diverse conditions. Based on recent emotion recognition research, we developed an attention-based BGRU RNN model. Our model shows competitive performance to the previous reported result (such as [10]) in four-way emotion classification, or binary classification in emotion evaluation in dubbing setup.

For future research, one direction will be developing more types of features to better simulate human rated holistic scores on video dubbing task, for example, utilizing intonation or prosody (which is related to emotion) features. In addition, more work is needed to investigate the robustness of pronunciation evaluation when facing emotional stimulus.

7. REFERENCES

[1] D Luo, R Luo, and L Wang, "Naturalness Judgement of L2 English Through Dubbing Practice.," in *Interspeech*, San Francisco, CA, 2016, pp. 200–203.

[2] Yi Hui Chiu, "Can film dubbing projects facilitate EFL learners' acquisition of English pronunciation?," *British Journal of Educational Technology*, vol. 43, no. 1, pp. E24–E27, 1 2012.

[3] Y. Jin and W. Jie, "The development methods and principles for generating china national english level table's spoken language part (in chinese)," *Foreign Language World*, , no. 2, pp. 10–19, 2017.

[4] P He and S Wasuntarasophit, "The Effects of Video Dubbing Tasks on Reinforcing Oral Proficiency for Chinese Vocational College Students," *Asian EFL Journal*, vol. 17, no. 2, pp. 106–133, 2015.

[5] Martine Danan, "Dubbing projects for the language learner: A framework for integrating audiovisual translation into task-based instruction," *Computer Assisted Language Learning*, vol. 23, no. 5, pp. 441–456, 2010.

[6] Zheng Lian, Ya Li, Jianhua Tao, and Jian Huang, "Investigation of multimodal features, classifiers and fusion methods for emotion recognition," *arXiv preprint arXiv:1809.06225*, 2018.

[7] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.

[8] Christos Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2 2012.

[9] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," in *Language Resources and Evaluation*, 12 2008, vol. 42, pp. 335–359.

[10] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.

[11] Kyunghyun Cho, Bart Van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *Computer Science*, 2014.

[12] Martn Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, and Matthieu Devin, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2015.

[13] François Chollet et al., "Keras," `https://keras.io`, 2015.

[14] Theodoros Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[15] Sergey Ioffe and Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the 32nd International Conference on Machine Learning*, Francis Bach and David Blei, Eds., Lille, France, 07–09 Jul 2015, vol. 37 of *Proceedings of Machine Learning Research*, pp. 448–456, PMLR.

[16] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy, "Hierarchical attention networks for document classification," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2017, pp. 1480–1489.

[17] Ilya Loshchilov and Frank Hutter, "Fixing weight decay regularization in adam," *arXiv preprint arXiv:1711.05101*, 2017.